HIDDEN MARKOV MODELS: MULTIPLE PROCESSES AND MODEL SELECTION

By
RACHEL J. MACKAY
B.A.Sc., University of Waterloo, 1996
M.S., Cornell University, 1999

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
Department of Statistics

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 19, 2003

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date June 16, 2003

DE-6 (2/88)

Abstract

This thesis considers two broad topics in the theory and application of hidden Markov models (HMMs): modelling multiple time series and model selection. Of particular interest is the application of these ideas to data collected on multiple sclerosis patients. Our results are, however, directly applicable to many different contexts in which HMMs are used.

One model selection issue that we address is the problem of estimating the number of hidden states in a HMM. We exploit the relationship between finite mixture models and HMMs to develop a method of consistently estimating the number of hidden states in a stationary HMM. This method involves the minimization of a penalized distance function.

Another such issue that we discuss is that of assessing the goodness-of-fit of a stationary HMM. We suggest a graphical technique that compares the empirical and estimated distribution functions, and show that, if the model is misspecified, the proposed plots will signal this lack of fit with high probability when the sample size is large. A unique feature of our technique is the plotting of both the univariate and multivariate distribution functions.

HMMs for multiple processes have not been widely studied. In this context, random effects may be a natural choice for capturing differences among processes. Building on the framework of generalized linear mixed models, we develop the theory required for implementing and interpreting HMMs with random effects and covariates. We consider the case where the random effects appear only in the conditional model for the observed data, as well as the more difficult setting where the random effects appear in the model for the hidden process. We discuss two methods of parameter estimation: direct maximum likelihood estimation and the EM algorithm. Finally, to determine whether the additional complexity introduced by the random effects is warranted, we develop a procedure for testing the significance of their variance components.

We conclude with a discussion of future work, with special attention to the problem of the design and analysis of multiple sclerosis clinical trials.

Contents

List of Tables

List of Figures

Acknowledgements

There are so many people to thank! But, I will try to make this shorter than an Academy Awards speech.

First of all, I thank my advisor, John Petkau, for his support, both academic and financial. John has truly been a mentor – as a researcher, consultant, and administrator – and I feel honoured to have had the chance to work with him.

The faculty in the Statistics Department is outstanding; their enthusiasm and approachability create a very special environment for learning and developing new ideas. Thanks especially to Bertrand Clarke for many helpful and interesting conversations, as well as to my committee members, Jim Zidek, Nancy Heckman, and Paul Gustafson, for their feedback on my work. I am also indebted to the office staff, Christine Graham, Rhoda Morgan, and Elaine Salameh, for all their help with administrative matters. Finally, muchísimas gracias to Ruben Zamar for stepping in at the last minute to act as a University Examiner at my defense.

I have greatly appreciated the opportunities that the department has offered me, particularly the chance to be involved with SCARL. Jim Zidek, Chuck Paltiel, and Rick White have been patient and inspirational teachers.

My fellow graduate students have contributed to my time here in many ways. Special thanks to the members of the Journal Club (Steven Wang, Yinshan Zhao, Rong Zhu, Weiliang Qiu, and Jochen Brumm), Jérôme "DJ" Asselin, Fatemah Alqallaf, and to everyone who came out to English Lunches!

Outside the department, I am grateful to Ya'acov Ritov of the Hebrew University of Jerusalem and to Jiahua Chen of the University of Waterloo for helpful correspondence. Thank you also to Paul Albert of NIH and Drs. Don Paty and David Li of the UBC MS/MRI Research Group for sharing their MS/MRI data. Finally, I thank Rolf Turner of the University of New Brunswick for his generous assistance with the faecal coliform data (provided by Geoff Coade of the NSW Environmental Protection Authority).

On a more personal note, I cannot express how grateful I am to my family and friends for their continual love and support. My dad has been an amazing role model for me, both academically and personally. Our "father-daughter" talks have helped me through many a difficult moment: I wouldn't have graduated before the age of 70 without him! My mom has put up with countless hours of "math talk", and has had an unwavering faith in me, even at those times when I lost faith in myself. I also need to thank my sister, "Vermin",

for always understanding me, and, of course, for her Markov jokes. The friends that have helped me are too numerous to mention here. But, most importantly, My Dang and Cindy Rejwan have been with me every step of the way. And, my Waterloo engineering buddies have always been sources of encouragement and much needed humour.

Finally, I thank my partner and future husband, Yevgeni Altman, with all my heart for his tireless love, patience, and support. I look forward to spending the rest of our lives together, thesis-free.

RACHEL MACKAY

The University of British Columbia
June, 2003

To my grandmother,

Denilde Gertrude Brodie (1924-2002),

whose love is with me always.

Chapter 1

Introduction

Hidden Markov models (HMMs) describe the relationship between two stochastic processes: an observed process and an underlying "hidden" (unobserved) process. These models form a class of mixture models where, given the hidden state at time t, the distribution of the observation at this time is fully specified. However, HMMs are more general than classical mixture models in that the hidden states are not assumed to be independent, but rather to have a Markovian structure. One consequence of this assumption is that the observed data are also correlated, with dependence between observations decreasing to zero as the distance between them increases to infinity. This correlation is long-range, in the sense that HMMs are not Markov chains.

In general, these models are used for two purposes. The first is to make inferences or predictions about an unobserved process based on the observed process. For example, HMMs have been used successfully for the purposes of prediction in the field of speech recognition (e.g. Levinson *et al.* 1983). In this context, the observed data – the acoustic signal – may be modelled as a function of unobserved articulatory configurations such as vocal tract shape or tongue movement. For each word from a vocabulary of size V, $V < \infty$, an acoustic signal is generated, and the parameters of the associated HMM estimated. Then, given a signal generated from an unknown word in this vocabulary, these V models can be used to predict which word was uttered. Some advanced systems based on HMMs now perform as accurately as their human counterparts (Juang & Rabiner 1991). Similarly, HMMs have been used in molecular biology for gene recognition (see, e.g., Krogh 1998). Here, sequenced strands of DNA are treated as functions of the underlying signals that comprise the structure of a gene. The models estimated from sequences with known genetic structure are then used to predict the location of genes in new sequences.

A second reason for using HMMs is to explain variation in the observed process based on

variation in a postulated hidden process. In this paradigm, a HMM captures over-dispersion (relative to a standard distribution) in the observed data. In particular, a HMM attributes this over-dispersion to the key model feature that observations come from one of several different marginal distributions, each associated with a different latent state. When physical meaning can be attributed to these states, a HMM provides a natural model for such data. As an illustration, Leroux & Puterman (1992) use a HMM to model the number of foetal lamb movements in consecutive 5-second intervals. The distribution of each observation is assumed to depend on whether the lamb is in a relaxed or excited state. As another example, Albert (1991) models the distribution of epileptic seizure frequencies according to whether the patient is in a high or low seizure activity state.

Magnetic resonance imaging (MRI) scans of relapsing-remitting multiple sclerosis (MS) patients are another source of data that may be appropriately modelled by HMMs. Patients afflicted with this type of MS experience lesions on the brain stem, with symptoms worsening and then improving in alternating periods of relapse and remission. Typical data amassed during clinical trials consist of lesion counts at regular time intervals for a collection of patients. It is now believed that exacerbations are associated with increased numbers of lesions on the brain stem. Thus, it may be reasonable to assume that the distribution of the lesion counts depends on the patient's (unobserved) disease state, i.e. whether the patient is in relapse or remission. Additionally, we might expect to see autocorrelation in this sequence of disease states. Indeed, Albert *et al.* (1994) use this idea in the development of a HMM for individual relapsing-remitting MS patients.

The study of HMM theory began in the late 1960's. One of the key papers, Baum *et al.* (1970), provides a method of obtaining the maximum likelihood estimates (MLEs). The asymptotic properties of the MLEs have subsequently been established (Leroux 1992a; Bickel *et al.* 1998; Douc & Matias 2001). Likelihood ratio tests for HMMs have been studied by Giudici *et al.* (2000), and Bickel *et al.* (2002) have determined bounds on the expectations of the HMM log-likelihood and its derivatives. However, theoretical gaps remain. This thesis addresses two such topics that have not been adequately studied in the literature: modelling multiple time series, and model selection techniques. We will use the MS/MRI context described above to illustrate many of our ideas. In particular, we will consider two MS/MRI data sets: that used by Albert *et al.* (1994), and another similar data set involving the 13 placebo patients from the Vancouver cohort of the PRISMS study (PRISMS Study Group 1998).

Most work to date on HMM theory has concentrated on models for a single observed process. In Chapter 2, we introduce some basic definitions and concepts in this setting. We then explore the model considered by Albert *et al.* (1994), as well as some simple extensions.

This exploration, which includes a discussion of the limitations of the theory surrounding these models, will serve to clarify the fundamental ideas behind HMMs, and will elucidate our questions of interest.

In Chapters 3 and 4, working in the context of a single, stationary HMM, we address two questions of critical importance in the application of HMMs: estimation of the number of hidden states and assessment of goodness-of-fit (GOF). In Chapter 3, we develop a method of consistently estimating the number of hidden states. Our method extends the work of Chen & Kalbfleisch (1996), who consider the use of a penalized distance function for estimating the number of components in a finite mixture model. We apply our procedure to the MS/MRI data collected by Albert *et al.* (1994), and carry out a small simulation study that suggests the method performs reasonably well for finite samples. This work was published in *The Canadian Journal of Statistics* (MacKay 2002). In Chapter 4, we propose a graphical technique for assessing the GOF of a HMM. Specifically, we show that plots comparing the empirical and estimated distribution functions – both univariate and multivariate – allow the detection of lack of fit in the model with high probability as the sample size grows. We use this technique to study the appropriateness of various HMMs for the two MS/MRI data sets.

Chapter 2 will also motivate the need for HMMs for multiple processes, which is the focus of Chapter 5. In that chapter, we propose the incorporation of random effects as a means of linking the different processes. Random effects provide an efficient way of modelling commonalities among patients while allowing for some inter-patient variability. Furthermore, including random effects permits greater flexibility in the modelling of the correlation structure of the observed data. Using the generalized linear mixed model (GLMM) framework, we consider the incorporation of random effects and covariates in both the conditional model for the observed process and the model for the hidden process. We discuss the estimation of these models, as well as their interpretation, with special attention to the impact of the random effects on the marginal moments of the observed data.

In Chapter 6, we address the issue of hypothesis tests for the parameters of the class of models developed in Chapter 5. We comment on the asymptotic properties of the MLEs, and suggest settings where standard test procedures may be appropriate. We then present a method for testing the significance of the variance components, which is a more challenging problem since the null hypothesis puts at least one parameter on the boundary of the parameter space. Our method has its inspiration in the score test proposed by Jacqmin-Gadda & Commenges (1995) in the GLMM context. We provide two illustrations of the theory in Chapters 5 and 6 to demonstrate the practicality of using our class of models in applications. The first involves the MS/MRI lesion count data from the Vancouver PRISMS study; the

second considers the analysis of repeated measurements of faecal coliform counts at several oceanic sites. We end this chapter with a modest simulation study to investigate the power of this method for finite samples.

We conclude with Chapter 7, where we summarize the work in this thesis and present ideas for future research in the field of HMMs. Of particular interest is the application of our theory to the design and analysis of MS/MRI clinical trials. Our discussion focuses on the issues that we anticipate will arise in this work.

Chapter 2

Hidden Markov Models for a Single Process

In this chapter, we provide the formal definition of a HMM for a single process, as well as some theory relevant to parameter estimation and hypothesis testing in this setting. We then illustrate these ideas with an application to a MS/MRI data set. The primary purposes of this chapter are to introduce basic concepts and to highlight our research questions of interest.

Throughout this thesis, we use the generic notation $f(x)$ to denote the density (or probability mass function) of a random variable (or vector), X. Usually, f will be a member of a parametric family with parameters ψ, in which case we will write $f(x;\psi)$. We will use bold face to indicate a vector, such as $\mathbf{Y}$ and $\mathbf{Z}$ to denote the vectors of observed responses and hidden states, respectively.

2.1 Definition of a Hidden Markov Model

Let Y_t be the observed response at time t, and let Z_t be the hidden state at time t, $t = 1, \ldots, n$. The process $\{Y_t\}$ is a discrete-time HMM if

- $\{Z_t\}$ is a Markov chain with transition probabilities $\{P^t_{k\ell}\}$ and initial probabilities $\{\pi_k\}$.
- $Y_t|Z_t$ is independent of $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_n$ and $Z_1, \ldots, Z_{t-1}, Z_{t+1}, \ldots, Z_n$.

Typically, the following assumptions are also made:

1. The density (or probability mass function) of $Y_t|Z_t$ is $h(\cdot\,; \theta_{Z_t}, \phi)$, where h is a parametric family indexed by the parameters $(\theta_{Z_t}, \phi) \in \Theta$.

2. $Z_t \in \{1, \ldots, K\}$, where K is known and finite.

3. The values of $\{\theta_k\}$ are distinct.

4. The time points $t = 1, \ldots, n$ are equally spaced.

5. $P_{k\ell}^t \equiv P_{k\ell}$, $k, \ell = 1, \ldots, K$, i.e. the transition probabilities are homogeneous.

6. $\{Z_t\}$ is stationary.

REMARK. Assumption 1 implies that the distribution of $Y_t|Z_t$ depends on t only through Z_t. Our notation indicates that some parameters (θ_{Z_t}) may vary with Z_t, whereas others (ϕ) are common across the hidden states. We relax Assumption 2 in Chapter 3, where we address the issue of estimating K. Assumption 3 is made in most applications of HMMs, with the notable exception of ion channel modelling (e.g. Chung *et al.* 1990). We discuss this issue in more detail in Section 3.2. Assumption 4 allows us to specify the model in terms of the 1-step transition probabilities, and hence is a useful simplification, but it is not strictly necessary (see, e.g., Section 6.4.3). Assumption 5 is also standard, though Hughes & Guttorp (1994) have considered non-homogeneous HMMs in applications. One advantageous consequence of Assumption 6 is that the random variables $\{Y_t\}$ are identically distributed – a feature that sometimes permits the extension of existing theory for iid random variables to the HMM setting (see, e.g., Chapters 3 and 4).

It is interesting to note that the marginal distribution of Y_t is a finite mixture:

$$f(y_t) = \sum_{k=1}^{K} f(y_t \mid Z_t = k)\ \pi_k.$$

However, the sequence of hidden states, $\{Z_t\}$, is allowed to have a Markovian, rather than independent, structure. Thus, we see that stationary HMMs are a generalization of traditional finite mixture models.

2.2 Maximum Likelihood Estimation

The likelihood associated with the model described in Section 2.1 is not a simple product of marginal distributions. Define $\psi = (\theta_1, \ldots, \theta_K, \phi, P_{11}, P_{12}, \ldots, P_{KK}, \pi_1, \ldots, \pi_K)$, and denote

the likelihood by $\mathcal{L}(\psi)$. Then, using the assumption that $\{Y_t\}$ are independent given $\{Z_t\}$,

$$\begin{aligned}
\mathcal{L}(\psi) &= f(\mathbf{y};\psi) \\
&= \sum_{\mathbf{z}} f(\mathbf{y}|\mathbf{z},\psi) f(\mathbf{z};\psi) \\
&= \sum_{\mathbf{z}} \prod_{t=1}^{n} f(y_t|z_t,\psi) \cdot f(z_1;\psi) \prod_{t=2}^{n} f(z_t|z_{t-1},\psi) \\
&= \sum_{\mathbf{z}} \prod_{t=1}^{n} h(y_t;\theta_{z_t},\phi) \cdot \pi_{z_1} \prod_{t=2}^{n} P_{z_{t-1},z_t}. \qquad (2.1)
\end{aligned}$$

Thus, we see that the likelihood involves a summation over the K^n possible values of $\mathbf{z}$, and hence is quite complicated.

We can simplify (2.1) somewhat by recognizing that for each t, the variable z_t appears in only a few factors. So

$$\mathcal{L}(\psi) = \sum_{z_1} \pi_{z_1} h(y_1;\theta_{z_1},\phi) \sum_{z_2} P_{z_1,z_2} h(y_2;\theta_{z_2},\phi) \cdots \sum_{z_n} P_{z_{n-1},z_n} h(y_n;\theta_{z_n},\phi).$$

This expression can then be written as a product of matrices (MacDonald & Zucchini 1997, Chapter 2). In particular, let A^1 be the vector with elements $A^1_k = \pi_k \; h(y_1;\theta_k,\phi)$, and let A^t be the matrix with elements $A^t_{k\ell} = P_{k\ell} \; h(y_t;\theta_\ell,\phi)$, $t > 1$. Let $\mathbf{1}$ be a the K-dimensional vector of 1's. Then

$$\mathcal{L}(\psi) = (A^1)' \left\{ \prod_{t=2}^{n} A^t \right\} \mathbf{1}, \qquad (2.2)$$

which is a very simple expression to compute. This form of the likelihood illustrates that the number of hidden states, K, has a far greater impact on the computational effort associated with maximum likelihood estimation than the number of observations, n.

Traditionally, the EM algorithm (Dempster *et al.* 1977) has been used to maximize HMM likelihoods. There are two likely reasons for the popularity of this algorithm. Firstly, for homogeneous HMMs (see Assumption 5 in Section 2.1) taking on only a finite number of values and with unknown initial probabilities, this algorithm reduces to an iterative procedure with simple, closed-form expressions for the parameter estimates at each iteration. In this context, the EM algorithm is often called the Forward-Backward algorithm and is credited to Baum *et al.* (1970). Details are provided in Appendix A.1. In terms of estimation of the parameters of a HMM, this case is the simplest possible, and will be a useful reference point when assessing the difficulty of estimating the parameters of the more complicated models we consider in Chapter 5.

A second reason is that derivatives of HMM likelihoods are somewhat difficult to compute, requiring iterative methods (e.g. Rynkiewicz 2001). The EM algorithm, unlike methods such as Newton-Raphson, does not require that derivatives be supplied.

However, in general, the steps of the EM algorithm do not involve closed-form expressions. Furthermore, this algorithm is notoriously slow to converge. Thus, we prefer direct numerical maximization of the likelihood, which is typically much more efficient (MacDonald and Zucchini 1997, Chapter 2). In particular, we have found that the quasi-Newton routine (Nash 1979) tends to locate maximum likelihood estimates (MLEs) more accurately and with far less computational effort than the EM algorithm. Even if the EM algorithm performs better than direct maximization under some circumstances (such as when we have a large number of parameters or poor starting values), repeating the direct maximization procedure using a variety of starting values still seems to be the most efficient means of parameter estimation.

Starting values are of critical importance since HMM likelihoods tend to have many local maxima. These values may be selected using, for example, the method suggested by Leroux & Puterman (1992). In addition, we recommend doing a grid search (over a variety of reasonable starting values) to improve our chances of locating the global maximum.

Another implementation issue concerns the parameters $\{\pi_k\}$, which are normally considered to be nuisance parameters. Three options exist for the dealing with these parameters. Firstly, we may assume values for $\{\pi_k\}$. In the absence of prior information, this option may not be reasonable. On the other hand, Leroux (1992a) shows that the consistency of the MLEs does not depend on the choice of initial distribution. Thus, for processes observed at a large number of time points, this option may be appealing.

Secondly, we may estimate $\{\pi_k\}$ from the data. This option is also undesirable for relatively small data sets since it would require the estimation of $K-1$ additional parameters, risking an increase in the standard errors of the parameters of interest.

Finally, if the hidden process can be assumed to be stationary, we can treat $\{\pi_k\}$ as functions of the transition probabilities. In this way, we can reduce the number of parameters to estimate. This option, while requiring the solution of a system of K linear equations at each iteration of the maximization procedure, seems to be the most attractive as long as the assumption of stationarity is appropriate. We thus use this approach in the examples we consider in this thesis.

2.3 Asymptotic Properties of the MLEs

Results on the properties of the MLEs require that the model (2.1) is identifiable, i.e.

$$f(\mathbf{y};\psi_1) = f(\mathbf{y};\psi_2) \text{ if and only if } \psi_1 = \psi_2.$$

Strictly speaking, a HMM is never identifiable, in the sense that we can always permute the labels of the hidden states without changing the likelihood. However, it is easy to overcome this obstacle by imposing appropriate restrictions on the parameters (such as an ordering of $\{\theta_k\}$ in the case where these values are distinct). Setting this point aside, determining sufficient conditions for identifiability is still a difficult problem, except when the HMM is stationary with distinct values of $\{\theta_k\}$. See Section 3.2 for details.

In the case of a stationary HMM where the hidden process takes on only a finite number of values, Leroux (1992a) and Bickel *et al.* (1998) establish the consistency and asymptotic normality, respectively, of the MLEs under quite general conditions. In addition to assuming model identifiability, these authors impose mild conditions on the transition probabilities and on the distribution h. These components of the model are usually quite simple (in contrast with the full likelihood), and hence these conditions are relatively easy to verify (and hold for most models). Douc & Matias (2001) show that the MLEs are also consistent and asymptotically normal in the case where $\{Z_t\}$ belongs to a compact set and is possibly non-stationary. Again, these authors impose conditions only on the Markov transition kernel and the distribution h. We are not aware of any results in the literature regarding the asymptotic properties of non-homogeneous HMMs.

With respect to inference about the unknown parameters, Bickel *et al.* (1998) and Douc & Matias (2001) show that the observed information converges in probability to the Fisher information matrix. Thus, if the HMM satisfies the conditions imposed by these authors, we can conduct Wald tests in the standard way, using the observed information to estimate the variance-covariance matrix of the MLEs. In addition, Giudici *et al.* (2000) show that, in the comparison of nested stationary HMMs with a common, known value of K, the likelihood ratio test statistic has the usual asymptotic χ^2 distribution. Their theory is applicable, for example, to test whether the hidden states have an independent, rather than Markovian, structure.

2.4 Application to MS/MRI Data

In this section, we discuss an interesting and unusual HMM developed by Albert *et al.* (1994) for MS/MRI data. We fit this model to their data and give the results in Section 2.4.1. We then develop several extensions, which we present in Sections 2.4.2–2.4.5. One purpose of this discussion is to solidify the concepts in Sections 2.1–2.3. Furthermore, with an eye towards future work on the design and analysis of MS/MRI clinical trials (see Chapter 7), this section will illustrate the type of questions that might be asked in this setting. Most

importantly, some of these questions will reveal gaps in existing theory for HMMs, which will motivate the research presented in this thesis.

The HMM proposed by Albert *et al.* (1994), to which we will henceforth refer as Albert's model, describes lesion counts on repeated MRI scans for a single relapsing-remitting MS patient. The authors apply the model individually to three patients, each of whom had monthly MRI scans for a period of approximately 30 months. The observed lesion counts range from 0 to 19, with a mean of 4.5 and a median of 4 lesions per scan. The data are displayed in Figure 2.1.

Albert's model is based on the idea that a patient is in an (unknown) state of deterioration or improvement at any time point. This underlying state will affect the mean number of lesions observed at that time. Specifically, it is assumed that if the patient's condition is deteriorating at time t, the mean lesion count at this time will be greater than the mean lesion count at time $t-1$ by a factor of θ. Similarly, if the patient's condition is improving at time t, the mean lesion count at this time will be less than the mean lesion count at time $t-1$ by a factor of θ.

Mathematically, the assumptions of the model can be stated as follows:

1. The hidden state, Z_t, is -1 if the patient's condition is improving at time t, and $+1$ if the patient's condition is deteriorating at time t. This process is modelled as a stationary Markov chain.

2. The transition probabilities are assumed to be homogeneous, with the probability of moving from deterioration to improvement equal to the probability of moving from improvement to deterioration. This common probability is denoted by γ.

3. Given Z_t, the lesion count, Y_t, is assumed to be independent of $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_n$, and distributed as Poisson(μ_t), where

$$\mu_t = \begin{cases} \theta\mu_{t-1}, & \text{if the patient is deteriorating at time } t \\ (1/\theta)\mu_{t-1}, & \text{if the patient is improving at time } t. \end{cases}$$

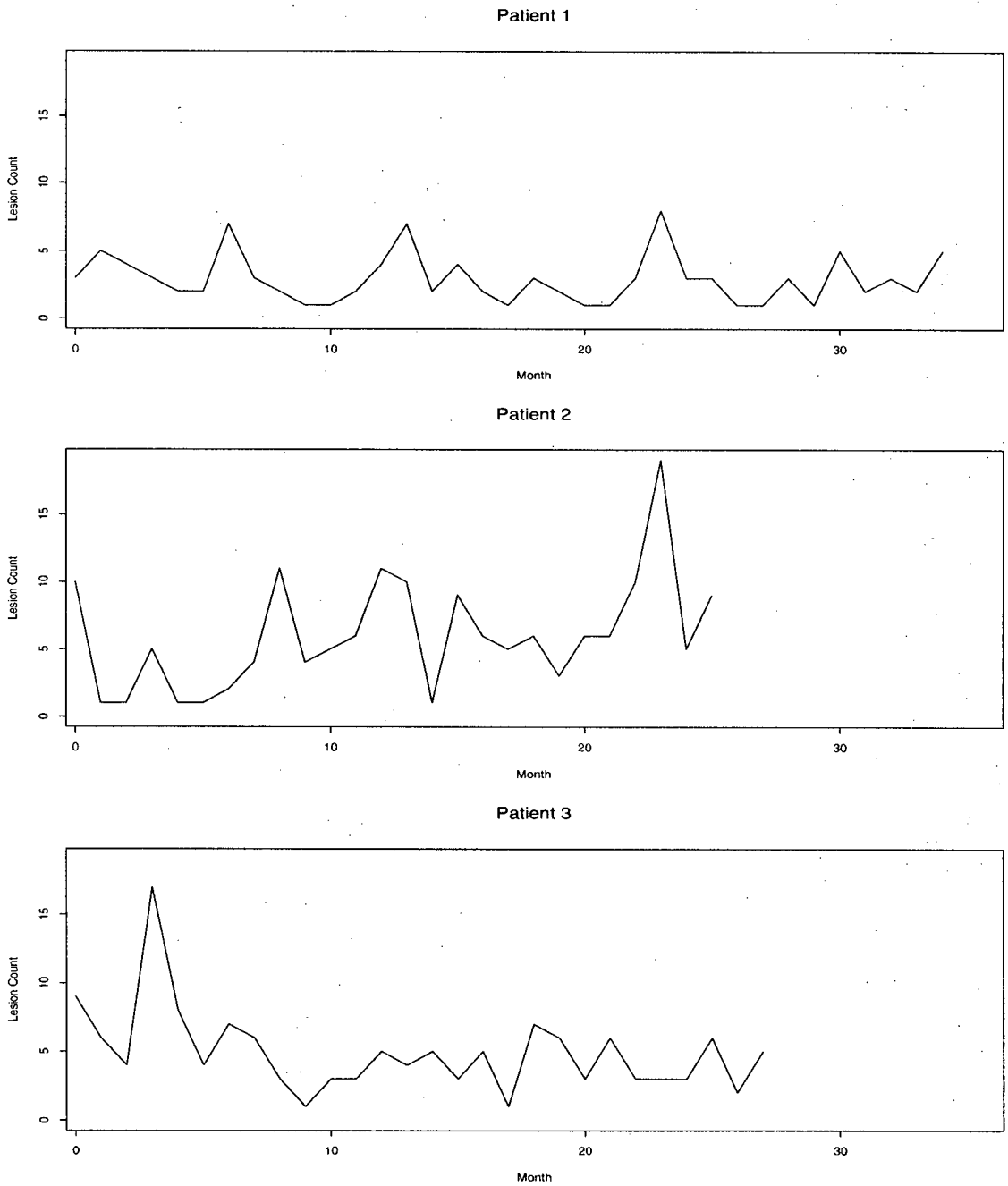
   This assumption can be rewritten as

$$\mu_t = \mu_0\theta^{S_t},$$

   where $S_t = \sum_{i=1}^{t} Z_i$ and μ_0 is the baseline mean lesion count.

REMARK. Assumptions 1 and 2 imply that the initial probabilities are $\mathrm{P}(Z_1 = -1) = \mathrm{P}(Z_1 = +1) = 0.5$. Under Assumption 3, when $\theta = 1$, the model reduces to that

Figure 2.1: MS/MRI data analyzed in Albert *et al.* (1994)



for independent observations distributed as Poisson(μ_0). Assumption 3 also leads to an identifiability problem, since the model with θ is equivalent to the model with $\frac{1}{\theta}$ if we reverse the labelling of the hidden states. To remedy this problem, we assume that $\theta \geq 1$.

On first glance, this model does not appear to be a HMM (according to the definition given in Section 2.1), since the distribution of Y_t depends on *all* previous hidden states, not just Z_t. However, by defining the hidden process as (S_{t-1}, S_t) with state space $\{(i,j) : i = -n, \ldots, n, j = -n, \ldots, n\}$, we see that Albert's model does, in fact, conform to the definition of a non-homogeneous HMM with countable state space.

Thus, the model can be fit in the manner described in Section 2.2. Albert *et al.* use the EM algorithm for this purpose, but we will maximize the likelihood directly. We have two reasons for this choice. First, as discussed in Section 2.2, direct maximization appears to be the more efficient of the two methods. Second, in our experience with this model, unlike the direct maximization method, the EM algorithm tends to converge to values other than the MLEs.

To make inferences about Albert's model and its extensions, we will use the methods outlined in Section 2.3. Because we lack theoretical results about the properties of these methods in the case of non-homogeneous HMMs, our conclusions should be considered only informal. These conclusions are nonetheless useful, as they will help to isolate and illustrate the issues of interest in this thesis. Furthermore, Albert's model is very similar to a stationary Poisson HMM. To see this, note that the mean lesion count at time t is restricted to a discrete number of values (evenly spaced on the log scale). If we assume that the observed process is stationary, it is reasonable to use a finite approximation to these mean values, i.e. to assume that the mean at time t is one of K values. This new model is simply a stationary Poisson HMM with K hidden states and with some restrictions on the transition probabilities. Hence, if more formal conclusions were desired, one could fit a stationary Poisson HMM with an appropriate value of K to the data. Then, the standard inference results discussed in Section 2.3 would certainly apply.

2.4.1 Albert's Model

We can facilitate the maximization of the likelihood by transforming the parameters so that their range is the entire real line. To this end, we use the following reparameterizations: $\mu_0^* = \log \mu_0$, $\theta^* = \log(\theta - 1)$, and $\gamma^* = \log(\gamma/(1-\gamma))$.

Table 2.1 gives the parameter estimates and approximate standard errors resulting from fitting Albert's model to the three patients' data. For Patient 1, θ is estimated as 1.000, i.e.

Table 2.1: Parameter estimates and standard errors for Albert's model

		Patient 1		Patient 2		Patient 3	
Parameter	Transformation	Estimate	SE	Estimate	SE	Estimate	SE
μ_0^*	$\log \mu$	1.070	0.091	1.362	0.178	2.223	0.244
θ^*	$\log(\theta - 1)$	-11.083	NA	-0.282	0.245	-1.128	0.560
γ^*	$\log\left(\frac{\gamma}{1-\gamma}\right)$	NA	NA	1.241	0.575	1.336	0.847
	$\log \mathcal{L}$	-66.814		-72.029		-65.363	

the model reduces to that for independent Poisson counts. In this case, the estimate for γ given by the quasi-Newton routine is, in fact, arbitrary. In addition, since $\theta = 1$ is on the boundary of the parameter space, there is no guarantee that the usual standard error for the estimate of θ^* is even approximately correct. For these reasons, we do not provide an estimate of γ^*, or a standard error for the estimate of θ^*.

One question of interest is whether the complexity of the HMM is warranted, or whether the simple model with independent Poisson counts is sufficient to describe the variability and correlation in the data. In principle, we should be cautious about making such inferences, since the test of $\theta = 1$ is a boundary problem. Informally, though, in the case of Patient 1, there is no evidence to suggest that the simpler model is inadequate. In the case of Patients 2 and 3, if we believe the 95% confidence intervals for θ ($[1.467, 2.219]$ and $[1.108, 1.970]$, respectively), then there is evidence against the null hypothesis that $\theta = 1$. Thus, the HMM structure seems to be more appropriate for these patients than the simpler model. These conclusions are consistent with those arrived at by Albert *et al.*

The results in Table 2.1 are essentially the same as those obtained by Albert *et al.* The primary differences are that we have omitted the estimate of γ^* in the case of Patient 1, and have achieved a higher value of the likelihood in the case of Patients 2 and 3. The latter emphasizes the importance of choosing both good starting values and an estimation method whose convergence properties are well-behaved in practice as well as in theory.

2.4.2 Generalization of the Transition Probabilities

The first extension we consider (for Patients 2 and 3) is the use of general transition probabilities. We do not apply this new model to the data from Patient 1 since the analysis in Section 2.4.1 indicates that the transition probabilities for this patient are arbitrary.

Table 2.2: Parameter estimates and standard errors for the model with general transition probabilities

		Patient 2		Patient 3	
Parameter	Transformation	Estimate	SE	Estimate	SE
μ_0^*	$\log \mu_0$	1.360	0.173	2.186	0.143
θ^*	$\log(\theta - 1)$	−0.281	0.244	−1.416	0.273
γ^*	$\log\left(\frac{\gamma}{1-\gamma}\right)$	1.489	0.801	0.975	0.515
β^*	$\log\left(\frac{\beta}{1-\beta}\right)$	1.039	0.690	18.570	NA
	$\log \mathcal{L}$	−71.921		−64.222	

In particular, we model the transition probabilities as

$$\begin{array}{c} \\ -1 \\ +1 \end{array} \begin{array}{c} \begin{array}{cc} -1 & +1 \end{array} \\ \begin{bmatrix} 1-\gamma & \gamma \\ \beta & 1-\beta \end{bmatrix} \end{array}.$$

This generalization may be of interest because Albert's model assumes that patients spend 50% of their time in a state of deterioration, and 50% in a state of improvement. This assumption seems too strong to make *a priori*. The parameter estimates for this model are given in Table 2.2. In the case of Patient 3, β is estimated as 1.000, which is on the boundary of the parameter space. We do not include a standard error for the estimate of β^* for this reason.

To test the validity of the assumption that $\gamma = \beta$, we note that Albert's model is nested within the more general model. We then use the likelihood ratio test (LRT) to compare the two models, assuming that the LRT statistic has an asymptotic χ_1^2 distribution. The p-values for these tests are

	Patient 2	Patient 3
p-value	0.642	0.131

Surprisingly, the more general model does not fit substantially better for either of the two patients.

We have two possible explanations for these results. Firstly, the standard errors of the estimates of γ^* given in Table 2.1 are quite large relative to the estimates themselves, and relative to the standard errors of the estimates of μ_0^* and θ^*. The same is true of the standard errors of the estimates of γ^* and β^* in Table 2.2. These examples show that making inferences about the hidden process is usually a difficult problem.

A second explanation may lie in the structure of μ_t. Note that the proportional increase

in the mean when the patient is deteriorating, θ, is, by assumption, equal to the proportional decrease in the mean when the patient is improving. In the case where there is no overall trend in the data (as is true for these particular patients, as well as for relapsing-remitting patients in general when observed over a short time period), the number of transitions from decreasing to increasing mean is forced to equal approximately the number of transitions from increasing to decreasing mean. This statement is equivalent to Albert's assumption that the patients spend equal proportions of time in the states of deterioration and improvement. These proportions can be expressed as

$$\left[\frac{\beta}{\beta+\gamma}, \frac{\gamma}{\beta+\gamma}\right], \tag{2.3}$$

so $\frac{\beta}{\beta+\gamma} = \frac{\gamma}{\beta+\gamma} = 0.5$ implies that $\beta = \gamma$. It is perhaps for this reason that the model with general transition probabilities does not provide an improved fit to these data.

Under some circumstances, we would expect to see a trend in the lesion counts, in which case the more general model might be appropriate. For example, patients with secondary progressive MS have lesion counts which may steadily increase over a given time period. Similarly, in a clinical trial setting where patients may be chosen for their relatively high level of disease activity, an initial downward trend ("regression to the mean") may be observed, even in the placebo patients. The simulated data in Figure 2.2 show a clear upward trend. These data were generated from this HMM with $\mu_0 = 1.5$, $\theta = 1.2$, $\gamma = 0.8$, and $\beta = 0.2$. In this case, the maximum value of the log-likelihood is -73.345 for the restricted model and -69.446 for the general model. The LRT leads to a p-value of 0.005, indicating that the general model provides a significantly improved fit to these data.

2.4.3 Generalization of the Conditional Mean Structure

In light of the discussion in Section 2.4.2, we might consider modelling μ_t more generally while leaving the transition probabilities as in Section 2.4.1. For example, we could fit Albert's model with the modification

$$\mu_t = \begin{cases} \theta_0 \mu_{t-1}, & \text{if patient is deteriorating at time } t \\ (1/\theta_1)\mu_{t-1}, & \text{if patient is improving at time } t \end{cases}. \tag{2.4}$$

The parameter estimates and approximate standard errors associated with fitting this model are given in Table 2.3. When $\theta_0 = 1/\theta_1 \equiv \theta$, the model implies that $\{Y_t\}$ are independent with Y_t distributed as Poisson$(\mu_0\theta^t)$, in which case γ is arbitrary. Thus, for Patient 1, we omit the estimate of γ^*.

This modification does not significantly improve the fit for Patient 1, but we observe

Figure 2.2: Simulated data with a trend

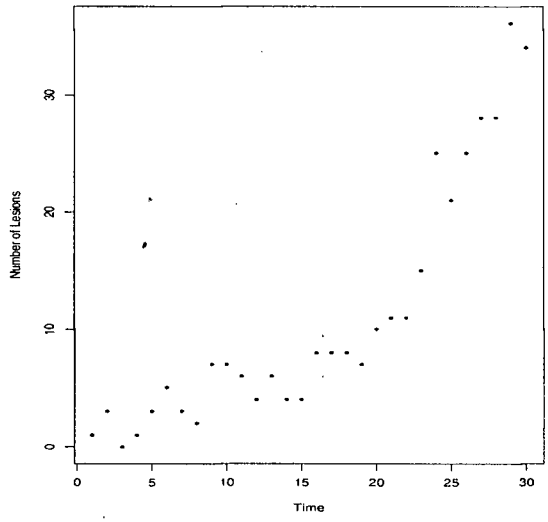


Table 2.3: Parameter estimates and standard errors for the model with a general conditional mean structure

		Patient 1		Patient 2		Patient 3	
Parameter	Transformation	Estimate	SE	Estimate	SE	Estimate	SE
μ_0^*	$\log \mu_0$	1.168	0.200	0.987	0.256	2.446	0.247
θ_0^*	$\log \theta_0$	0.006	0.089	0.484	0.621	0.466	0.166
θ_1^*	$\log \theta_1$	−0.006	0.089	0.621	0.120	0.398	0.158
γ^*	$\log\left(\frac{\gamma}{1-\gamma}\right)$	NA	NA	0.974	0.708	2.384	0.813
	$\log \mathcal{L}$	−66.652		−70.401		−63.904	

some evidence of an improved fit for Patients 2 and 3. The p-values associated with these tests are

	Patient 1	Patient 2	Patient 3
p-value	0.569	0.071	0.088

2.4.4 Generalization of the Conditional Mean Structure and of the Transition Probabilities

It was anticipated that, in the case of Patients 2 and 3, the fit of the model might be further improved by incorporating general transition probabilities (i.e. by combining the models proposed in Sections 2.4.2 and 2.4.3). In fact, the LRTs comparing this model to the model with $\gamma = \beta$ (i.e. the model in Section 2.4.3) yield the following p-values:

	Patient 2	Patient 3
p-value	0.065	1.000

Thus, there is some support for the expanded model in the case of Patient 2.

The estimates of the transformed parameters and approximate standard errors are given in Table 2.4.

2.4.5 Addition of a Third Hidden State

Our final question of interest regarding Albert's model involves the choice of the number of hidden states. Albert's model is quite restrictive, in the sense that it forces the mean lesion count to either increase or decrease from time $t-1$ to time t. One would imagine that,

Table 2.4: Parameter estimates and standard errors for model with general transition probabilities and conditional mean structure

		Patient 2		Patient 3	
Parameter	Transformation	Estimate	SE	Estimate	SE
μ_0^*	$\log \mu_0$	1.889	0.235	2.445	0.233
θ_0^*	$\log \theta_0$	0.811	0.169	0.466	0.154
θ_1^*	$\log \theta_1$	0.383	0.071	0.398	0.149
γ^*	$\log\left(\frac{\gamma}{1-\gamma}\right)$	2.022	1.134	2.372	1.045
β^*	$\log\left(\frac{\beta}{1-\beta}\right)$	−0.399	0.505	2.396	1.052
	$\log \mathcal{L}$	−68.695		−63.904	

especially during periods of remission, the mean lesion count would remain stable. Thus, we consider the addition of a third hidden state, state 0, where the patient's condition is neither deteriorating nor improving. Mathematically, this modification can be expressed as

$$\begin{aligned}\mu_t &= \begin{cases} \theta\mu_{t-1}, & \text{if patient is deteriorating at time } t \\ \mu_{t-1}, & \text{if patient is stable at time } t \\ (1/\theta)\mu_{t-1}, & \text{if patient is improving at time } t \end{cases} \\ &= \mu_0\theta^{S_t}\end{aligned}$$

We represent the transition probabilities as follows:

$$\begin{array}{c} \\ -1 \\ 0 \\ +1 \end{array} \begin{array}{c} \begin{array}{ccc} -1 & 0 & +1 \end{array} \\ \begin{bmatrix} p_1 & p_2 & 1-p_1-p_2 \\ p_3 & p_4 & 1-p_3-p_4 \\ p_5 & p_6 & 1-p_5-p_6 \end{bmatrix} \end{array}$$

One disadvantage of such an extension is the introduction of the problem of computing the stationary probabilities, which are used as initial probabilities for the hidden Markov chain. In the two-dimensional case, we have the simple, closed form (2.3) for the stationary distribution. In order to compute the stationary distribution in the three-dimensional case, however, a system of three linear equations must be solved at each iteration of the quasi-Newton algorithm. Another disadvantage of this model is the large number of unknown parameters. However, we can reduce this number if we are willing to place restrictions on the transition probabilities (as in Albert's model), for example by assuming that the transition probability matrix is symmetric.

The parameter estimates and standard errors are given in Table 2.5. In this case, the likelihood functions are quite flat (likely due to the large number of parameters and relatively small sample sizes) and hence difficult to maximize. The parameter estimates are not entirely reliable, and may correspond to a local maximum. Turning our attention to the likelihood, when we compare the results in Table 2.5 with those in Table 2.1, we see that substantial decreases occur for Patient 2 in particular. Thus, we might surmise that this 3-state model is more appropriate than Albert's model.

It would be a mistake, however, to use the χ^2_5 distribution to gauge the extremity of the LRT statistic. The test comparing models with differing numbers of hidden states amounts to the hypothesis that some of the transition probabilities are zero. Thus, this test is a boundary problem and hence does not satisfy the conditions required for the usual results on LRTs. Moreover, we cannot assume that methods such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) provide consistent estimates of the number of hidden states. Formal hypothesis testing and estimation of the number of hidden states is still an open problem.

Table 2.5: Parameter estimates and standard errors for 3-state model

		Patient 1		Patient 2		Patient 3	
Parameter	Transformation	Estimate	SE	Estimate	SE	Estimate	SE
μ_0^*	$\log \mu_0$	1.061	0.139	1.372	0.155	2.119	0.110
θ^*	$\log(\theta - 1)$	0.003	0.028	0.671	0.130	0.395	0.068
p_1^*	$\log\left(\frac{p_1}{1-p_1-p_2}\right)$	0.417	0.476	−3.139	0.229	0.184	0.167
p_2^*	$\log\left(\frac{p_2}{1-p_1-p_2}\right)$	−0.536	0.816	4.316	0.315	0.126	0.195
p_3^*	$\log\left(\frac{p_3}{1-p_3-p_4}\right)$	0.173	0.537	−0.142	0.643	−6.825	6.846
p_4^*	$\log\left(\frac{p_4}{1-p_3-p_4}\right)$	−0.370	0.250	−6.376	1.247	2.210	0.349
p_5^*	$\log\left(\frac{p_5}{1-p_5-p_6}\right)$	0.547	0.856	−1.757	1.002	7.741	6.168
p_6^*	$\log\left(\frac{p_6}{1-p_5-p_6}\right)$	−1.525	0.136	7.079	1.051	−1.170	1.401
	$\log \mathcal{L}$	−65.931		−69.448		−63.514	

2.5 Summary

The analyses in this chapter illustrate some of the questions that still remain in our understanding and application of HMMs.

We have seen the difficulties involved in comparing the fit of models with differing numbers of hidden states in Section 2.4.5. Likewise, determining the number of hidden states requires non-standard results. We address this issue in Chapter 3, where we develop a method for consistently estimating the number of hidden states in a single, stationary HMM.

The analyses in this chapter also reveal the challenges of selecting an appropriate HMM for a given data set. Tools for examining the fit of both the conditional model for the observed data and the model for the hidden process are needed. We discuss the problem of assessing the goodness-of-fit of stationary HMMs in Chapter 4.

One important common element of our analyses is the relatively large standard errors associated with the estimates of the parameters of the hidden process. Assuming the same model for each patient is one means of reducing this uncertainty. However, the behaviour of lesion counts in MS patients is often highly variable. It thus seems more reasonable to allow at least some model parameters to vary across patients. Random effects are a useful means of capturing between-patient differences while still borrowing strength across patients. Their incorporation in HMMs is discussed in the Chapters 5 and 6.

Chapter 3

Estimating the Number of States of a HMM

Consider the case where we have a single, stationary HMM, with observed data $\{Y_t\}_{t=1}^n$ and hidden states $\{Z_t\}_{t=1}^n$. We will assume that Z_t takes values in the set $\{1, \ldots, K^0\}$, and will denote the stationary probabilities by $\{\pi_k^0\}$ and the transition probabilities by $\{P_{k\ell}^0\}$, $k, \ell = 1, \ldots, K^0$. We assume that $P(Y_t \leq y | Z_t) = H(y; \theta_{Z_t}^0, \phi^0)$, where $(\theta_{Z_t}^0, \phi^0) \in \Theta$.

As discussed in Section 2.3, estimation of the model parameters in the case where K^0 is known has already been studied extensively. However, the problem of consistently estimating K^0 has not yet been satisfactorily resolved. Maximum likelihood estimation cannot be used because the likelihood is non-decreasing in K^0. Most authors applying HMMs, including Leroux & Puterman (1992), Hughes & Guttorp (1994), Albert *et al.* (1994), and Wang & Puterman (1999), simply use the AIC or BIC, but these methods have not been justified in the context of HMMs (MacDonald & Zucchini 1997).

Several authors have attempted to address this problem using penalized likelihood methods. Included in this group are Baras & Finesso (1992), who develop a consistent estimator of K^0 when the observed process takes on only finitely many values. Rydén (1995) relaxes this assumption, but at the expense of consistency. He shows that a class of penalized likelihood estimators provides, in the limit, an upper bound on K^0. Dortet–Bernadet (2001) proves that Rydén's method in fact leads to a consistent estimator of K^0, but under fairly restrictive conditions: he assumes the existence of a known, non-zero lower bound on the transition probabilities. His method of proof appears to require only that this bound apply to the stationary transition probabilities, so perhaps this weaker assumption would suffice.

Other authors have used an information-theoretic approach to estimate K^0. Kieffer (1993) proposes a method involving maximum likelihood codes to find a consistent estimator

of K^0. Liu & Narayan (1994) also give a consistent estimator by describing the observed data in terms of a uniquely decodable code. However, both of these estimators rely on the assumption that the observed process takes on only finitely many values.

Poskitt & Chung (1996) assume a specific form for the HMM, namely that $Y_t = Z_t + \epsilon_t$, where $\{Z_t\}$ is a finite-state Markov chain, and $\{\epsilon_t\}$ is a white noise process. Under these conditions, they suggest an algorithm based on least-squares type calculations that provides an efficient means of consistently estimating K^0.

Robert, Rydén & Titterington (2000) use reversible jump Markov chain Monte Carlo techniques to estimate K^0 in a Bayesian setting. It appears, however, that no frequentist method currently exists for consistently estimating K^0 in the general setting where $\{Y_t\}$ is a stationary, identifiable HMM. In this paper, we approach this problem by extending the ideas of Chen & Kalbfleisch (1996), henceforth called CK. These authors develop a penalized minimum-distance method that gives, under certain conditions, a consistent estimate of the number of components in a finite mixture model. We will use the fact that the marginal distribution of Y_t is also a finite mixture in order to show that a variation of CK's method is applicable to stationary HMMs as well.

3.1 Notation

Under our assumptions, $Y_1, \ldots, Y_n$ are identically distributed with common distribution function

$$F_0(y) \equiv F(y, G_0) = \sum_{k=1}^{K^0} \pi_k^0 H\left(y; \theta_k^0, \phi^0\right) = \int H\left(y; \theta, \phi\right) dG_0\left(\theta, \phi\right), \tag{3.1}$$

where the mixing distribution G_0 is defined by

$$G_0(\theta, \phi) = \sum_{k=1}^{K^0} \pi_k^0 I\left(\theta_k^0 \leq \theta, \phi^0 \leq \phi\right).$$

For ease of exposition, we will assume that the $\{\theta_k^0\}$ and ϕ^0 are scalars so that $\theta_k^0 \leq \theta$ and $\phi^0 \leq \phi$ have the usual interpretations. However, the theory we present easily extends to the more general setting where these parameters are multidimensional. In addition, we treat the pairs (θ_k^0, ϕ^0) as the support points of G_0, even though ϕ^0 is common across states and so would normally be excluded from the definition of the mixing distribution. The treatment of ϕ^0 in this manner will facilitate our discussion of both identifiability and the consistency of the estimator of this parameter.

Similarly, using the notation $\mathbf{y}_1^m = (y_1, \ldots, y_m)$ and $\theta_1^m = (\theta_1, \ldots, \theta_m)$, we will express

the $m-$dimensional distributions of (Y_t) as

$$F_0^m(\mathbf{y}_1^m) \equiv F^m(\mathbf{y}_1^m, G_0^m) = \sum_{z_1=1}^{K^0} \cdots \sum_{z_m=1}^{K^0} \left\{ \prod_{t=1}^{m} H\left(y_t; \theta_{z_t}^0, \phi^0\right) \right\} \pi_{z_1}^0 P_{z_1,z_2}^0 \cdots P_{z_{m-1},z_m}^0 \tag{3.2}$$

with

$$G_0^m(\theta_1^m, \phi) = \sum_{z_1=1}^{K^0} \cdots \sum_{z_m=1}^{K^0} \pi_{z_1}^0 P_{z_1,z_2}^0 \cdots P_{z_{m-1},z_m}^0 I\left(\theta_{z_1}^0 \leq \theta_1, \ldots, \theta_{z_m}^0 \leq \theta_m, \phi^0 \leq \phi\right)$$

3.2 Identifiability

Before presenting the proposed method, we address the issue of model identifiability. First, we need to define K^0 more carefully. We have stated above that K^0 is the number of hidden states. This value is not, in general, equal to the number of values of $\theta_1^0, \ldots, \theta_{K^0}^0$, since these values may not be distinct. Indeed, we can always construct a HMM with $K^0 + 1$ hidden states and distribution F_0^k by choosing an additional state with $\theta_{K^0+1}^0 \in \{\theta_1^0, \ldots, \theta_{K^0}^0\}$ and an appropriate lumpable underlying Markov chain. (For a discussion of lumpability, see White, Mahony & Brushe 2000.) For this reason, we define K^0, the order of the HMM, as *the minimum number of hidden states such that* $\{Y_t\}$ *is a HMM.* We will denote the number of distinct values of the $\{\theta_k^0\}$ by K', where necessarily $K' \leq K^0$.

We now state the mild regularity conditions that we assume.

Condition 1. The transition probability matrix of $\{Z_t\}$ is irreducible and aperiodic.

Condition 2. The parameter space Θ is compact.

Condition 3. $H(y; \theta, \phi)$ is continuous in θ and ϕ.

Condition 4. Given $\epsilon > 0$, there exists $A > 0$ such that for all $(\theta, \phi) \in \Theta$, $H(A; \theta, \phi) - H(-A; \theta, \phi) \geq 1 - \epsilon$.

Condition 5. The family of finite mixtures of $\{H(y; \theta, \phi)\}$ is identifiable, i.e.,

$$F(y, G_1) = F(y, G_2) \quad \Rightarrow \quad G_1 = G_2.$$

Condition 6. Either we know that $\{\theta_k^0\}$ are distinct, or we know an upper bound, M, on the number of hidden states.

REMARK. Conditions 1 and 2 are also assumed by Dortet–Bernadet (2001). Condition 1 implies that the stationary distribution of $\{Z_t\}$ is unique, and that $\pi_k^0 > 0$ for all k. Conditions 3 and 4 are satisfied by commonly used distributions including the normal, Poisson, exponential, binomial, and gamma distributions. Condition 5 is also assumed by Leroux (1992a), Rydén (1995), and Dortet–Bernadet (2001), and is satisfied by many common distributions. Prakasa Rao (1992) provides a good discussion. Condition 6 is weaker than that assumed by Rydén (1995) and Dortet–Bernadet (2001), who postulate a known upper bound on the number of hidden states regardless of the distinctness of $\{\theta_k^0\}$.

3.2.1 Parameter Identifiability

To obtain well-behaved parameter estimates – using either the maximum likelihood or penalized minimum-distance method – model identifiability is required. In the usual case where the values of $\{\theta_k^0\}$ are distinct and Condition 5 is satisfied, the model parameters are identifiable up to permutations of the labels of the hidden states (Wang & Puterman 1999). Petrie (1969) discusses identifiability in the case where Y_t takes on only a finite set of values.

However, we are unaware of the existence of sufficient conditions for parameter identifiability when $\{Y_t\}$ is a general HMM with possibly non-distinct values of $\{\theta_k^0\}$. Determining such conditions appears to be quite a difficult problem. Following Wang & Puterman (1999), we may use the series of $m-$dimensional distributions, $m \geq 1$, to obtain a sequence of equations involving the model parameters. For example, let $\{\theta_{(q)}^0\}$ be the set of K' distinct values among $\{\theta_k^0\}$, with $\theta_{(1)}^0 < \cdots < \theta_{(K')}^0$. Let $S_q = \{k : \theta_k^0 = \theta_{(q)}^0\}$. The one-dimensional distributions of $\{Y_t\}$ are given by

$$F(y, G_0) = \sum_{k=1}^{K^0} \pi_k^0 H\left(y; \theta_k^0, \phi^0\right) = \sum_{q=1}^{K'} H\left(y; \theta_{(q)}^0, \phi^0\right) \sum_{k \in S_q} \pi_k^0.$$

Condition 5 allows us to identify G_0 from this equation. By Condition 1, all support points of G_0 have positive mass, and hence $\{\theta_{(q)}^0\}$, ϕ^0, and $\{\sum_{k \in S_q} \pi_k^0\}$ may also be identified.

Similarly, the two-dimensional distributions of (Y_t) are given by

$$\begin{aligned} F^2(\mathbf{y}_1^2, G_0^2) &= \sum_{k=1}^{K^0} \sum_{\ell=1}^{K^0} \pi_k^0 P_{k\ell}^0 H\left(y_1; \theta_k^0, \phi^0\right) H\left(y_2; \theta_\ell^0, \phi^0\right) \\ &= \sum_{q=1}^{K'} \sum_{r=1}^{K'} H\left(y_1; \theta_{(q)}^0, \phi^0\right) H\left(y_2; \theta_{(r)}^0, \phi^0\right) \sum_{k \in S_q} \pi_k^0 \sum_{\ell \in S_r} P_{k\ell}^0. \end{aligned}$$

Teicher (1967) shows that mixtures of products of distributions from a given family are identifiable if this family satisfies Condition 5. Using this result, we may identify G_0^2, and

hence

$$\left\{ \sum_{k \in S_q} \pi_k^0 \sum_{\ell \in S_r} P_{k\ell}^0 \right\},$$

from this equation. In particular, if $\{\theta_k^0\}$ are distinct, we see that the two-dimensional distributions are sufficient to allow the identification of the parameters up to permutations of the labels of the hidden states.

In the case where $\{\theta_k^0\}$ are not distinct, parameter identifiability can be explored by applying Teicher's result in this manner to the higher-dimensional distributions. However, the equations obtained in this fashion are highly non-linear – in part due to the complicated relationship between $\{P_{k\ell}^0\}$ and $\{\pi_k^0\}$ – and difficult to analyze. Hence, at this time, we lack a means of assessing, in general, whether the parameters of a given model are identifiable. In addition, some of these equations are redundant, so it is unclear even how to specify the minimum number of dimensions that must be considered in order to determine parameter identifiability.

Fortunately, Rydén (1995) shows that the finite-dimensional distributions of $\{Y_t\}$ are determined by the $2(K^0 - K' + 1)-$dimensional distribution when Condition 5 is satisfied. Thus, under the assumption that K^0 is minimal, if M is an upper bound on the number of hidden states, then $2M$ is an upper bound on the required number of dimensions. Letting $\psi = (\theta_1, \ldots, \theta_K, \phi, P_{11}, P_{12}, \ldots, P_{KK}, \pi_1, \ldots, \pi_K)$ as before, we will use Rydén's equivalence relation: $\tilde{\psi}^0$ denotes the equivalence class of ψ^0, with $\psi \in \tilde{\psi}^0$ if and only if ψ induces the same law for $\{Y_t\}$ as ψ^0.

In conclusion, if Condition 5 holds, then $\tilde{\psi}^0$ is determined by the $2M-$dimensional distributions, where we take $M = 1$ if the values of $\{\theta_k^0\}$ are distinct. We will use this result in the development of the penalized minimum-distance method for HMMs.

3.2.2 Sufficient Conditions for CK's Identifiability Criterion

When K^0 is minimal, Condition 5 is the standard notion of identifiability of a mixture model. CK, however, assume an identifiability criterion of a different form. In particular, let d be a distance measure on the space of probability distributions, and let G_n be a sequence of one-dimensional mixing distributions. CK assume that

$$\lim_{n \to \infty} d\{F(y, G_n), F(y, G_0)\} = 0 \quad \Rightarrow \quad \lim_{n \to \infty} d(G_n, G_0) = 0.$$

This criterion should in fact read: for distance functions d_1 and d_2, G_n converges to G_0 weakly when $d_2(G_n, G_0) \to 0$ as $n \to \infty$, and

$$\lim_{n \to \infty} d_1\{F(y, G_n), F(y, G_0)\} = 0 \quad \Rightarrow \lim_{n \to \infty} d_2(G_n, G_0) = 0, \tag{3.3}$$

(Chen, personal communication).

It is of interest to develop sufficient conditions for CK's criterion, both as an extension to CK's work, and for use in our procedure to estimate the parameters of a HMM. The following lemma, which will be proved in Appendix B.1, states that, for specific choices of d_1 and d_2, this criterion is satisfied when Conditions 2–5 hold. Normally, it will be easier to verify these conditions than the original CK criterion.

In particular, we assume that d_2 corresponds to weak convergence, and that $d_1 \equiv d_{KS}$ is the Kolmogorov–Smirnov distance, i.e., for distribution functions F_1 and F_2,

$$d_{KS}(F_1, F_2) = \sup_y |F_1(y) - F_2(y)| .$$

Lemma 3.1 *Let (G_n^m) be a sequence of m-dimensional mixing distributions with associated parameters $\{(\theta_k^n, \phi^n)\} \in \Theta$. Then, under Conditions 2–5, if $d_{KS}\{F^m(\mathbf{y}_1^m, G_n^m), F^m(\mathbf{y}_1^m, G_0^m)\} = o(1)$, then G_n^m converges weakly to G_0^m.*

3.3 Parameter Estimation

Let $\{c_n\}$ be a sequence of positive constants with $c_n = o(1)$, and let $\bar{F}_n$ be the empirical distribution function of Y_t. If $\{Y_t\}$ are independent observations from a finite mixture model, F_0, then CK estimate the parameters of F_0 (including the number of components) by minimizing the penalized distance function

$$D\left(\bar{F}_n, F\right) = d_1\left(\bar{F}_n, F\right) - c_n \sum_{k=1}^{K} \log \pi_k$$

over all F, where F is a finite mixing distribution with K components and mixing probabilities $\{\pi_k\}$, and d_1 is any distance function satisfying (3.3). The authors prove the consistency of the parameter estimates obtained in this fashion.

The novelty of this approach is that the penalty term is a function of $\sum_{k=1}^{K} \log \pi_k$. Models with large values of K are penalized since the requirement that $\pi_1 + \cdots + \pi_K = 1$ forces π_k to 0 for some values of k as $K \to \infty$. In addition, models for which some states have small values of π_k are penalized. In these two ways, the estimated number of components is indirectly controlled. In contrast, most other penalized methods, including the AIC and BIC, attempt to control K directly.

From the discussion in Section 3.2, it is clear that CK's estimation procedure, which is based on one-dimensional distributions, is not sufficient to estimate all the parameters of

a HMM. However, by modifying the method to incorporate multiple dimensions, we may obtain a procedure that is appropriate to our problem.

In particular, for $n \geq m$, we consider the $m-$dimensional process $\{\mathbf{Y}_t^{t+m-1}\}_{t=1}^{n-m+1}$, where $\mathbf{Y}_t^{t+m-1} = (Y_t, \ldots, Y_{t+m-1})$. We then define the penalized distance based on the distribution of this process as

$$D\left(\bar{F}_n^m, F^m\right) = d_1\left(\bar{F}_n^m, F^m\right) - c_n \sum_{k=1}^{K} \log \pi_k, \tag{3.4}$$

where $\bar{F}_n^m$ is the $m-$dimensional empirical distribution function,

$$\bar{F}_n^m(\mathbf{y}_1^m) = \frac{\sum_{t=1}^{n-m+1} I\left(Y_t \leq y_1, \ldots, Y_{t+m-1} \leq y_m\right)}{n-m+1}. \tag{3.5}$$

The dimension m should be chosen based on identifiability considerations, as discussed in Section 3.2. In addition, it is desirable to choose m as small as possible to minimize the computational burden. Thus, we will use $m = 2M$.

CK's proof of the consistency of the parameter estimates in the case of finite mixture models does not require the independence of the observed data. In fact, this proof depends only on their identifiability criterion (see Section 3.2.2) and the assumption that $d_1(\bar{F}_n, F_0) = o(c_n)$ a.s. We will show that, if d_1 is chosen such that $d_1(\bar{F}_n^m, F_0^m) = o(c_n)$ a.s. and Conditions 1–6 are satisfied, then the theorems developed by CK also hold when the underlying model is a HMM and the parameter estimates minimize the penalized distance function (3.4). Hence, we will obtain consistent estimators of the model parameters, including K^0.

Before stating our first theorem we require the definition of the concept of an α-mixing multidimensional process (see, for instance, Ould-Saïd 1994).

Definition 3.1 *A stationary sequence $(\mathbf{V}_t)$ of $m-$dimensional vectors is $\alpha-$mixing or strong mixing if for any s,*

$$\alpha_\ell = \sup_{A \in \mathcal{F}_1^s, B \in \mathcal{F}_{s+\ell}^\infty} |\mathrm{P}(AB) - \mathrm{P}(A)\mathrm{P}(B)| \to 0 \quad as \quad \ell \to \infty$$

where $\mathcal{F}_a^b = \sigma(\mathbf{V}_t, a \leq t \leq b)$. The α_ℓ's are called the mixing coefficients.

Theorem 3.1 *For a stationary HMM,*

$$d_{KS}\left(\bar{F}_n^m, F_0^m\right) = O\left(n^{-\frac{1}{2}}\sqrt{\log\log n}\right) \quad a.s.$$

Proof. Ould-Saïd (1994) proves that for a stationary, $m-$dimensional, $\alpha-$mixing process with mixing coefficients $\alpha_\ell = O(\ell^{-\nu})$ for some $\nu > 2m+1$,

$$\limsup \left[\left(\frac{n}{2 \log\log n} \right)^{1/2} \sup_{\mathbf{y}_1^m} \left| \bar{F}_n^m (\mathbf{y}_1^m) - F_0^m (\mathbf{y}_1^m) \right| \right] = 1 \quad \text{a.s.}$$

Thus to prove that the empirical distribution of a HMM converges at this rate, it is sufficient to show that the HMM satisfies the condition $\alpha_\ell = O(\ell^{-\nu})$ for $\nu > 4M+1$.

First, the mixing coefficients α_ℓ^z of the Markov chain $\{Z_t\}$ satisfy the inequality $\alpha_\ell^z \leq c\rho^\ell$, where c is a finite, positive constant, and $0 < \rho < 1$ (see, for instance, Doukhan 1994). Now, without loss of generality, take $\ell > m$. Define $\mathcal{F}_a^b = \sigma(Y_a, \ldots, Y_{b+m-1})$. Following Lindgren (1978), for stationary HMMs we have that

$$\begin{aligned}
\alpha_\ell &= \sup_{A \in \mathcal{F}_1^s, B \in \mathcal{F}_{s+\ell}^\infty} |\mathrm{P}(AB) - \mathrm{P}(A)\mathrm{P}(B)| \\
&= \sup_{A \in \mathcal{F}_1^s, B \in \mathcal{F}_{s+\ell}^\infty} \left| \mathrm{E}\left\{ \mathrm{P}\left(AB | Z_1^{s+m-1}, Z_{s+\ell}^\infty \right) \right\} - \mathrm{E}\left\{ \mathrm{P}\left(A | Z_1^{s+m-1} \right) \right\} \mathrm{E}\left\{ \mathrm{P}\left(B | Z_{s+\ell}^\infty \right) \right\} \right| \\
&= \sup_{A \in \mathcal{F}_1^s, B \in \mathcal{F}_{s+\ell}^\infty} \left| \mathrm{E}\left\{ \mathrm{P}\left(A | Z_1^{s+m-1} \right) \mathrm{P}\left(B | Z_{s+\ell}^\infty \right) \right\} - \mathrm{E}\left\{ \mathrm{P}\left(A | Z_1^{s+m-1} \right) \right\} \mathrm{E}\left\{ \mathrm{P}\left(B | Z_{s+\ell}^\infty \right) \right\} \right|
\end{aligned}$$

Since $\mathrm{P}(A|Z_1^{s+m-1})$ and $\mathrm{P}(B|Z_{s+\ell}^\infty)$ are bounded and measurable with respect to $\sigma(Z_1^{s+m-1})$ and $\sigma(Z_{s+\ell}^\infty)$, respectively, we may apply Theorem 17.2.1 of Ibragimov & Linnik (1971) to obtain

$$\alpha_\ell \leq 4\alpha_{\ell-m+1}^z \leq 4c\rho^{\ell-m+1}.$$

We have thus proved that $\alpha_\ell = o(\ell^{-\nu})$ for $\nu < \infty$. □

COROLLARY. *Let $\hat{F}^m$ be the function that minimizes* (3.4), *where the minimization is over all parameters, including K. Under Conditions* 1–6, *if we choose $d_1 \equiv d_{KS}$, then $d_{KS}(\hat{F}^m, F_0^m) \to 0$ a.s. and $\hat{G}^m$ converges weakly to G_0^m a.s. Furthermore, $\hat{\phi}^0 \to \phi^0$ a.s., and $\hat{\theta}_k^0 \to \theta_*$ a.s., where θ_* is one of the support points of G_0^m.*

Proof. In light of Theorem 3.1 and Lemma 3.1, the technique of proof used in Theorems 1 and 3 of CK also applies in the HMM case. Although the class of mixture models considered by CK does not incorporate parameters that are common to all components (i.e., ϕ^0), their Theorem 3 is also sufficient to show that $\hat{\phi} \to \phi^0$ a.s. □

We now investigate the asymptotic behaviour of $\hat{K}^0$ using a method inspired by the work of Leroux (1992b) in the context of penalized maximum likelihood estimation of mixture

models. In particular, we study properties of the estimated $m-$dimensional distribution when we fix the number of hidden states at some known value K, while the other parameters are estimated by minimizing (3.4). We denote this distribution by $\hat{F}_K^m$, and the associated parameter estimates by $\hat{\psi}$, so that

$$\hat{F}_K^m(\mathbf{y}_1^m) = \sum_{z_1=1}^{K} \cdots \sum_{z_m=1}^{K} H\left(y_1; \hat{\theta}_{z_1}, \hat{\phi}\right) \cdots H\left(y_m; \hat{\theta}_{z_m}, \hat{\phi}\right) \hat{\pi}_{z_1} \hat{P}_{z_1,z_2} \cdots \hat{P}_{z_{m-1},z_m}.$$

We will need the following two technical results.

Lemma 3.2 *If we know the value of K^0 and estimate the remaining parameters by minimizing (3.4), then we have that*

$$d_{KS}\left(\hat{F}_{K^0}^m, \bar{F}_n^m\right) = O(c_n) \quad a.s. \quad \text{and} \quad \liminf \sum_{k=1}^{K^0} \log \hat{\pi}_k^0 > -\infty \quad a.s.$$

Proof. This lemma is similar to Theorems 1 and 2 of CK, and can be proved in the same way. □

Lemma 3.3 *If K is chosen such that $K < K^0$ and the remaining parameters estimated by minimizing (3.4), then*

$$\liminf d_{KS}(\hat{F}_K^m, \bar{F}_n^m) > 0 \quad a.s.$$

Proof. By the triangle inequality,

$$d_{KS}\left(\hat{F}_K^m, \bar{F}_n^m\right) \geq d_{KS}\left(\hat{F}_K^m, F_0^m\right) - d_{KS}\left(\bar{F}_n^m, F_0^m\right).$$

By Theorem 3.1, $d_{KS}(\bar{F}_n^m, F_0^m) = o(1)$ a.s. For this reason, it is enough to show that $\liminf d_{KS}(\hat{F}_K^m, F_0^m) > 0$.

Choose a realization of the sequence $\{\hat{F}_K^m\}$, indexed by n. Towards a contradiction, assume that $\liminf d_{KS}(\hat{F}_K^m, F_0^m) = 0$ for this realization. Then there exists a subsequence such that $\lim d_{KS}(\hat{F}_K^m, F_0^m) = 0$. For this subsequence, $\hat{F}_K^m$ converges pointwise to F_0^m, and there exists a further subsequence such that each parameter estimate converges to some limiting value, ψ. Since H is continuous in θ and ϕ, we have that

$$\lim \hat{F}_K^m(\mathbf{y}_1^m) = \sum_{z_1=1}^{K} \cdots \sum_{z_m=1}^{K} H(y_1; \theta_{z_1}, \phi) \cdots H(y_m; \theta_{z_m}, \phi) \pi_{z_1} P_{z_1,z_2} \cdots P_{z_{m-1},z_m}.$$

Let $S^* \equiv \{\theta_k\}$. From our discussion of identifiability in Section 3.2, the set of distinct elements of S^* must be equal to $\{\theta^0_{(k)}\}$, and we must have that $\phi = \phi^0$. So, defining $S^*_k = \{i : \theta_i = \theta^0_{(k)}\}$, $k = 1, \ldots, K'$,

$$\begin{aligned} \lim \hat{F}^m_K(\mathbf{y}^m_1) &= \sum_{z_1=1}^{K'} \cdots \sum_{z_m=1}^{K'} H(y_1; \theta^0_{(z_1)}, \phi^0) \cdots H\left(y_k; \theta^0_{(z_m)}, \phi^0\right) \\ &\quad \times \sum_{q_1 \in S^*_{z_1}} \cdots \sum_{q_m \in S^*_{z_m}} \pi_{q_1} P_{q_1,q_2} \cdots P_{q_{m-1},q_m}. \end{aligned}$$

This limiting distribution is the $m-$dimensional distribution of a HMM with K hidden states. But, now it is clear that $\lim \hat{F}^m_K \neq F^m_0$, since K^0 was defined as minimal. This contradicts our hypothesis that $\lim d_{KS}(\hat{F}^m_K, F^m_0) = 0$. □

Theorem 3.2 *Assume that $c_n = o(1)$ and Conditions 1–6 are satisfied. Then* $\liminf \hat{K}^0 \geq K^0$ *a.s. If $\{\theta^0_k\}$ are distinct, then we also have that* $\limsup \hat{K}^0 \leq K^0$ *a.s.*

Proof. If $\{\theta^0_k\}$ are distinct, then Theorem 4 of CK applies. Hence, $\hat{K}^0 \to K^0$ a.s. If $\{\theta^0_k\}$ are not distinct, we require a different technique of proof. It is clear that $\liminf \hat{K}^0 \geq K'$ a.s. However, some work is required to prove that $\liminf \hat{K}^0 \geq K^0$ a.s. The key idea is that if the limiting value of $\inf_{F^m} D(\bar{F}^m_n, F^m)$ is less when we fix $K = K^0$ than when we fix $K < K^0$, we will know that $D(\bar{F}^m_n, F^m)$ is minimized, in the limit, by a HMM with at least K^0 hidden states.

By Lemmas 3.2 and 3.3, for $K' \leq K < K^0$, we have that

$$\frac{d_{KS}(\hat{F}^m_{K^0}, \bar{F}^m_n) - d_{KS}(\hat{F}^m_K, \bar{F}^m_n)}{c_n} \to -\infty \quad \text{a.s.}$$

Also by Lemma 3.2, $\sum_{k=1}^{K^0} \log \hat{\pi}^0_k$ is bounded away from $-\infty$ a.s. Thus, with probability 1, for all n sufficiently large,

$$\begin{aligned} \frac{d_{KS}(\hat{F}^m_{K^0}, \bar{F}^m_n) - d_{KS}(\hat{F}^m_K, \bar{F}^m_n)}{c_n} &< \sum_{k=1}^{K^0} \log \hat{\pi}^0_k, \\ d_{KS}(\hat{F}^m_{K^0}, \bar{F}^m_n) - d_{KS}(\hat{F}^m_K, \bar{F}^m_n) &< c_n \left(\sum_{k=1}^{K^0} \log \hat{\pi}^0_k - \sum_{k=1}^{K} \log \hat{\pi}_k \right), \\ D(\bar{F}^m_n, \hat{F}^m_{K^0}) &< D(\bar{F}^m_n, \hat{F}^m_K). \end{aligned}$$

In other words, for n sufficiently large, the estimated distribution function will have at least K^0 hidden states with probability 1. □

Our final theorem demonstrates the consistency of $\{\hat{P}_{k\ell}\}$ in the quotient topology generated by our equivalence relation. In particular, if $\mathcal{A}$ is an open set containing $\tilde{\psi}^0$, then

$\hat{\psi}^0 \in \mathcal{A}$ for large n a.s. We will use the notation $\hat{\psi}^0 \to \tilde{\psi}^0$ a.s. This notion was also used by Leroux (1992a) and Rydén (1995).

Theorem 3.3 *Assume that $c_n = o(1)$ and Conditions 1–6 are satisfied. Then $\hat{\psi}^0 \to \tilde{\psi}^0$ a.s. If $\{\theta_k^0\}$ are distinct, then we also have that $\lim \hat{P}_{k\ell}^0 \to P_{k\ell}^0$ a.s. (ignoring possible permutations of the labels of the hidden states).*

Proof. If $\{\theta_k^0\}$ are distinct, the model parameters are identifiable up to permutations of the labels of the hidden states. Thus, by Theorem 3.2 and the corollary to Theorem 3.1, $\hat{P}_{k\ell}^0 \to P_{k\ell}^0$ a.s. for all k, ℓ (ignoring possible permutations).

Otherwise, from the corollary to Theorem 3.1, we know that $\hat{F}^m$ converges pointwise to F_0^m a.s. In addition, from the corollary to CK's Theorem 2, we know that $\limsup \hat{K}^0 < \infty$ a.s. The remainder of the proof will be restricted to the event of probability 1 where these limits hold.

Consider a realization of the sequence $\{\hat{K}^0\}$ indexed by n. Since $\hat{K}^0$ takes on only integer values, we can find neighbourhoods around each limit point of $\{\hat{K}^0\}$ such that there is only one limit point in each neighbourhood, and all points of $\{\hat{K}^0\}$ are in at least one neighbourhood. Since $\limsup \hat{K}^0 < \infty$, there are only a finite number of these neighbourhoods. Let $\{n_j^i\}$ be the subsequence defined by the set of points in the ith neighbourhood, and let K^i be the limit point of $\{\hat{K}^0\}$ associated with this neighbourhood. For this subsequence, and for all j sufficiently large, $\hat{K}^0 = K^i$ and hence $\hat{F}_{K^i}^m = \hat{F}^m$. We can then interchange the limit and the summation in the expression for $\lim \hat{F}_{K^i}^m$ (as in the proof of Lemma 3.3) and use the fact that $\hat{F}_{K^i}^m = \hat{F}^m \to F_0^m$ to conclude that, for this subsequence, $\hat{\psi}^0 \to \tilde{\psi}^0$. However, since the union of all the subsequences is the original sequence, it is clear that, in fact, $\hat{\psi}^0 \to \tilde{\psi}^0$ a.s. for the original sequence. □

3.4 Application to MS/MRI Data

In this section, under the assumption that a stationary HMM adequately captures the behaviour of the lesion counts observed on relapsing-remitting MS patients, we estimate the number of underlying disease states using the data of Albert *et al.* (1994) described in Section 2.4. For the purposes of illustrating our technique, we assume the same model for each patient. Specifically, we assume that $Y_{it}|Z_{it} \sim \text{Poisson}(\mu_{Z_{it}}^0)$. Here Y_{it} is the number of lesions recorded for patient i at time t, and Z_{it} is the associated disease state.

Table 3.1: Penalized minimum-distances for different numbers of hidden states

Number of states	Estimated Poisson means	Minimum distance
1	4.03	0.1306
2	2.48, 6.25	0.0608
3	2.77, 2.62, 7.10	0.0639
4	2.05, 2.96, 3.53, 7.75	0.0774
5	1.83, 3.21, 3.40, 3.58, 8.35	0.0959

We then use the method of penalized minimum-distance to estimate K^0, the number of states in the hidden process. We are willing to assume that the values of $\{\mu_k^0\}$ are distinct, and thus use the bivariate distributions to compute the distance function. As suggested in CK, we use $c_N = 0.01N^{-1/2}\log N$, where N is the total number of observations across all patients. Using a variety of starting values, we calculate the penalized minimum-distance for $K^0 = 1, \ldots, 5$ using a quasi-Newton minimization routine (Nash 1979). These distances are displayed in Table 3.1. Note that the overall penalized minimum-distance occurs at $K^0 = 2$, which becomes our estimate for the number of hidden states. These two states may correspond to relapse and remission, which is consistent with qualitative observations of the behaviour of this disease.

The estimates of the parameters of the hidden process were

$$\hat{\pi^0} = [0.594, 0.406], \qquad \hat{P^0} = \begin{bmatrix} 0.619 & 0.381 \\ 0.558 & 0.442 \end{bmatrix}$$

indicating that, on average, patients spent a higher proportion of their time in remission (59.4%) than in relapse (40.6%).

3.5 Performance of the Penalized Minimum-Distance Method

The appropriateness of the penalized minimum-distance estimator for finite sample sizes is another topic of interest. To address this issue, we generated 100 individual time series of two different lengths (30 and 100) from various Poisson HMMs. We then estimated K^0 using three different methods: the AIC, BIC, and penalized minimum-distance method.

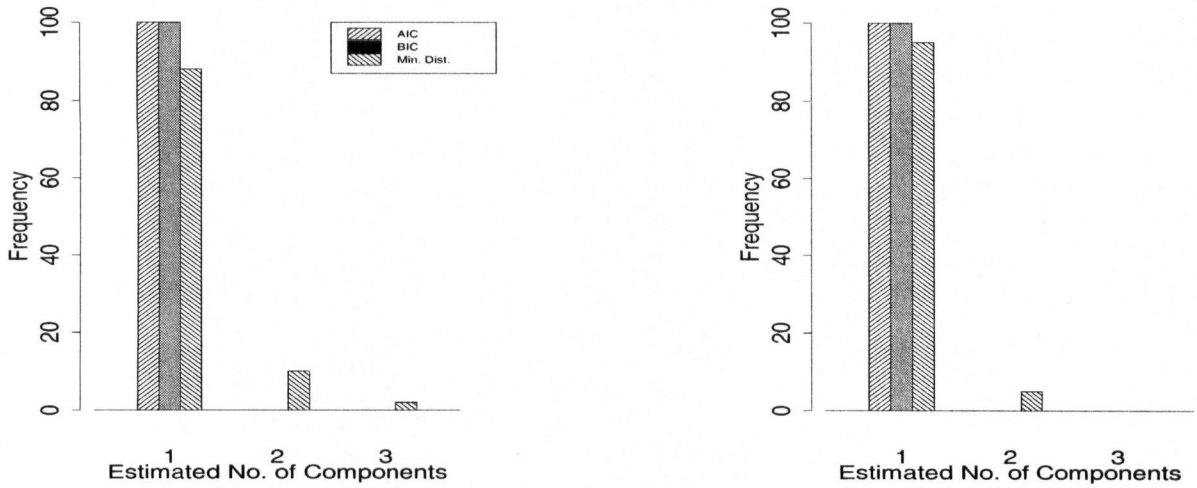
Table 3.2: Parameter values used in the simulation study

K^0	Poisson Means	Transition Probabilities
1	$\mu = [4]$	$P = [1]$
2	$\mu = \begin{bmatrix} 1 \\ 7 \end{bmatrix}$	$P = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$
	$\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$	$P = \begin{bmatrix} 0.269 & 0.731 \\ 0.119 & 0.881 \end{bmatrix}$
3	$\mu = \begin{bmatrix} 1 \\ 5 \\ 9 \end{bmatrix}$	$P = \begin{bmatrix} 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}$
	$\mu = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$	$P = \begin{bmatrix} 0.140 & 0.231 & 0.629 \\ 0.212 & 0.212 & 0.576 \\ 0.030 & 0.366 & 0.604 \end{bmatrix}$

The simulation was conducted in the spirit of an experimental design with four factors: number of components (1, 2, or 3), sample size (30 or 100), separation of components (well-separated or close together), and proportion of time in each state (balanced or unbalanced among states). The sample sizes of 30 and 100 were chosen to reflect the sizes of typical and large MS/MRI data sets, respectively. The parameters selected for each case are given in Table 3.2.

As in the previous section, the models were fit using a quasi-Newton minimization routine (Nash 1979). For the penalized minimum-distance method, we again chose $c_n = 0.01 n^{-1/2} \log n$. To reduce the computational burden, when the true value of K^0 was 1 or 2, we fit only models with 1, 2, and 3 components. For $K^0 = 3$, we fit models with 1, 2, 3, and 4 components. Histograms of the resulting estimated values of K^0 appear in Figures 3.1-3.5. The legend is the same for all plots, and is included only in the first.

Figure 3.1: Distribution of $\hat{K}^0$ when $K^0 = 1$



a) $n = 30$ b) $n = 100$

The histograms show that the penalized minimum-distance method seems to perform well relative to the AIC and BIC, especially when the problem is "hard," i.e., $K^0 = 2$ or $K^0 = 3$, the components are not well-separated or the proportion of time in each state is not balanced. For $K^0 = 1$, the method tends to overestimate slightly the number of components.

It is interesting that the performances of the methods do not improve substantially when the sample size is increased from 30 to 100. This may not be surprising in the cases of the AIC and BIC (which have not been proved to be consistent methods of estimation). However, in the case of the penalized minimum-distance method, we might expect to see a greater improvement. This result may be due to the fact that K^0 is equal to the number of non-zero stationary transition probabilities. Since these probabilities are parameters of the unobserved process, large data sets are required for their precise estimation. Consequently, estimating K^0 precisely is also a difficult problem.

3.6 Discussion

One of the practical difficulties with the method of penalized minimum-distance is that the resulting objective function has many local minima at which the algorithm tends to converge. Using a variety of different starting values can help, but this increases the required computational effort, and we can never be certain that we have located the global minimum.

Although, in general, locating the global minimum of the penalized distance function

Figure 3.2: Distribution of $\hat{K}^0$ when $K^0 = 2$, $n = 30$

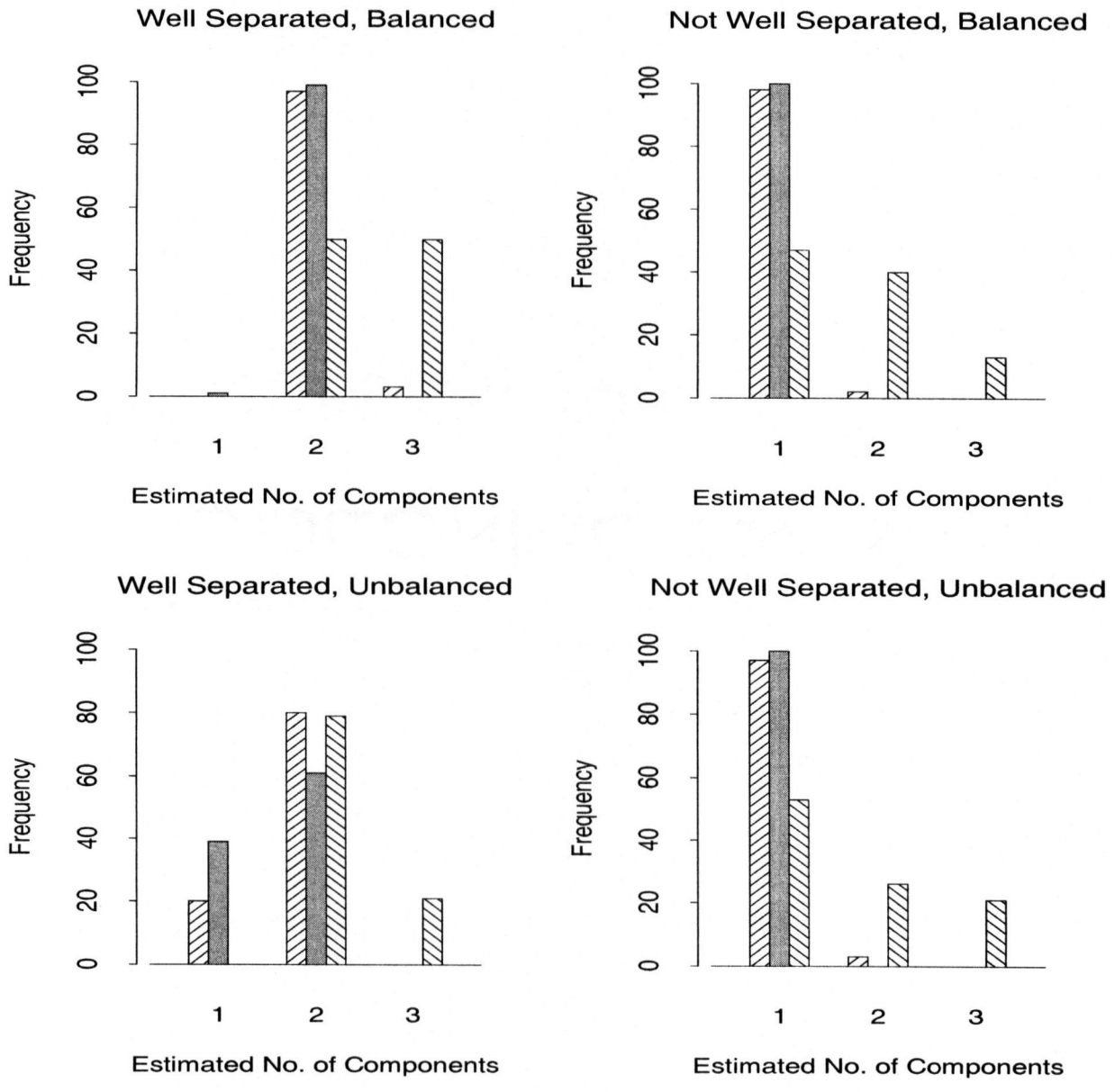


Figure 3.3: Distribution of $\hat{K}^0$ when $K^0 = 2$, $n = 100$

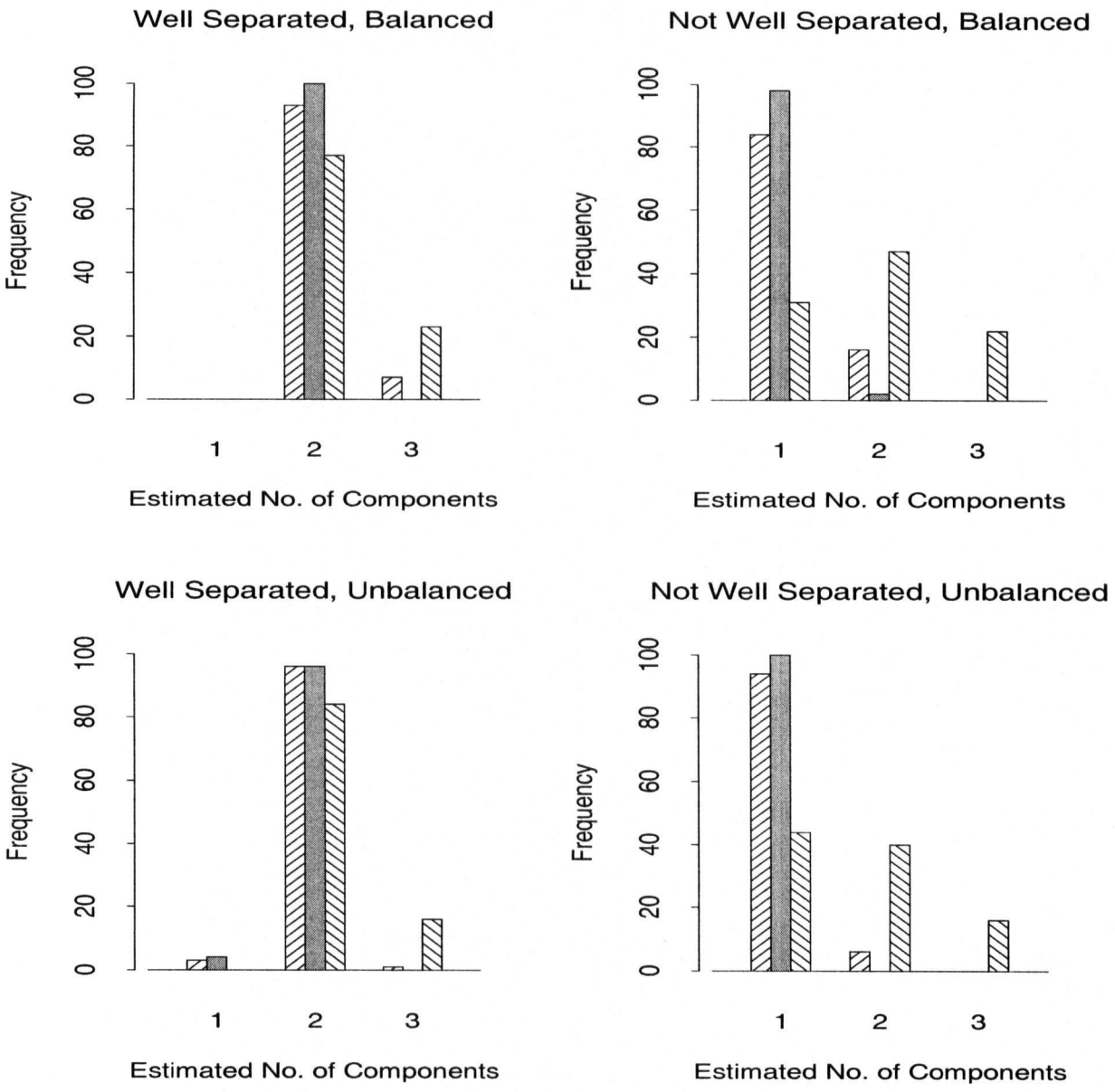


Figure 3.4: Distribution of $\hat{K}^0$ when $K^0 = 3$, $n = 30$

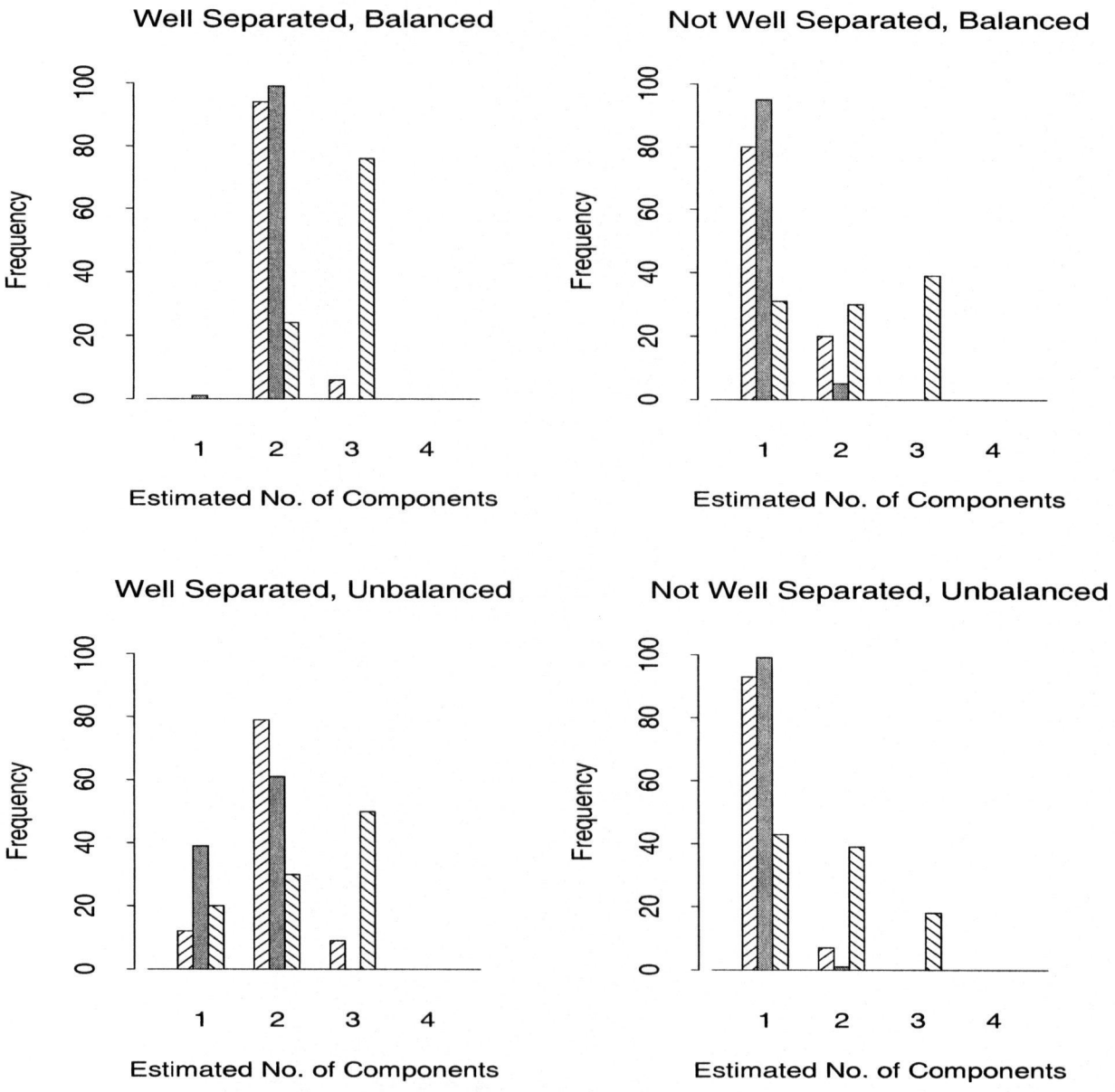
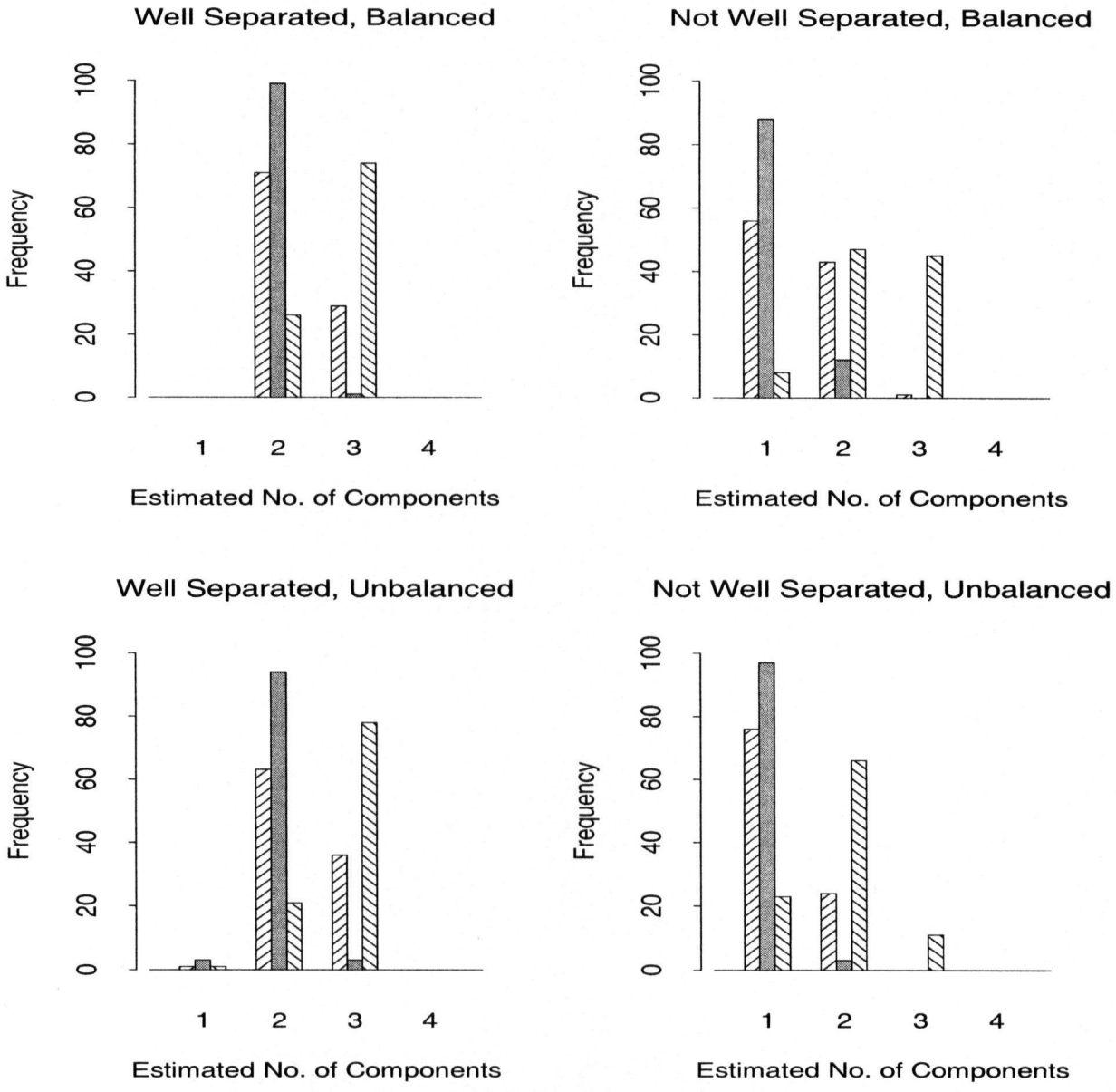


Figure 3.5: Distribution of $\hat{K}^0$ when $K^0 = 3$, $n = 100$



is critical for obtaining good parameter estimates, it may not be necessary for determining the number of hidden states. In the examples we have considered, the local minima found for each value of K^0 were typically both close to one another and reasonably well-separated from the local minima found for other values of K^0.

For this reason, a two-stage estimation procedure may be appropriate. First, K^0 could be estimated using the method of penalized minimum-distance. Then, assuming this value of K^0, the remaining parameters could be estimated using maximum likelihood estimation. The benefit of this approach is that, based on our experience, it seems easier to locate the maximum value of the likelihood function than the minimum value of the distance function. In addition, MLEs have appealing properties, and approximate standard errors are readily available. In contrast, we do not have this information for the estimates obtained using the penalized minimum-distance method. The disadvantage of the two-stage method is that the standard errors for our MLEs do not take into account our uncertainty about K^0. Further exploration of the implications of this estimation procedure is in order.

We may be able to improve the performance of the penalized minimum-distance method through different choices of the distance function, c_n, or the penalty term. For example, our method might also be valid for other distance functions. The Cramér-von Mises distance or Kullback-Leibler information are options (CK), as is the Hellinger distance, which has the advantage of robustness (Hettmansperger & Thomas 2000). These distance functions – and others – may be valid choices in the HMM setting, and may result in improved performance. In addition, as discussed in CK, there are many possibilities for the form of the penalty term. Finally, the choice of c_n is critical. When $c_n = Cn^{-1/2}\log n$, for some positive constant C, the penalty term converges to zero at approximately the same rate as the distance function. This seems reasonable. However, in simulation studies similar to those discussed in Section 3.5, we found that the resulting estimate of K^0 was very sensitive to the value of C. Of the possible values $C = 0.1, 0.01, 0.001$, the optimal value of C, as judged by the frequency with which the correct value of K^0 was predicted, tended to decrease with K^0. In other words, in the cases considered, the (subjective) choice of C appeared to have a considerable influence on the estimate of K^0. Clearly, the form of c_n is a topic requiring further investigation.

Chapter 4

Assessment of Goodness-of-Fit

In this chapter, we will again be working in the setting where we have a single, stationary HMM, and will use the same notation as in Chapter 3.

As with most data analysis problems, it is desirable to find methods for assessing the goodness-of-fit (GOF) of a given HMM. As discussed in Section 2.3, Giudici *et al.* (2000) show that the likelihood ratio test can be used to compare nested stationary HMMs with a common, known value of K^0. However, the comparison of non-nested models, including both HMMs (with possibly unknown values of K^0) and models outside the class of HMMs, is more challenging, and this is the problem we consider here.

Lystig (2001) provides a comprehensive overview of existing literature on GOF techniques for HMMs. Turner *et al.* (1998), working with Poisson HMMs, create a diagnostic plot by overlaying the predicted mean responses at each time point on the observed data. A limitation of this method is its focus on means rather than on distributions: it is not suitable for detecting violations of the Poisson assumption. Zucchini & Guttorp (1991) also look at plots of the predicted mean responses but for binary data. In the case where the observed data take on only a finite number of values, Albert (1991) suggests qualitatively comparing the observed and expected frequencies of each value. However, neither of these methods allows the investigation of deviations from the assumed model for the hidden process. Hughes & Guttorp (1994), working with a non-homogeneous HMM with finite state space, consider the comparison of the observed frequency of each response, y, to

$$\frac{1}{n}\sum_{t=1}^{n}\hat{\mathrm{P}}(Y_t = y)$$

where $\hat{\mathrm{P}}$ is the estimated probability under the fitted HMM. In a similar way, these authors compare the observed and estimated survival functions (i.e. the probability that the observed process is in state r for at least k days), as well as the observed and estimated correlations.

However, since $P(Y_t = y)$ depends on t in this setting, the advantage of averaging over the n observations is unclear. Finally, Lystig (2001) develops a formal test based on the score process for use in the context where there are n responses (from a finite state space) on each of N independent individuals, and N is large.

In addition to providing guidance about choosing among models, it is desirable that a GOF technique will, with high probability, detect a lack of fit as n gets large – when either the marginal distribution or the correlation structure of the observed data is misspecified. However, none of the methods described above has been shown to have this property.

With these goals in mind, we consider an alternative method of assessing GOF in this chapter. Our method is similar to that of Hughes & Guttorp (1994), but we exploit the fact that, in the stationary case, we have identically distributed observations. In particular, we propose a graphical approach to analyzing the fit of the HMM: we plot the estimated cumulative distribution function (cdf), $F(\cdot, \hat{G}_0)$, given by Equation 3.1 against the empirical cdf, $\bar{F}_n(\cdot)$. Recall that the empirical cdf is based solely on the observed data:

$$\bar{F}_n(y) = \frac{\sum_{t=1}^{n} I(Y_t \leq y)}{n}$$

If $\{Y_t\}$ is discrete, we might also consider plotting the estimated probability distribution function (pdf) against the empirical pdf. However, we will restrict our discussion to the general case, and henceforth, the word "distribution" will refer to the cdf.

Under regularity conditions, if we have correctly specified the model, the empirical and estimated distributions will both be consistent estimates of the true distribution, so as n increases, this plot will converge to a 45° line through the origin.

By plotting the univariate distributions in this way, we will be able to assess the fit of the assumed marginal distribution for Y_t, i.e. the mixture distribution given by Equation 3.1. However, this plot provides no information about the fit of the assumed correlation structure. In light of the comment by Hughes & Guttorp (1994) that, at least in their setting, "it is generally not difficult to get a good fit to the empirical marginal probabilities", checking the correlation structure may be of primary interest. Thus, making use of the ideas in Section 3.2, we propose the construction of an additional plot: the estimated bivariate distribution, $F^2(\cdot, \hat{G}_0^2)$ (see Equation 3.2), against the empirical bivariate distribution, $\bar{F}_n^2(\cdot)$ (given by Equation 3.5). If the values of $\{\theta_j^0\}$ are not distinct, we may also wish to make plots of the higher dimensional distributions. Again, we would expect these plots to converge to a straight line if the assumed model is correct. In this way, as n increases, we will be able to make a better assessment of the fit of both the marginal model and the correlation structure of the observed data. Moreover, we will be able to compare the fit of several proposed models by overlaying plots constructed by fitting these models to the same data set.

When Y_t is discrete, we plot $F^m(\mathbf{y}, \hat{G}_0^m)$ versus $\bar{F}_n^m(\mathbf{y})$ for a finite number of points, focusing on values of $\mathbf{y}$ over which these functions tend to concentrate. When Y_t is continuous, we plot $F^m(\mathbf{y}, \hat{G}_0^m)$ versus $\bar{F}_n^m(\mathbf{y})$ over the entire range of $\mathbf{y}$. In the case of the plot of the univariate distributions, the functions are monotonic in $\mathbf{y}$, and hence their values are necessarily ordered with respect to the values of $\mathbf{y}$. The points of the multidimensional distributions, however, are not ordered in this way.

The requirements that we impose to ensure that the plot of the m-dimensional distributions has the above convergence property are as follows:

Requirement 1. $\{Y_t\}$ is strictly stationary.

Requirement 2. $F^m(\cdot, \hat{G}_0^m)$ converges to $F_0^m(\cdot)$.

Requirement 3. $\bar{F}_n^m(\cdot)$ converges to $F_0^m(\cdot)$.

REMARK. Requirement 1 implies that the joint distribution of $(Y_t, \ldots, Y_{t+\ell})$ is the same for all t. Requirement 2 will be satisfied (in the sense of pointwise convergence) if $F_0^m(\cdot)$ is continuous in the parameters and the parameter estimates are consistent. We may obtain consistent parameters by using either the method of maximum likelihood (when K^0 is known) or the penalized minimum-distance method described in Chapter 3 (when K^0 is known or unknown).

We discuss Requirement 3 in more detail in Section 4.1, and present two alternative sets of sufficient conditions for this requirement. We use these to show that our proposed graphical method is valid for stationary HMMs. Thus, we will be able to graphically compare different (including non-nested) HMMs for the observed data by examining how close each estimated distribution is to the empirical distribution.

More generally, we would like to know that the empirical distribution is converging to the true distribution *regardless* of whether the true distribution is a HMM. Since the examples in this thesis have focused on count data, in Section 4.2 we discuss other models for stationary series of count data. It turns out that these models meet at least one of our conditions. Thus, if the true underlying model is a member of the broad class that we consider, our method will allow us to determine whether the HMM in question is a reasonable model for our data. As an additional advantage, if consistent estimates of these alternative distributions are available, we will also be able to use our method to compare the fit of the HMM with that of the other models by overlaying the appropriate plots.

We apply our method to our two MS/MRI data sets in Section 4.3. These examples illustrate the type of deviations that we might see when a HMM does not represent the data

well, or when our choice of the conditional model for the observed data is not appropriate.

4.1 Convergence Conditions

The conditions for Requirement 3 that we develop are based on the concept of α-mixing sequences of random variables (see Definition 3.1). The idea is that for the empirical distribution to converge to the true distribution, α_ℓ must converge to 0 quickly enough. The two theorems that we present give sufficient rates of convergence. The first is due to Ould-Saïd (1994), and is applicable to plots of the multidimensional distributions. We cited this result in the proof of Theorem 3.1, but we repeat it here in the particular form that we require.

Theorem 4.1 *For a stationary, m-dimensional, α-mixing process with mixing coefficients $\alpha_\ell = O(\ell^{-\nu})$ for some $\nu > 2m + 1$,*

$$\lim_{n\to\infty} \sup_{\mathbf{y}\in\mathcal{R}^m} |\bar{F}_n^m(\mathbf{y}) - F_0^m(\mathbf{y})| = 0 \tag{4.1}$$

almost surely.

In the proof of Theorem 3.1, we showed that the mixing coefficients of stationary HMMs satisfy the condition in Theorem 4.1. Thus, our graphical method (in any dimension) is valid for these models.

If we consider the pointwise, rather than uniform, convergence of the empirical distribution to the true distribution, Requirement 3 amounts to a law of large numbers (LLN) for dependent variables. Lin & Lu (1996) provide general information in this context for processes satisfying various mixing conditions (e.g. α-mixing, ρ-mixing, ψ-mixing, and others). Theorem 4.2 below is an example of one such LLN. We focus on this particular result because of its relative simplicity in our context. In particular, for some models the conditions of Theorem 4.2 may be easier to verify than those of Lin & Lu (1996) (or of Theorem 4.1) because the calculation of the mixing coefficients is not required.

Theorem 4.2 *Assuming that $\{Y_t\}$ is stationary, let*

$$\beta_\ell(y) = |\mathrm{P}(Y_t \le y, Y_{t+\ell} \le y) - \mathrm{P}(Y_t \le y)\mathrm{P}(Y_{t+\ell} \le y)|.$$

Then for each y, $\bar{F}_n(y)$ converges in probability to $F_0(y)$ if

$$\sum_{\ell=1}^{n-1} (n-l)\beta_\ell(y) = o(n^2). \tag{4.2}$$

Proof. The proof follows from Chebyshev's inequality. Let $N(y) = \sum_{t=1}^{n} I(Y_t \leq y)$. Then for a given value of ϵ,

$$\begin{aligned}
&\mathrm{P}(\left|\bar{F}_n(y) - F_0(y)\right| \geq \epsilon) \\
&\leq \frac{1}{\epsilon^2}\mathrm{E}\left[\frac{N(y)}{n} - F_0(y)\right]^2 \\
&= \frac{1}{\epsilon^2}\mathrm{E}\left[\frac{N^2(y)}{n^2} - (F_0(y))^2\right] \\
&= \frac{1}{\epsilon^2}\left\{\frac{1}{n^2}\mathrm{E}\left[\sum_{t=1}^{n} I(Y_t \leq y) + 2\sum_{s<t} I(Y_t \leq y, Y_s \leq y)\right] - (F_0(y))^2\right\} \\
&= \frac{1}{\epsilon^2}\left\{\frac{1}{n^2}\left[nF_0(y) + 2\sum_{s<t} P(Y_t \leq y, Y_s \leq y)\right] - (F_0(y))^2\right\} \\
&\leq \frac{1}{\epsilon^2}\left\{\frac{F_0(y)}{n} + \frac{2}{n^2}\sum_{s<t}[(F_0(y))^2 + \beta_{t-s}(y)] - (F_0(y))^2\right\} \\
&= \frac{1}{\epsilon^2}\left\{\frac{F_0(y)}{n} + \frac{2n(n-1)}{2n^2}(F_0(y))^2 + \frac{2}{n^2}\sum_{\ell=1}^{n-1}(n-\ell)\beta_\ell(y) - (F_0(y))^2\right\} \\
&= \frac{1}{\epsilon^2}\left\{\frac{F_0(y)}{n} - \frac{(F_0(y))^2}{n} + \frac{2}{n^2}o(n^2)\right\} \\
&\to 0 \qquad \square
\end{aligned}$$

Although Theorem 4.2 is stated in terms of the univariate distributions, it can easily be extended to the multidimensional case. For example, to show the pointwise convergence of the empirical bivariate distribution, we would define

$$\beta_{t-s}(x,y) = |\mathrm{P}(Y_s \leq x, Y_{s+1} \leq y, Y_t \leq x, Y_{t+1} \leq y) - \mathrm{P}(Y_s \leq x, Y_{s+1} \leq y)\mathrm{P}(Y_t \leq x, Y_{t+1} \leq y)|$$

and replace the condition (4.2) by the requirement that

$$\sum_{\ell=1}^{n-2}(n-\ell-1)\beta_\ell(x,y) = o(n^2)$$

for all x and y.

4.2 Other Models for Count Data

Work to date on models for stationary time series of count data is nicely summarized in MacDonald and Zucchini (1997). In addition to HMMs, other possible models are

1. Markov models, including

- Integer-valued autoregressive (INAR) models (e.g. Alzaid & Al-Osh 1993; McKenzie 1988)
- "Observation-driven processes" of the form $Y_t \sim \text{Poisson}(\mu_t)$, where μ_t is modelled as $\log \mu_t = g(Y_{t-p}, \ldots, Y_{t-1})$, and g is some function (e.g. Albert *et al.* 1994)

2. m-dependent time series, including
   - Models of the form $Y_t = g(\gamma_t)$, where g is some function, $\{\gamma_t\}$ is a MA(m) process, and $Y_t \mid \gamma_t$ is independent of $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_n$
   - INMA(m) (integer-valued moving average) models (e.g. Alzaid & Al-Osh 1993; McKenzie 1988)
3. "Parameter-driven processes" of the form $Y_t \sim \text{Poisson}(\mu_t)$, where $\log \mu_t = g(\mathbf{x}) + \epsilon_t$, and ϵ_t is a stationary, autocorrelated error process (e.g. Chen & Ibrahim 2000)

We now show that all of these models, in fact, satisfy the convergence criterion given in Theorem 4.1, with the exception of that described by Albert *et al.* (1994), since this model is not stationary and is thus beyond the scope of this chapter.

4.2.1 Markov Models

In the proof of Theorem 3.1 we cite the result that the mixing coefficients of a Markov chain satisfy $\alpha_\ell \leq c\rho^\ell$, where c is a positive constant, and $0 < \rho < 1$ (see, e.g., Doukhan 1994). Thus, it is clear that stationary Markov chains satisfy the condition in Theorem 4.1, and hence our graphical method is valid for these processes. The INAR(p) process is a (p-order) example of such a process. Observation-driven processes, such as Poisson regression models with lagged dependent variables, are also p-order Markov processes, since in this case, the distribution of $Y_t | Y_{t-p}, \ldots, Y_{t-1}$ is independent of $Y_1, \ldots, Y_{t-p-1}$.

4.2.2 m-Dependent Time Series

Since, for an m-dependent time series, $\alpha_l = 0$ for $l > m$, time series of this type clearly satisfy the condition of Theorem 4.1. Included in this class are MA(m) and INMA(m) processes.

4.2.3 Parameter-Driven Processes

The next model we consider assumes that $Y_t \sim \text{Poisson}(\mu_t)$, with

$$\log \mu_t = g(\mathbf{x}) + \epsilon_t$$

where g is some function of the covariates, $\mathbf{x}$. These covariates are assumed to be constant over time so as not to violate the requirement that $\{Y_t\}$ be stationary. The error terms, $\{\epsilon_t\}$, are modelled as an ARMA(p, q) process, and $Y_t|\epsilon_t$ is assumed to be independent of $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_n$.

Blais *et al.* (2000) show that if $\{\epsilon_t\}$ is an α-mixing sequence with mixing coefficients α_ℓ, and $\{Y_t\}$ is a process such that $Y_t|\epsilon_t$ is independent of $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_n$, then the process $\{Y_t\}$ is also α-mixing, with mixing coefficients $4\alpha_\ell$. The proof is similar to that of Theorem 3.1.

From Liebscher (1996) we have that an ARMA(p, q) process is α-mixing with exponential rate, i.e. the mixing coefficients satisfy $\alpha_\ell \leq c\rho^\ell$ for some ρ, $0 < \rho < 1$, and some c, $0 < c < \infty$.

It is now obvious that the condition of Theorem 4.2 holds for models of this type.

4.3 Application to MS/MRI Data

In this section, assuming the same model for each patient, we compare the fit of five different stationary HMMs to both Albert's data and to data on the placebo patients from the Vancouver cohort of the PRISMS study (PRISMS Study Group 1998). Based on the results from Section 3.4, we assume that each HMM has two hidden states and use the method of maximum likelihood to obtain estimates of the other parameters. We model the conditional distribution of Y_t given Z_t as one of the following five distributions:

1. Poisson: $\text{P}(Y_t = y \mid Z_t = k) = \frac{e^{-\lambda_k}\lambda_k^y}{y!}$, $\lambda_k > 0$
2. Negative binomial: $\text{P}(Y_t = y \mid Z_t = k) = \binom{p_k+y-1}{y}\alpha_k^{p_k}(1-\alpha_k)^y$, $0 < \alpha_k < 1$, $p_k > 0$
3. Logarithmic: $\text{P}(Y_t = y \mid Z_t = k) = \frac{-\theta_k^{y+1}}{y\log(1-\theta_k)}$, $0 < \theta_k < 1$
4. Generalized Poisson: $\text{P}(Y_t = y \mid Z_t = k) = \frac{\lambda_k(\lambda_k+\theta_k y)^{y-1}e^{-\lambda_k-\theta_k y}}{y!}$, $\lambda_k > 0$, $\theta_k \geq 0$

5. Zero-extended Poisson:

$$\mathrm{P}(Y_t = y \mid Z_t = k) = w_k \delta_{\{y=0\}} + (1 - w_k)\frac{e^{-\lambda_k}\lambda_k^y}{y!}, \quad 0 < w_k < 1, \quad \lambda_k > 0$$

4.3.1 Albert's Data

Figures 4.1 and 4.2 show the fit of these models to Albert's data. In Figure 4.1, we plot the estimated univariate distribution of Y_t under each model versus the empirical distribution of Y_t over the range $0, \ldots, 20$. Figure 4.2 is the corresponding plot of the bivariate distributions over the range $(0,0), (0,1), \ldots, (20,20)$.

Note that the HMM involving the zero-extended Poisson distribution was not fit to Albert's data. Since there were, in fact, no zeroes in this data set, the parameter w could not be estimated for any of the components of the HMM.

Figure 4.1 shows that these models seem to capture the univariate behaviour of Albert's data quite well, with the exception of the logarithmic model. Since the Poisson model seems to be reasonable, it is not surprising that the negative binomial and generalized Poisson models also provide good fits, since these are generalizations of the Poisson model. In contrast, Figure 4.2, shows that none of the models is a good choice for representing the bivariate behaviour of the data. In particular, the estimated probabilities tend to be lower than the empirical probabilities throughout almost the entire range. Thus, it would appear that a 2-state HMM cannot fully capture the correlation structure of the data, and hence is not an adequate model in this case.

4.3.2 Vancouver PRISMS Data

The Vancouver data have the same format as Albert's, but consist of 13 rather than 3 patients. Each patient has between 2 and 26 observations. The lesion counts are most frequently zero, but range up to 15.

Figures 4.3 and 4.4 give the plots of the univariate and bivariate distributions for the Vancouver data, with the functions plotted over the ranges $0, \ldots, 15$ and $(0,0), (0,1), \ldots, (15,15)$, respectively. These figures show that the Poisson model seems to provide quite a good fit to these data, both for the univariate and bivariate distributions, although it may underestimate somewhat the probability of seeing pairs of small lesion counts. Again, the logarithmic model seems inappropriate for these data.

Figure 4.1: Comparison of the Estimated and Empirical Univariate Distributions (Albert's Data)

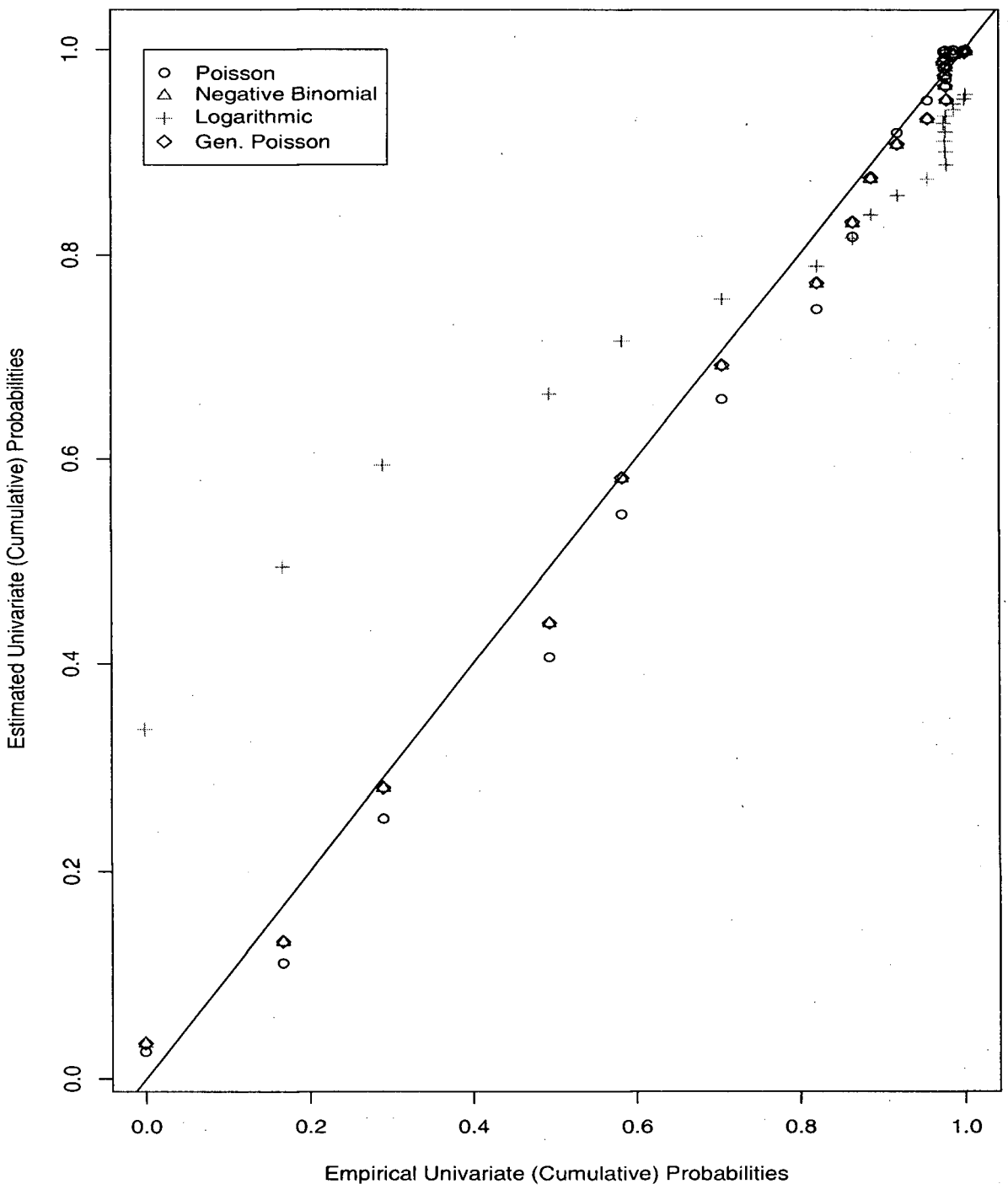


Figure 4.2: Comparison of the Estimated and Empirical Bivariate Distributions (Albert's Data)

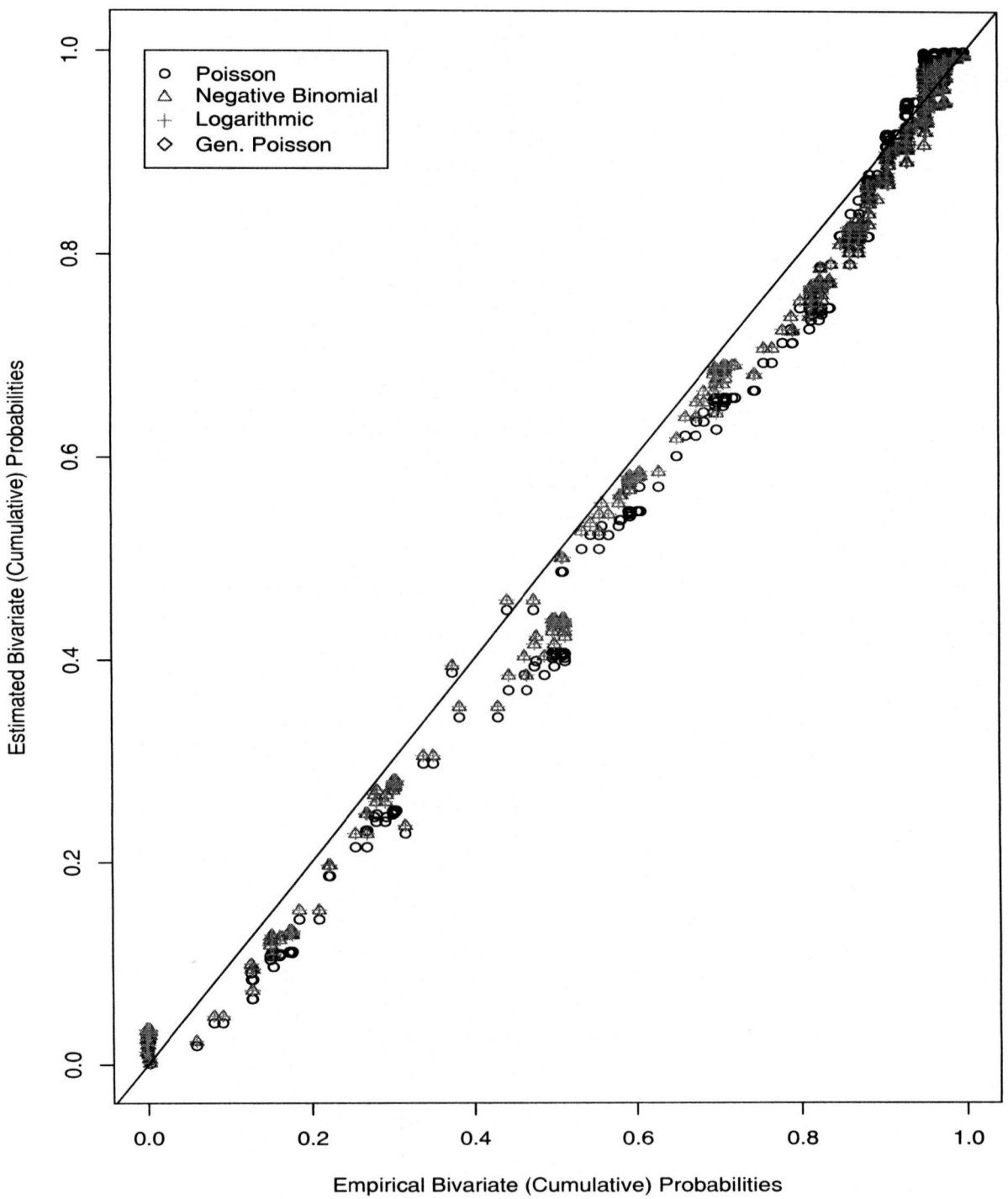


Figure 4.3: Comparison of the Estimated and Empirical Univariate Distributions (Vancouver Data)

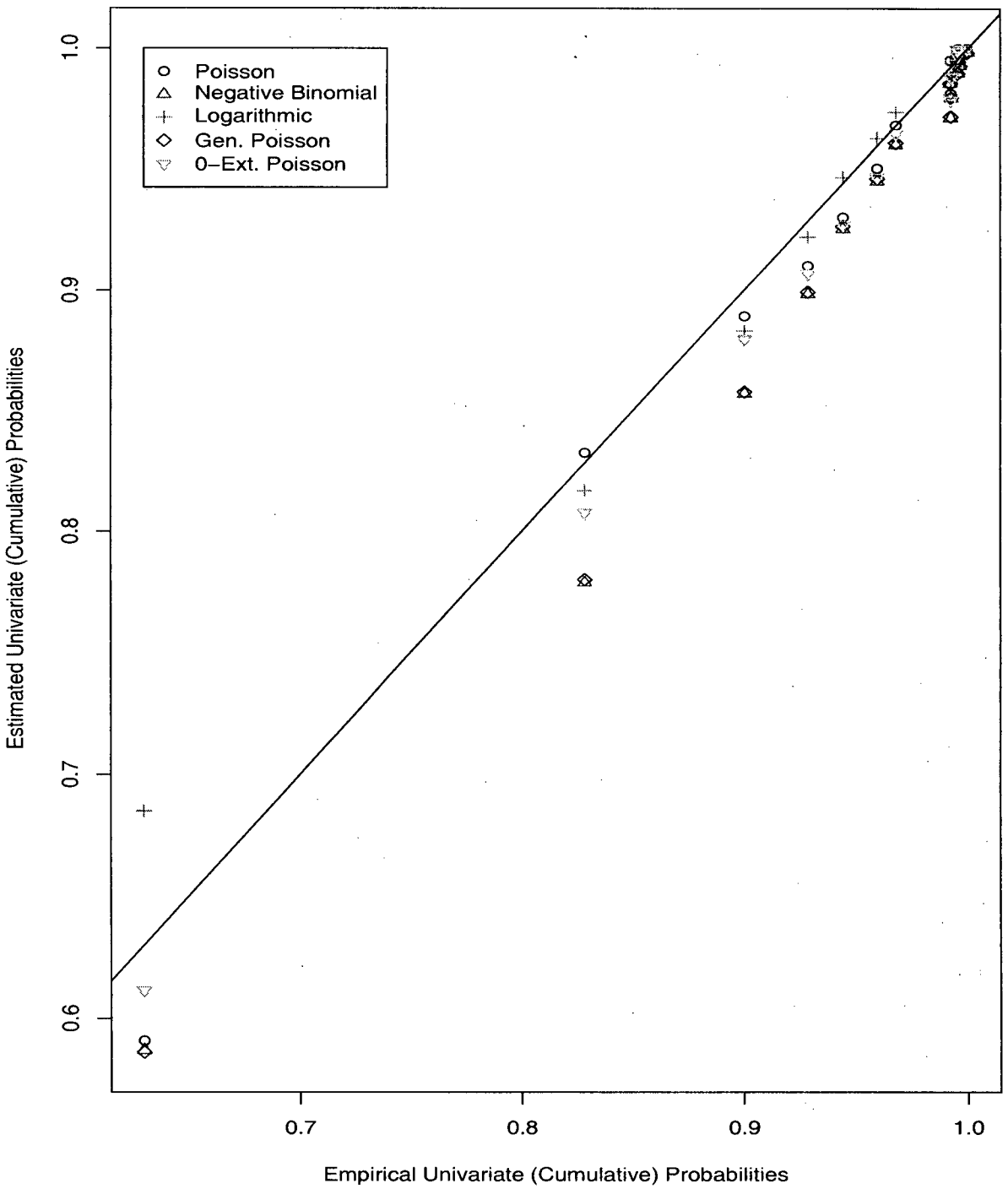
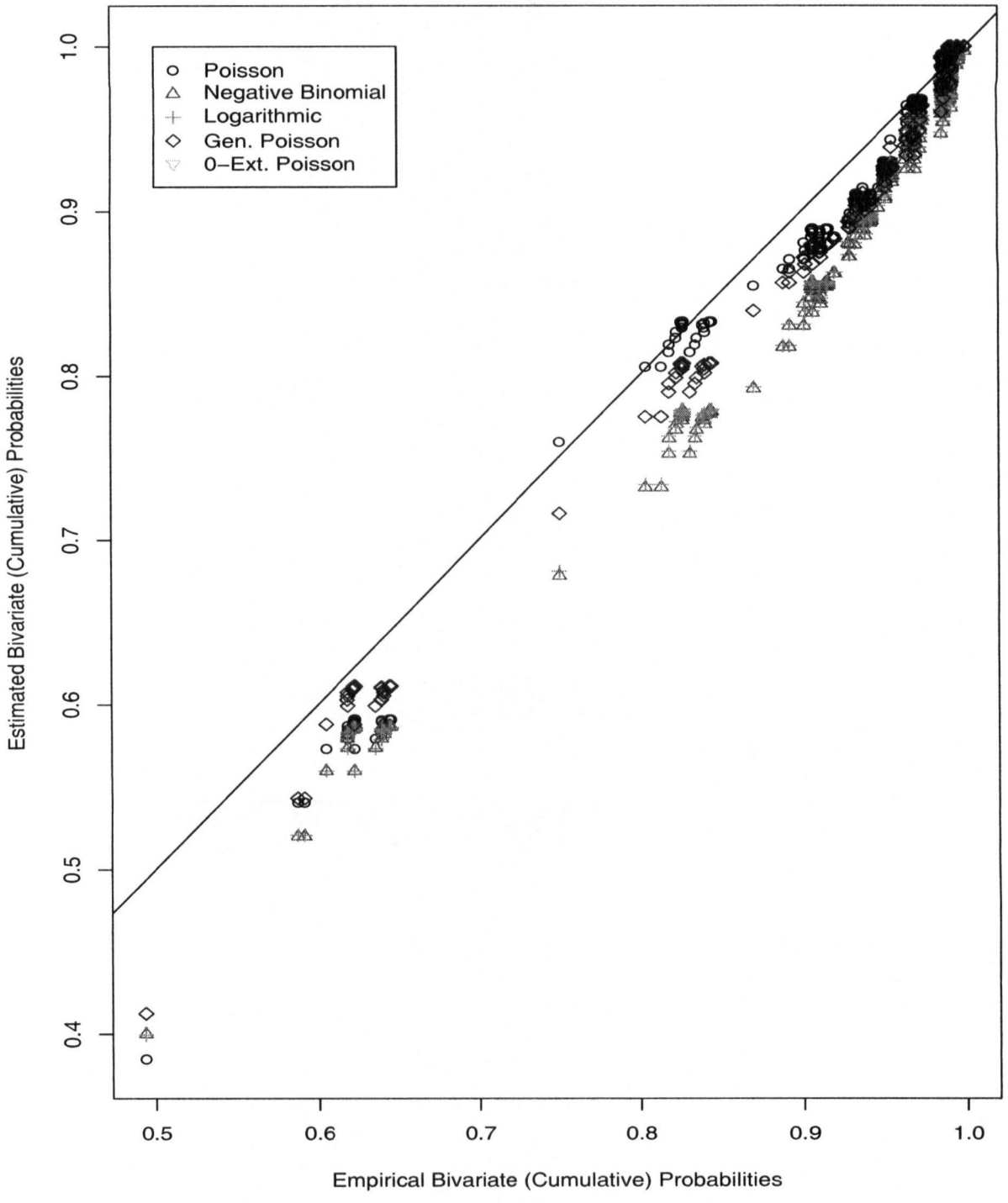


Figure 4.4: Comparison of the Estimated and Empirical Bivariate Distributions (Vancouver Data)



4.4 Formal Assessment of the GOF Plots

The examples in Section 4.3 show that our GOF method is useful both for comparing different models, and for detecting when a proposed HMM is not appropriate for the data. We have also proved in Section 4.1 that if we have correctly specified the model, then the plot will converge to a 45° line through the origin as $n \to \infty$. It is also of interest to develop a formal method of assessing the degree of variability in the observed plot. In other words, it would be desirable to have a theoretical means of determining whether the observed scatter around the 45° line is "acceptable" for a given sample size, n.

One way in which other authors have assessed this variability is by computing the correlation coefficient of the two plotted variables, and then deriving the distribution of a test statistic based on this coefficient under the null hypothesis that the model fits. This derivation is simplified considerably if one of the variables is fixed rather than random. Lockhart & Stephens (1998) provide a good example in this setting. They investigate the use of probability plots, where the n observations are ordered and plotted against the values $F^{-1}\{k/(n+1)\}$, $k = 1, \ldots, n$. Here F is an arbitrary distribution in the proposed family of distributions. Under the assumptions that F is in the location-scale family (usually with the values of the location and scale parameters chosen as 0 and 1, respectively) and that the observations are iid, the asymptotic distribution of their test statistic has a nice form. However, typically, the exact distribution of this statistic is not easily derived. Furthermore, in our setting, where we plot two different estimates of the cdf and our observations are not independent, computing the distribution of the associated correlation coefficient, even asymptotically, seems like a very challenging problem.

Alternatively, in the context where the proposed distribution is completely specified under the null hypothesis, Raubertas (1992) considers envelope plots as a formal GOF test. He suggests simulating s independent samples of size n from this distribution, and preparing a plot of the estimated distribution against the true distribution for each. These plots are then superimposed and summarized by displaying only their upper and lower envelopes. Points corresponding to the observed (as opposed to simulated) data that fall outside this envelope indicate lack of fit in the model. The advantage of this method is that the true distribution need not be limited in its complexity. For example, we could easily simulate observations from a HMM. However, Raubertas (1992) points out that the power of this test may be undesirably low, and does not recommend it when other options are available. An envelope plot in our case would have even more inherent variability, since we would need to sample from the estimated, rather than true, distribution. For the same reason, the computational burden would also be quite heavy. Given these concerns, we have not attempted to conduct

this test on our data sets.

In conclusion, a feasible, theoretical method of assessing the variability in our GOF plots is not yet available at this time. Further research on this topic is required.

Chapter 5

Hidden Markov Models for Multiple Processes

In Chapters 3 and 4, we discussed the problems of estimating the number of hidden states and assessing the goodness-of-fit of a single HMM. We now move on to extensions of the basic model, in particular, the use of the HMM framework for modelling multiple processes.

Few such HMMs currently exist. Most have been developed in the context of specific applications, and hence have not been posed in their full generality. Not surprisingly, little is known about the theory surrounding these models.

Hughes & Guttorp (1994) present one example of such a model: a multivariate HMM for data consisting of daily binary rainfall observations (rain or no rain) at four different stations. These time series are assumed to be functions of the same underlying process. Given the hidden state at time t, the observed measurements at this time are considered to be independent of all measurements taken at other time points. They may, however, depend on one another.

Turner *et al.* (1998) and Wang & Puterman (1999), working in the setting of Poisson count data, develop models for independent processes, each of which depends on a different underlying process. To account for between-subject differences, these authors incorporate covariates into the conditional model for the observed process. MacDonald & Zucchini (1997, Chapter 3) also provide a brief discussion of this idea. Between-subject differences can also occur in the transition probabilities, and Hughes & Guttorp (1994) address this issue in the context of precipitation modelling by including covariates in the model for the hidden process.

The addition of random effects is a natural extension of these models. In the subject-

area literature, HMMs with random effects have appeared in a limited way. For instance, Humphreys (1997, 1998) suggests such a model for representing labour force transition data. He works with binary observations on employment status (employed or unemployed) which are assumed to contain reporting errors. The misclassification probabilities, as well as the transition probabilities, depend on a subject-specific random effect. Seltman (2002) proposes a complex biological model for describing cortisol levels over time in a group of patients. The initial concentration of cortisol in each patient is modelled as a random effect.

The goal of this chapter is to develop a new class of models, HMMs with random effects, which will unify existing HMMs for multiple processes and will provide a general framework for working in this context. We will build on the ideas of generalized linear mixed models (GLMMs), and will address theoretical questions regarding interpretation, estimation, and hypothesis testing. For concreteness, we will present our ideas in the setting where the observed processes are repeated measurements on a group of patients. This choice will not, however, limit the generality of our models.

The advantages of HMMs with random effects are numerous. Most importantly, modelling multiple processes simultaneously permits the estimation of population-level effects. For example, to study the impact of a new MS drug, the individual models proposed by Albert *et al.* (1994) would not be sufficient. Estimating the treatment effect would require the assumption of some commonality among the responses of patients in the same treatment group while allowing for differences in patients in different treatment groups. A second advantage of these models is that they allow more efficient estimation of patient-level effects while recognizing inter-patient differences. This feature is particularly important in view of our examples and discussion in Section 2.4. Finally, HMMs with random effects permit greater flexibility in modelling the correlation structure of the data. Standard HMMs assume that the serial dependence induced by the hidden process completely characterizes the correlation structure of the observed data. While this may be reasonable for a single process, in multiple processes, the correlation structure could be more complex.

In our presentation of these models, we first consider the case where the random effects appear only in the conditional model for the observed data. We then explore the more difficult case where the random effects also appear in the model for the transition probabilities. We will show that interpretation and estimation of HMMs with random effects in the conditional model for the observed data is quite straightforward, but is more challenging when there are random effects in the model for the hidden process. We defer the study of the properties of these estimators and hypothesis testing to Chapter 6.

5.1 Notation and Assumptions

We will denote the observation on patient i, $i = 1, \ldots, N$, at time t, $t = 1, \ldots, n_i$, by Y_{it}, and the corresponding hidden state by Z_{it}. As in Section 2.1, we will assume that the time points are equally spaced. Furthermore, we will assume that Z_{it} takes on values from a finite set, $\{1, 2, \ldots, K\}$, where K is known.

Let $\sum_{i=1}^{N} n_i = n_T$. Then, we denote by $\mathbf{Y}_i$ the n_i-dimensional vector of observations on patient i, and by $\mathbf{Y}$ the n_T-dimensional vector of all observations. Similarly, let $\mathbf{Z}_i$ be the n_i-dimensional vector of hidden states for patient i, and let $\mathbf{Z}$ be the n_T-dimensional vector of all hidden states.

We assume that, conditional on the random effects, $\{Z_{it}\}_{t=1}^{n_i}$ is a Markov chain. These processes may or may not be stationary. If $\{Z_{it}\}$ are conditionally stationary with unique stationary distributions, we will use these distributions for the initial probabilities. In other words, we will compute the initial distributions based on the transition probabilities, so that these probabilities may vary among patients. Otherwise, we will assume that the initial probabilities are fixed parameters and are the same for all patients. We generally have very little information with which to estimate the initial probabilities, so allowing for further complexity does not seem necessary.

Finally, we assume that, conditional on the random effects, the ith process, $\{Y_{it}\}_{t=1}^{n_i}$, is a HMM, and that these HMMs are independent of one another.

5.2 Model I: HMMs with Random Effects in the Conditional Model for the Observed Process

In this section, we focus on the addition of random effects to the conditional model for the observed data, and assume that random effects do not appear in the model for the hidden processes. In particular, we assume that the hidden processes are homogeneous with transition probabilities, $\{P_{k\ell}\}$, and initial probabilities, $\{\pi_k\}$, common to all subjects.

Borrowing from the theory of GLMMs (see, e.g., McCulloch & Searle 2001) and again using ψ to represent the vector of all model parameters, we assume that, conditional on the random effects, $\mathbf{u}$, and the hidden states $\mathbf{Z}$, $\{Y_{it}\}$ are independent with distribution in the exponential family, i.e.

$$f(y_{it} \mid Z_{it} = k, \mathbf{u}, \psi) = \exp\{(y_{it}\eta_{itk} - c(\eta_{itk}))/a(\phi) + d(y_{it}, \phi)\}. \tag{5.1}$$

Here

$$\eta_{itk} = \tau_k + \mathbf{x}'_{it}\beta_k + \mathbf{w}'_{itk}\mathbf{u}, \tag{5.2}$$

where τ_k is the fixed effect when $Z_{it} = k$, $\mathbf{x}'_{it}$ are the covariates for patient i at time t, and $\mathbf{w}'_{itk}$ is the row of the model matrix for the random effects for patient i at time t in state k. We will denote the distribution of the random effects by $f(\mathbf{u};\psi)$, and will assume that the random effects are independent of the hidden states. The notation $\mathbf{u}$ (as opposed to $\mathbf{u}_i$) indicates that a random effect could be common to more than one patient. The form of the exponential family (5.1) assumes that the link function is canonical, an assumption that is not strictly necessary, but one that will facilitate our exposition. We will henceforth refer to this model as Model I.

The notation in (5.2) was selected for its generality, but in most applications we would expect significant simplification. For example, consider the MS/MRI setting where we have two treatment groups, with group membership indicated by the explanatory variable x_i. We might postulate the existence of two hidden states, one associated with relapse (state 1) and the other associated with remission (state 2). We prefer this interpretation of the hidden states to that of Albert *et al.* (1994) since it implies that patients' mean lesion counts can remain stable (at either a high or low level of activity) while assuming only two hidden states (in contrast with the model in Section 2.4.5). If we believe that the mean lesion count in state k varies randomly among patients, we might choose the model

$$\eta_{itk} = \tau_k + x_i\beta_{1k} + t\beta_{2k} + u_{ik},$$

where u_{ik} is independent of $u_{jk'}$, $i \neq j$, and (u_{i1}, u_{i2}) has some bivariate distribution. This model allows the mean lesion count in state k to vary not only with patient, but also with treatment group and with time.

The model (5.2) could be made even more general by allowing the parameter ϕ to vary among patients. We do not consider this case here, but provide a brief discussion in Chapter 7.

The likelihood for Model I is

$$\begin{aligned}
\mathcal{L}(\psi) &= f(\mathbf{y};\psi) \\
&= \int_{\mathbf{u}} \sum_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{z}, \mathbf{u}, \psi) f(\mathbf{z} \mid \mathbf{u}, \psi) f(\mathbf{u};\psi)\, d\mathbf{u} \\
&= \int_{\mathbf{u}} \sum_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{z}, \mathbf{u}, \psi) f(\mathbf{z};\psi) f(\mathbf{u};\psi)\, d\mathbf{u} \\
&= \int_{\mathbf{u}} \sum_{\mathbf{z}} \left\{ \prod_{i=1}^{N} \prod_{t=1}^{n_i} f(y_{it} \mid z_{it}, \mathbf{u}, \psi) \right\} \left\{ \prod_{i=1}^{N} \pi_{z_{i1}} \prod_{t=2}^{n_i} P_{z_{i,t-1},z_{it}} \right\} f(\mathbf{u};\psi)\, d\mathbf{u} \\
&= \int_{\mathbf{u}} \sum_{\mathbf{z}} \left\{ \prod_{i=1}^{N} \pi_{z_{i1}} f(y_{i1} \mid z_{i1}, \mathbf{u}, \psi) \prod_{t=2}^{n_i} P_{z_{i,t-1},z_{it}} f(y_{it} \mid z_{it}, \mathbf{u}, \psi) \right\} f(\mathbf{u};\psi)\, d\mathbf{u}. \qquad (5.3)
\end{aligned}$$

As in (2.2), we may simplify this expression by writing the summation as a product of matrices. For a given value of $\mathbf{u}$, let A^{i1} be the vector with elements $A^{i1}_k = \pi_k \, f(y_{i1} \mid Z_{i1} = k, \mathbf{u})$, and let A^{it} be the matrix with elements $A^{it}_{k\ell} = P_{k\ell} \, f(y_{it} \mid Z_{it} = \ell, \mathbf{u})$, $t > 1$. Again, let $\mathbf{1}$ be the K-dimensional vector of 1's. Then

$$\mathcal{L}(\psi) = \int_{\mathbf{u}} \prod_{i=1}^{N} \left\{ (A^{i1})' \left(\prod_{t=2}^{n_i} A^{it} \right) \mathbf{1} \right\} f(\mathbf{u}; \psi) \, d\mathbf{u}. \tag{5.4}$$

From (5.4) it becomes clear that the only impact of the random effects in (5.2) on the computation of the likelihood is the introduction of the integration over the distribution of the random effects. In other words, it is only the complexity of this integral that makes evaluating (5.4) more difficult than the likelihood for standard HMMs. In most applications, (5.4) reduces to a very simple form. We consider two special cases.

EXAMPLE 1 (Single, patient-specific random effect). We assume that u_i is a random effect associated with the ith patient, $i = 1, \ldots, N$, and that $\{u_i\}$ are iid. Under this model, observations on different patients are independent. Furthermore, our model specifies that, given u_i and the sequence of hidden states, the observations collected on patient i are independent. In this case, (5.3) and (5.4) simplify to a one-dimensional integral:

$$\begin{aligned} \mathcal{L}(\psi) &= \prod_{i=1}^{N} \int_{u_i} \left\{ \sum_{\mathbf{z}} \pi_{z_{i1}} f(y_{i1} \mid z_{i1}, u_i, \psi) \prod_{t=2}^{n_i} P_{z_{t-1}, z_{it}} f(y_{it} \mid z_{it}, u_i, \psi) \right\} f(u_i; \psi) \, du_i \\ &= \prod_{i=1}^{N} \int_{u_i} \left\{ (A^{i1})' \prod_{t=2}^{n_i} A^{it} \right\} \mathbf{1} f(u_i; \psi) \, du_i. \end{aligned} \tag{5.5}$$

EXAMPLE 2 (Patient-specific and centre-specific random effects). We incorporate patient-specific random effects as in Example 1. However, we also allow multiple treatment centres, with a random effect associated with each (assumed independent of the patient-specific random effects). Consequently, observations on patients from different centres are independent, and, given the centre-specific random effects, observations on different patients at the same centre are independent. We use the same notation as before, but with an additional index, c, $c = 1, 2, \ldots, C$, to indicate treatment centre. We assume that treatment centre c has m_c patients. Then (5.3) and (5.4) reduce to a two-dimensional integral:

$$\begin{aligned} \mathcal{L}(\psi) &= \prod_{c=1}^{C} \int_{v_c} \prod_{i=1}^{m_c} \int_{u_{ci}} \sum_{\mathbf{z}_{ci}} \left\{ \pi_{z_{ci1}} f(y_{ci1} \mid z_{ci1}, u_{ci}, v_c, \psi) \prod_{t=2}^{n_{ci}} P_{z_{c,i,t-1}, z_{cit}} f(y_{cit} \mid z_{cit}, u_{ci}, v_c, \psi) \right\} \\ &\qquad \cdot f(u_{ci}; \psi) \, du_{ci} \, f(v_c; \psi) \, dv_c \\ &= \prod_{c=1}^{C} \int_{v_c} \prod_{i=1}^{m_c} \int_{u_{ci}} \left\{ (A^{ci1})' \prod_{t=2}^{n_{ci}} A^{cit} \right\} \mathbf{1} f(u_{ci}; \psi) \, du_{ci} \, f(v_c; \psi) \, dv_c. \end{aligned} \tag{5.6}$$

In the usual case where the integral in (5.4) is of low dimension, numerical methods of evaluation seem to work well. For common choices of distribution for $\mathbf{u}$, Gaussian quadrature methods offer both accuracy and efficiency. For example, the Gauss-Laguerre formula gives good approximations to integrals of the form $\int f(x)x^{\alpha}e^{-x}dx$. The case where there is one patient-specific random effect, u_i, with $f(u_i)$ in the gamma family of distributions, fits into this framework. In the special cases considered by Humphreys (1997, 1998), (5.4) can actually be evaluated analytically.

The EM algorithm may seem like a natural choice for parameter estimation in this setting if we think of the random effects as "missing" along with the values of $\{Z_{it}\}$. However, again, we do not recommend this algorithm because of efficiency considerations (see Section 2.2). Nonetheless, for completeness, we include the details of the procedure in Appendix A.2 for the usual case where the random effects are patient-specific. We show that, when random effects are included in this way, the increase in difficulty in the implementation of this algorithm (compared to the simple case described in Appendix A.1) essentially depends only on the complexity of the integral.

Monte Carlo (MC) methods allow the circumvention of the evaluation of the integral, and hence may also prove useful for either maximizing the likelihood directly or implementing the EM algorithm. For example, the MC Newton-Raphson and MCEM algorithms presented by McCulloch (1997) in the GLMM context may be applicable to our models as well.

5.3 Moments Associated with Model I

It is easy to create complex models by postulating the existence of latent variables. However, caution must be exercised in order to avoid unwanted implications of our modelling choices. In particular, it is important to be able to interpret both the fixed and random effects. One way of understanding their impact on the model is to examine the resulting marginal moments of the observed process. We use the property of exponential families that $\mathrm{E}[Y_{it} \mid Z_{it} = k, \mathbf{u}] = c'(\eta_{itk})$ and $\mathrm{Var}[Y_{it} \mid Z_{it} = k, \mathbf{u}] = c''(\eta_{itk})a(\phi)$ (see, e.g., McCullagh & Nelder 1989). In addition, we use our assumption that $\mathrm{Cov}[Y_{it}, Y_{js} \mid Z_{it} = k, Z_{js} = \ell, \mathbf{u}] = 0$. Then, under Model I,

$$\mathrm{E}[Y_{it}] = \mathrm{E}[c'(\eta_{itk})] \tag{5.7}$$

$$\mathrm{Var}[Y_{it}] = \mathrm{E}[c''(\eta_{itk})]a(\phi) + \mathrm{Var}[c'(\eta_{itk})] \tag{5.8}$$

$$\mathrm{Cov}[Y_{it}, Y_{js}] = \mathrm{Cov}[c'(\eta_{itk}), c'(\eta_{js\ell})], \quad s < t. \tag{5.9}$$

Here the expectations are over both $\mathbf{Z}$ and $\mathbf{u}$.

In general, the moments (5.7)-(5.9) do not have closed forms. Even if $\{Z_{it}\}_{t=1}^{n_i}$ is stationary, we can not always evaluate these expressions analytically. However, the Laplace method (e.g. Evans & Swartz 1995) can provide good approximations when the variance components of the random effects are reasonably small. And, for certain common distributions of $Y_{it} \mid Z_{it}, \mathbf{u}$ (e.g., Poisson, normal) and of the random effects (e.g. multivariate normal), closed forms do exist.

For instance, consider the case where $\mathbf{u} \equiv (u_1, \ldots, u_N)$ is a vector of iid patient-specific random effects, as in Example 1. Let $Y_{it} \mid Z_{it}, u_i \sim \text{Poisson}(\mu_{it})$ with

$$\log \mu_{it} = \tau_{Z_{it}} + u_i, \tag{5.10}$$

where $u_i \sim N(0, \sigma^2)$, and $\{Z_{it}\}_{t=1}^{n_i}$ is stationary with stationary distribution $\{\pi_k\}$. Since $a(\phi) = 1$ for the Poisson distribution, and since $\text{Var}[c'(\eta_{itk})] \geq 0$, we can see by comparing (5.7) and (5.8) that this model allows for overdispersion in the Poisson counts. In this case, we can evaluate (5.7)-(5.9) as

$$\begin{aligned}
\text{E}[Y_{it}] &= e^{\frac{1}{2}\sigma^2} \sum_k e^{\tau_k} \pi_k \\
\text{Var}[Y_{it}] &= e^{\frac{1}{2}\sigma^2} \sum_k e^{\tau_k} \pi_k + e^{2\sigma^2} \sum_k e^{2\tau_k} \pi_k - e^{\sigma^2} \left(\sum_k e^{\tau_k} \pi_k \right)^2 \\
\text{Cov}[Y_{is}, Y_{jt}] &= \begin{cases} 0, & i \neq j \\ e^{2\sigma^2} \sum_k \sum_\ell e^{\tau_k + \tau_\ell} P_{k\ell}(t-s) \pi_k - e^{\sigma^2} \left(\sum_k e^{\tau_k} \pi_k \right)^2, & i = j \end{cases}
\end{aligned}$$

where $P_{k\ell}(t-s)$ denotes the $(t-s)$-step transition probability, $\text{P}(Z_{it} = \ell \mid Z_{is} = k, \psi)$.

One feature of the moments (5.7)-(5.9) is that they depend on the covariates, $\mathbf{x}_{it}$, in a very specific way. These relationships should be carefully considered in terms of their scientific justifiability before applying the proposed model.

Thinking of s as fixed, it is interesting to look at the covariance (5.9) as $t \to \infty$ when $i = j$. Consider the case where $\{Z_{it}\}$ is stationary and $\mathbf{x}_{it} \equiv \mathbf{x}_i$ and $\mathbf{w}_{itk} \equiv \mathbf{w}_i$ are independent of t and k. We have that

$$\begin{aligned}
\text{Cov}[Y_{it}, Y_{is}] &= \text{E}\left\{ \sum_{k,\ell} c'(\tau_k + \mathbf{x}_i\beta_k + \mathbf{w}_i'\mathbf{u}) c'(\tau_\ell + \mathbf{x}_i\beta_\ell + \mathbf{w}_i'\mathbf{u}) P_{k\ell}(t-s)\pi_k \right\} \\
&\quad - \left\{ \text{E}\left[\sum_k c'(\tau_k + \mathbf{x}_i\beta_k + \mathbf{w}_i'\mathbf{u})\pi_k \right] \right\}^2.
\end{aligned}$$

Now, for each k and ℓ, $P_{k\ell}(t-s) \to \pi_\ell$ as $t \to \infty$. Setting

$$X_t \equiv \sum_{k,\ell} c'(\tau_k + \mathbf{x}_i\beta_k + \mathbf{w}_i'\mathbf{u}) c'(\tau_\ell + \mathbf{x}_i\beta_\ell + \mathbf{w}_i'\mathbf{u}) P_{k\ell}(t-s)\pi_k$$

and

$$X \equiv \sum_k c'(\tau_k + \mathbf{x}_i\beta_k + \mathbf{w}_i'\mathbf{u})\pi_k,$$

we have $X_t \to X^2$ as $t \to \infty$. Assuming that the dominated convergence theorem (DCT) holds, we see that

$$\begin{aligned} \mathrm{Cov}[Y_{it}, Y_{is}] &\to \mathrm{E}[X^2] - \{\mathrm{E}[X]\}^2 \\ &= \mathrm{Var}[X] \\ &\geq 0, \end{aligned}$$

with equality occurring if and only if the distribution of $\mathbf{w}_i'\mathbf{u}$ is degenerate. Since

$$|X_t| \leq \sum_{k,\ell} |c'(\tau_k + \mathbf{x}_i\beta_k + \mathbf{w}_i'\mathbf{u})c'(\tau_\ell + \mathbf{x}_i\beta_\ell + \mathbf{w}_i'\mathbf{u})|\,, \tag{5.11}$$

the DCT will hold if $c(\cdot)$ and $f(\mathbf{u};\psi)$ are chosen such that the right-hand side of (5.11) is integrable. This is true, for example, in the case where $f(y_{it}|z_{it},\mathbf{u},\psi)$ is in the Poisson or normal family and $\mathrm{E}[\mathbf{u}] < \infty$. Thus, referring again to our example in the Poisson context (5.10), we have that

$$\mathrm{Cov}[Y_{is}, Y_{it}] \to (e^{2\sigma^2} - e^{\sigma^2}) \left(\sum_k e^{\tau_k}\pi_k \right)^2 \geq 0,$$

with equality occurring if and only if $\sigma = 0$.

In contrast, when we remove the random effects from this model (i.e. we assume the same model for each patient), $\mathrm{Cov}[Y_{it}, Y_{is}] \to 0$ as $t \to \infty$. Thus, the random effects impact the correlation structure of the observed data in the expected way. In particular, they allow a long-range, positive dependence in each patient's observations, as we observe in the linear mixed model setting. This is an additional layer of dependence, superimposed on the usual correlation structure of a HMM. Considering the heterogeneity of MS patients, this model may be more realistic for that context than models where the same HMM is applied to all patients.

5.4 Model II: HMMs with Random Effects in the Model for the Hidden Process

It may also be desirable to allow the hidden Markov chain to vary randomly among patients. For example, in the MS/MRI context, patients may spend differing proportions of time in relapse and remission.

However, allowing random effects in the hidden process of the HMM is a challenging problem, regardless of whether random effects also appear in the conditional model for the observed data. To explore this class of models, we will again specify the conditional model for the observed data by (5.1) and (5.2), but we will now allow the model for the hidden process to vary among patients.

In particular, we will assume that $\{Z_{it} \mid \mathbf{u}\}_{t=1}^{n_i}$ is a Markov chain and that $Z_{it} \mid \mathbf{u}$ is independent of $Z_{js} \mid \mathbf{u}$ for $i \neq j$. Naturally, any model for these Markov chains must satisfy the constraints that the transition probabilities lie between 0 and 1, and that the rows of the transition probablity matrices sum to 1. Thus, we propose modelling the transition probabilities as

$$\mathrm{P}(Z_{it} = \ell \mid Z_{i,t-1} = k, \mathbf{u}, \psi) = \frac{\exp\left\{\tau^*_{k\ell} + \mathbf{x}^{*\prime}_{it}\beta^*_{k\ell} + \mathbf{w}^{*\prime}_{itk\ell}\mathbf{u}\right\}}{\sum_{h=1}^{K} \exp\left\{\tau^*_{kh} + \mathbf{x}^{*\prime}_{it}\beta^*_{kh} + \mathbf{w}^{*\prime}_{itkh}\mathbf{u}\right\}}. \tag{5.12}$$

We use asterisks here to distinguish the model matrices and parameters from those in (5.2). The vector $\mathbf{u}$ now contains the random effects associated with the hidden process as well as those associated with the conditional model for the observed data. To prevent over-parameterization, we define $\tau^*_{kK} \equiv \beta^*_{kK} \equiv 0$ for all k, and set $\mathbf{w}^{*\prime}_{itkK}$ to be a row of 0's for all i, t, k. We will refer to this model as Model II.

The likelihood associated with Model II is very similar to (5.4). Again, we may write

$$\mathcal{L}(\psi) = \int_{\mathbf{u}} \prod_{i=1}^{N} \left\{ (A^{i1})' \left(\prod_{t=2}^{n_i} A^{it} \right) \mathbf{1} \right\} f(\mathbf{u}; \psi)\, d\mathbf{u}, \tag{5.13}$$

where now we define the quantities A^{i1}_k and $A^{it}_{k\ell}$ as $A^{i1}_k = \pi^i_k\ f(y_{i1} \mid Z_{i1} = k, \mathbf{u}, \psi)$ and $A^{it}_{k\ell} = \mathrm{P}(Z_{it} = \ell \mid Z_{i,t-1} = k, \mathbf{u}, \psi)\ f(y_{it} \mid Z_{it} = \ell, \mathbf{u}, \psi)$, $t > 1$.

At first glance, (5.4) and (5.13) look the same. However, in the case of (5.13), the integral will typically be quite complicated, even in simple situations such as that presented in Example 3 below.

EXAMPLE 3 (Patient-specific random effects). We assume that the random effects are patient-specific, so that observations on different patients are independent. In particular, for patient i, we model the transition probabilities as

$$\mathrm{P}(Z_{it} = \ell \mid Z_{i,t-1} = k, u^*_{ik\ell}, \psi) = \frac{\exp\left\{\tau^*_{k\ell} + u^*_{ik\ell}\right\}}{\sum_{h=1}^{K} \exp\left\{\tau^*_{kh} + u^*_{ikh}\right\}},$$

where $\tau^*_{kK} \equiv u^*_{ikK} \equiv 0$ for all i, k. The likelihood associated with this model can be simplified as in (5.5). However, instead of requiring only one random patient effect, we will need $K(K-1)$ (possibly correlated) random effects for each patient. There is no obvious way

to reduce this number; simplifications such as $u^*_{ik\ell} \equiv u^*_i$ do not have sensible interpretations in terms of the transition probabilities for patient i. The restriction $u^*_{ik\ell} \equiv u^*_{i\ell k}$ for all k and ℓ may be appropriate in some cases, but still results in quite a large number of random effects. Moreover, adding a random effect to the conditional model for the observed data (as in Example 1) would further increase the dimension of the integral; if this random effect were correlated with those in the model for the hidden process, the resulting integral could be quite complex.

We see that the estimation of the parameters in this setting is a difficult problem, even in simple cases such as Example 3. The high-dimensional integrals in these models not only create a potentially prohibitive computational burden, but also raise questions about the accuracy with which we can evaluate (5.13). For similar reasons, computational difficulties may arise in the implementation of the EM algorithm as well. See Appendix A.3 for details in the case where the random effects are patient-specific.

As a final note, the number of parameters associated with the distribution of $\mathbf{u}$ may be undesirably large. For example, consider the model with $K = 2$ where there are two random effects associated with the transition probabilities, and one associated with the conditional model for the observed data, and all are correlated (as in Example 3). If we model the distribution of these three random effects as multivariate normal with mean 0 and variance-covariance matrix $\mathbf{Q}$, this distribution will have 6 unknown parameters. We could reduce this number by assuming that $\{u_i\}$ are independent with distribution $N(0, \sigma_i^2)$, $i = 1, 2, 3$, or even that $\mathbf{Q} = \sigma^2\mathbf{I}$ (i.e. that the random effects are iid), but these assumption would not be reasonable in most applications.

5.5 Moments Associated with Model II

Another problem with adding random effects to the model for the hidden process is that it becomes difficult to assess their impact on the model in general, and on the marginal moments in particular.

The expressions for the moments are the same as in (5.7)-(5.9). However, the integration is considerably more difficult in this setting. For instance, consider (5.7). Under Model I, this equation becomes

$$\mathrm{E}[Y_{it}] = \sum_{k=1}^{K} \int_{\mathbf{u}} c'(\eta_{itk}) f(\mathbf{u}; \psi)\, d\mathbf{u}\ \mathrm{P}(Z_{it} = k; \psi).$$

At least in the case where $\{Z_{it}\}$ is stationary, this expression can be easily interpreted. In contrast, under Model II (assuming the random effects appear only in the model for the

hidden process), (5.7) becomes

$$\mathrm{E}[Y_{it}] = \sum_{k=1}^{K} c'(\eta_{itk}) \int_{\mathbf{u}} \mathrm{P}(Z_{it} = k \mid \mathbf{u}, \psi) f(\mathbf{u}; \psi)\, d\mathbf{u}.$$

Even if $\{Z_{it} \mid \mathbf{u}\}$ is stationary, it will be difficult to compute (or approximate) this integral because $\mathrm{P}(Z_{it} = k \mid \mathbf{u}, \psi)$ is a complex function of the transition probabilities when $K > 2$. This is also true of the integral in (5.8).

The expression (5.9) is even harder both to evaluate and interpret when we have random effects in the model for the hidden process. Specifically, we need to integrate the function

$$\mathrm{P}(Z_{it} = \ell, Z_{js} = k \mid \mathbf{u}, \psi) = \begin{cases} \mathrm{P}(Z_{it} = \ell \mid \mathbf{u}, \psi)\ \mathrm{P}(Z_{js} = k \mid \mathbf{u}, \psi), & i \neq j \\ \mathrm{P}(Z_{it} = \ell \mid Z_{is} = k, \mathbf{u}, \psi)\ \mathrm{P}(Z_{is} = k \mid \mathbf{u}, \psi), & i = j \end{cases}.$$

The $(t - s)$-step transition probabilities, $\mathrm{P}(Z_{it} = \ell \mid Z_{is} = k, \mathbf{u}, \psi)$, are, like the stationary transition probabilities, complicated functions of the transition probabilities. Thus, evaluating these integrals – even by approximation methods – does not seem feasible.

In the case where $\{Z_{it} \mid \mathbf{u}\}$ is homogeneous with transition probabilities $\{P^i_{k\ell}\}$, $k, \ell = 1, 2$, and stationary, the stationary distribution has the simple form

$$[\pi^i_1 \quad \pi^i_2] = \left[\frac{P^i_{21}}{P^i_{12} + P^i_{21}} \quad \frac{P^i_{12}}{P^i_{12} + P^i_{21}} \right].$$

Then we may compute approximations to (5.7) and (5.8). However, evaluating (5.9), even approximately, is still problematic because the $(t - s)$-step transition probabilities do not have a simple form.

Interestingly, we are nonetheless able to make statements about $\lim_{t\to\infty} \mathrm{Cov}[Y_{it}, Y_{is}]$ when the hidden Markov chains are homogeneous and stationary. By the same argument given in Section 5.3, we have that

$$\lim_{t\to\infty} \mathrm{Cov}[Y_{it}, Y_{is}] \geq 0$$

if condition (5.11) holds. Again, this will be true if $f(y_{it} \mid z_{it}, \mathbf{u}, \psi)$ is in the Poisson or normal family and $\mathrm{E}[\mathbf{u}] < \infty$. Thus, under these conditions, we see that adding random effects to the model for the hidden process also affects the correlation structure in the expected way (see the discussion in Section 5.3).

As a way of circumventing the problems associated with adding random effects to the transition probabilities, if the underlying Markov chains are homogeneous and stationary, we might instead consider incorporating the random effects into the stationary transition probabilities. We would then be able to compute $\mathrm{E}[Y_{it}]$, $\mathrm{Var}[Y_{it}]$, and $\mathrm{Cov}[Y_{it}, Y_{js}]$, $i \neq j$, explicitly. A simpler (though not explicit) expression for $\mathrm{Cov}[Y_{it}, Y_{is}]$ would also result.

However, it is important to consider the implications of these random effects in terms of the distribution of the transition probabilities.

For example, in the simple case where $K = 2$, it is easy to derive the constraint that the transition probabilities for this model must satisfy, namely

$$P_{12}^i = \frac{\pi_2^i P_{21}^i}{\pi_1^i} \tag{5.14}$$

for each i. It is reasonable to insist that, in the absence of prior information, all the transition probabilities have the same distribution. But, it is unclear whether there exists a distribution for $\mathbf{u}$ so that this is possible, given the constraint (5.14). Even if an appropriate distribution does exist, extending this method to the case where $K > 2$ seems unrealistic because the system of equations relating the transition probabilities to the stationary transition probabilities becomes increasingly complex with K.

In conclusion, we may add random effects to the transition probabilities, but are unlikely to be able to evaluate any moments. Adding random effects to the stationary transition probabilities results in tractable expressions in (5.7), (5.8), and (5.9) for $i \neq j$, but may not be sensible in terms of the resulting distributions of the transition probabilities. Of course, these problems are not unique to HMMs: incorporating random effects in Markov chains is equally difficult.

5.6 Summary

In summary, the addition of random effects to the conditional model for the observed data results in a tractable, interpretable model. Marginal moments can be evaluated or approximated, and parameter estimation is not much more complicated than in the case of a standard HMM. The primary change resulting from the inclusion of random effects is the need for integration (in most cases, numerical) of the likelihood. In addition, we will need to estimate the extra parameters associated with the distributions of the random effects, but this task should be straightforward as long as we have a suitable number of patients.

On the other hand, including random effects in the model for the hidden process is hard from the perspective of both interpretation and estimation. Typically, marginal moments cannot even be approximated because of the complex nature of the integrands involved. Parameter estimation also may be problematic due to the high-dimensional integrals, the dependence structure of the random effects, and the large number of unknown parameters.

However, the limitations of Model II are perhaps not as serious as one might imagine.

As discussed in Section 2.5, we have relatively little information about the parameters of the hidden process. Extending the model to allow patient-to-patient differences on this level may explain very little additional variation in the observed data, and hence may not be worthwhile from a statistical standpoint. Moreover, capturing inter-patient heterogeneity in the hidden processes is still possible by incorporating covariates in this part of the model (though, of course, the issue of the reasonableness of the resulting marginal moments remains).

Chapter 6

Hypothesis Testing

In Chapter 5, we presented two models for multiple processes, Models I and II, as well as some theory regarding their interpretation and estimation. We did not, however, discuss the properties of these estimators. In the present chapter, we address this issue, along with the problem of hypothesis testing. We begin, in Sections 6.1 and 6.2, by considering model identifiability and the asymptotic properties of the MLEs, respectively. Although we do not provide formal results, our discussion will suggest that standard tests are appropriate for most hypotheses concerning the model parameters. We then move on to the focus of this chapter: tests of the significance of the variance components of the random effects. In particular, we wish to develop a test of whether the additional complexity associated with introducing random effects into HMMs is warranted. Hypotheses of this nature do not lend themselves to standard testing procedures, and pose quite different challenges.

Specifically, the difficulty surrounding tests of the variance components is that the null hypothesis puts at least one parameter on the boundary of the parameter space. Without loss of generality, we can assume that the variance-covariance matrix of the random effects, $\mathbf{Q}$, is a function of some d-dimensional parameter D, $D \geq 0$, and that $\mathbf{Q}(D)$ is a matrix of zeroes if $D = 0$. One of the conditions for the validity of the usual Wald or likelihood ratio test (LRT) is that the true parameter be in the interior of the parameter space. In the test with null hypothesis $D = 0$, this condition clearly does not hold.

There is a substantial literature on the asymptotic distribution of the LRT statistic when the true parameter is on the boundary of the parameter space and when the observed data are iid. In this case, this distribution can often be determined analytically. For example, when the observed data are iid and normally distributed, Self & Liang (1987) show that, for a variety of boundary problems, the LRT statistic is asymptotically a mixture of χ^2 distributions. Stram & Lee (1994) apply these results to the LRT of the variance components

in a linear mixed model. Feng & McCulloch (1992) extend this work to the general iid case, and show that a variation of the LRT statistic (based on an enlarged parameter space) has the usual asymptotic χ^2 distribution.

However, in other settings, including that where the observed data are not independent, the distribution of the LRT statistic when some parameters are on the boundary is far more complicated. For this reason, other tests are used more commonly in the GLMM context. In particular, Jacqmin-Gadda & Commenges (1995), Lin (1997), and Hall & Præstgaard (2001) have proposed tests based on the score vector

$$\left(\frac{\partial}{\partial D_k} \log \mathcal{L}(\psi) \Big|_{D=0} \right), \quad k = 1, \ldots, d. \tag{6.1}$$

This test has also been suggested for detecting overdispersion (as captured by a random effect) in Poisson and binomial regression models (e.g. Dean 1992). In both cases, conditional on the random effects, the observations are independent with distributions in the exponential family. It turns out that this conditional independence leads to a nice form for (6.1), and hence its asymptotic distribution is easily derived. This is not true in the HMM setting, but we can still use this framework to conduct tests of the significance of the variance components. We provide a detailed explanation in Section 6.3.

We apply the theory developed in the previous and current chapters to two data sets: lesion count data collected on the Vancouver PRISMS patients (PRISMS Study Group 1998), and faecal coliform counts in sea water (Turner *et al.* 1998). These analyses are presented in Section 6.4. We include the results of a small simulation study that investigates the performance of our variance component test in Section 6.5.

6.1 Identifiability of Models I and II

In Sections 2.3 and 3.2, we raised the issue of the identifiability of a HMM for a single process, along with the fact that this topic has not been adequately addressed in the literature. Nonetheless, identifiability is a critical concern, since it is a required condition for the usual asymptotic properties of the parameter estimates to hold (e.g. Leroux 1992a; Bickel *et al.* 1998; Douc & Matias 2001).

Likewise, the identifiability of Models I and II is also needed in order to study the asymptotic properties of their MLEs. However, the necessary theory is even less well-developed in the multiple process setting. In fact, we are aware of only one reference on this subject. Wang & Puterman (1999) show that Poisson HMMs with (possibly time-varying) covariates in the conditional model for the observed data are identifiable if for each patient

i. The model matrix for the covariates is of full rank.

ii. The Poisson means associated with each hidden state at each time point are distinct.

The method of proof used to establish this result follows Leroux (1992a), and relies on the fact that the marginal distribution of each observation is an identifiable mixture distribution. This method easily extends to the case where the conditional distribution of the observed data is in the exponential family, (i) holds, and the support points of the mixing distributions are distinct (see the discussion in Section 3.1).

In contrast, in the case where the model contains random effects, the approach of Wang & Puterman (1999) does not readily apply. Consider Model I. The marginal distribution of Y_{it} is

$$f(y_{it};\psi) = \sum_{k=1}^{K} \mathrm{P}(Z_{it} = k;\psi) \int f(y_{it} \mid Z_{it} = k, \mathbf{u}, \psi) f(\mathbf{u};\psi)\, d\mathbf{u},$$

which is a finite mixture. Prakasa Rao (1992) provides a summary of sufficient conditions for the identifiability of both finite and arbitrary mixture models. However, it may be difficult to verify these conditions for Model I because the kernel $\int f(y_{it} \mid Z_{it} = k, \mathbf{u}, \psi) f(\mathbf{u};\psi)\, d\mathbf{u}$ will typically not be a standard distribution. Similarly, the marginal distribution of Y_{it} under Model II is

$$f(y_{it};\psi) = \sum_{k=1}^{K} \int f(y_{it} \mid Z_{it} = k, \mathbf{u}, \psi) \mathrm{P}(Z_{it} = k \mid \mathbf{u}, \psi)\ f(\mathbf{u};\psi)\ du_i.$$

This distribution is also technically a mixture (with both discrete and continuous mixing distributions), but its complicated form makes verifying identifiability even more challenging.

Clearly, further research is required to establish sufficient conditions for the identifiability of this new class of models. In the meantime, for the purposes of the discussion in this section, we will follow Bickel *et al.* (1998) and Douc & Matias (2001) and simply assume that the models in question are identifiable.

6.2 Asymptotic Properties of the MLEs of Models I and II

At this time, we have not formally established the asymptotic properties of the MLEs of Models I and II. However, Bradley & Gart (1962) develop conditions for the consistency and asymptotic normality of MLEs in a very general setting. We describe these conditions in this section, since they illustrate the work that may be involved in showing that these properties hold in the context of HMMs with random effects.

In addition to model identifiability, these authors require that the data be sampled from independent populations. They consider two broad classes of problems:

i. The number of populations is finite, with the number of observations on each approaching infinity.

ii. The number of populations approaches infinity, with a finite number of observations on each.

In the context of MS/MRI clinical trials, we would expect to follow each patient for a limited time period, but to collect data on large numbers of patients. Thus, we are primarily interested in the second scenario.

The conditions that Bradley & Gart (1962) impose are chosen to ensure that for all k, ℓ, and m, as $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\psi_k}\log f(\mathbf{y}_i;\psi) = o_p(1), \tag{6.2}$$

and that there exists a positive definite matrix with (k,ℓ)th entry $\mathbf{J}_{k\ell}(\psi)$ and with finite determinant, as well as a constant $C < \infty$, such that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\psi_k\partial\psi_\ell}\log f(\mathbf{y}_i;\psi) + \mathbf{J}_{k\ell}(\psi) = o_p(1) \tag{6.3}$$

$$\text{and} \quad \lim_{N\to\infty} \mathrm{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^3}{\partial\psi_k\partial\psi_\ell\partial\psi_m}\log f(\mathbf{y}_i;\psi)\right| > C\right) = 0. \tag{6.4}$$

Unfortunately, the conditions required for (6.2)-(6.4) are hard to verify in the HMM setting due to the complicated form of the likelihood and its derivatives, especially when random effects are involved. However, it is very likely that the MLEs have the usual asymptotic properties in the case where we have N independent patients with $n_i \leq n < \infty$ for all i, and $N \to \infty$. For this reason, in this context, we feel comfortable conducting standard Wald tests, using the observed information to approximate the standard errors of the parameter estimates.

6.3 Variance Component Testing

In this section, we develop a test, based on the score function, for testing the significance of the variance components of the random effects in our proposed models for multiple processes. For our purposes in this section, we specify the values of the variance components under the

null hypothesis, but the other model parameters remain unknown. Thus, we treat the other model parameters as nuisance parameters. Little is known about the behaviour of the score function in the multiple HMM setting. For single HMMs, asymptotic properties of the score function have been established (Bickel *et al.* 1998; Douc & Matias 2001), but not in the case when the true parameters lie on the boundary, nor when there are nuisance parameters.

Jacqmin-Gadda & Commenges (1995) provide a general result about the asymptotic distribution of a statistic based on the function $S_{(N)}(\xi) \equiv \sum_{i=1}^{N} S_i(\xi)$, where $\{S_i(\xi)\}$ are independent, and ξ is a vector of nuisance parameters. In their derivation of this distribution, the conditions that they impose ensure the validity of the central limit theorem and law of large numbers that these authors apply to certain key sums. When $S_{(N)}(\xi)$ has the form (6.1) and the observed data arise from a GLMM with a single, patient-specific random effect, they show that these conditions are easily verified. Perhaps not surprisingly given our discussion in Section 6.2, the same is not true of these conditions in the HMM setting.

Thus, while we follow the approach of Jacqmin-Gadda & Commenges (1995) in the formulation of our test procedure, we will require different results in order to verify the necessary regularity conditions. In particular, we will need bounds on the expectation of the product of derivatives of $\log f(\mathbf{y}_i \mid u_i, \psi)$. To this end, we use the methods of Bickel *et al.* (2002), henceforth called BRR. These authors consider the case where $\{Y_t\}$ is a single, stationary HMM, and establish bounds on the expectation of the derivatives of $\log f(\mathbf{y}; \psi)$ for a given sample size n. We show that the techniques used in the development of these bounds are also applicable to our problem.

The beauty of BRR's work is that their conditions depend only on the conditional distribution of the observed data and the model for the hidden process, not on the full likelihood. For this reason, verifying these conditions is reasonably easy. BRR's bounds are derived with an eye towards determining the properties of the likelihood and its derivatives as $n \to \infty$, and are quite refined. We are working in a much simpler context (i.e. where observations on different patients are independent, and, for each i, n_i remains bounded while $N \to \infty$) and so presumably far cruder results would suffice. However, we are unaware of the existence of such results at this time.

Our test is valid for a variety of models, and we discuss these at the end of this section. For the purposes of developing the relevant theory, we present only the simple example of Model I with no covariates, but with one patient-specific random effect, u_i, in the conditional model for the observed data (see, e.g., (5.5)). In other words, we assume that $\{u_i\}$ are iid with mean zero, and that

$$f(y_{it} \mid Z_{it} = k, u_i, \psi) = \exp\{(y_{it}\eta_{itk} - c(\eta_{itk}))/a(\phi) + d(y_{it}, \phi)\},$$

with $\eta_{itk} = \tau_k + u_i$. Recall that, for Model I, the hidden Markov chains are assumed homogeneous and stationary with transition probabilities common to all patients.

Without loss of generality, we model the random effect as $u_i = \sqrt{D}v_i$, where the $\{v_i\}$ are independent and identically distributed with zero mean and unit variance. One advantage of our test is that we do not need to specify the distribution of u_i.

Let ξ be the p-dimensional vector of all the model parameters except D, the variance of the random effect. Then, we will form our test statistic based on $S_{(N)}(\xi) \equiv \sum_{i=1}^{N} S_i(\xi)$, where

$$S_i(\xi) \equiv \frac{1}{2}\left\{ \frac{\partial^2}{\partial u_i^2} \log f(\mathbf{y}_i \mid u_i, \psi)\Big|_{u_i=0} + \left[\frac{\partial}{\partial u_i} \log f(\mathbf{y}_i \mid u_i, \psi)\Big|_{u_i=0} \right]^2 \right\}, \tag{6.5}$$

$i = 1, \ldots, N$. Other authors have claimed that $S_i(\xi)$ is equal to $\frac{\partial}{\partial D} \log f(\mathbf{y}_i; \psi)\Big|_{D=0}$, using either L'Hôpital's rule (e.g. Chesher 1984; Jacqmin-Gadda & Commenges 1995) or a Taylor series expansion of $f(\mathbf{y}_i \mid u_i, \psi)$ about u_i (e.g. Dean 1992; Lin 1997) as justification. Both methods depend on (unstated) regularity conditions for their validity. Specifically, these conditions permit, in the case of the former, two interchanges of derivatives and integrals, and in the case of the latter, the interchange of a derivative and an infinite summation. It is not clear that these conditions hold; as usual, they are particularly difficult to verify in our setting. For this reason, we will not make this claim, but simply note that it may be true in some cases.

We now introduce some notation, mostly borrowed from BRR. All expectations and probabilities are calculated under the null hypothesis that $D = 0$.

First, we use the operator $\mathbf{D}_m$ to indicate a partial derivative of order m, taken with respect to any combination of u_i and the parameters $\{\xi_k\}$. The bounds that we develop on the expectation of these derivatives will not depend on the particular combination of these variables, hence the generality of our notation.

Second, let $\mathbf{Z}_+$ be the set of positive integers. We define

$$\mathcal{J} \equiv \{(J_1, \ldots, J_k) : k > 0,\ J_j \in \mathbf{Z}_+,\ j = 1, \ldots, k\}$$

and

$$\mathcal{J}_m^n(k) \equiv \{(I_1, \ldots, I_k) : I_j \in \mathbf{Z}_+,\ m \le I_j \le n,\ j = 1, \ldots, k\}.$$

For $\mathbf{J} \in \mathcal{J}$, let $|\mathbf{J}|$ be the length of $\mathbf{J}$, and let $|\mathbf{J}|_+ = \sum_{j=1}^{|\mathbf{J}|} J_j$. In addition, let $\mathcal{J}^+(k) \equiv \{\mathbf{J} \in \mathcal{J} : |\mathbf{J}|_+ = k\}$. Also, we define $U(\mathbf{I})$ as the ordered set of unique elements of the vector $\mathbf{I}$.

Finally, we define quantities $C_m^i(y_{it}, \psi)$ and $B_{m_1 \cdots m_d}^i(\psi)$. These quantities relate to the conditional distribution $h_i(t) \equiv \log f(y_{it}, z_{it} \mid \mathbf{y}_{i1}^{t-1}, \mathbf{z}_{i1}^{t-1}, u_i, \psi)$, where $\mathbf{y}_{i1}^t \equiv (y_{i1}, \ldots, y_{it})$,

and are critical in determining our bounds of stated interest. They are somewhat difficult to understand in isolation, but their purpose should become clear in the context of the proof of Theorem 6.1.

In particular, we define

$$C_m^i(y_{it},\psi) \equiv \max_{\mathbf{D}_m}\max_{k,\ell}\left\{\left|\mathbf{D}_m \log P_{k\ell}\right| + \left|\mathbf{D}_m \log f(y_{it}|Z_{it}=k,u_i,\psi)\Big|_{u_i=0}\right| + \left|\mathbf{D}_m \log \pi_k\right|\right\},$$

where $\max_{\mathbf{D}_m}$ should be interpreted as the maximum over all derivatives $\mathbf{D}_m$. Our model assumptions imply that $h_i(t) = \log P_{z_{i,t-1},z_t} + \log f(y_{it} \mid z_{it}, u_i, \psi)$, so $C_m^i(y_{it},\psi)$ is a bound for the mth order derivatives of $h_i(t)$.

Letting the empty product be 1, we define

$$B_m^i(\psi) \equiv \max\left\{\prod_{s=1}^{r}\mathrm{E}\left(\prod_{\substack{j\in\{1,\ldots,|\mathbf{J}|\}:\\ t_s=I_j}}\frac{C_{J_j}^i(Y_{i,t_s},\psi)}{J_j!}\,\middle|\, Z_{i,t_s}=z_{i,t_s}\right) : \mathbf{J}\in\mathcal{J}^+(m),\right.$$
$$\left.\mathbf{I}\in\mathcal{J}_1^{n_i}(|\mathbf{J}|),\ r=|U(\mathbf{I})|,\ (t_1,\ldots,t_r)=U(\mathbf{I}),\ z_{i,t_s}\in\{1,\ldots,K\}\right\}.$$

The value $B_m^i(\psi)$ will be used to bound the expectation of the product of derivatives of $h_i(t)$. We also need the quantities $B_{m_1m_2}^i(\psi)$, $B_{m_1m_2m_3}^i(\psi),\ldots$, where

$$B_{m_1m_2}^i(\psi) \equiv \max\left\{\prod_{s=1}^{r}\mathrm{E}\left(\prod_{\substack{j_1\in\{1,\ldots,|\mathbf{J}^1|\}:\\ t_s=I_{j_1}^1}}\frac{C_{J_{j_1}^1}^i(Y_{i,t_s},\psi)}{J_{j_1}^1!}\cdot\prod_{\substack{j_2\in\{1,\ldots,|\mathbf{J}^2|\}:\\ t_s=I_{j_2}^2}}\frac{C_{J_{j_2}^2}^i(Y_{i,t_s},\psi)}{J_{j_2}^2!}\,\middle|\, Z_{i,t_s}=z_{i,t_s}\right) :\right.$$
$$\mathbf{J}^1\in\mathcal{J}^+(m_1),\ \mathbf{J}^2\in\mathcal{J}^+(m_2),\mathbf{I}^1\in\mathcal{J}_1^{n_i}(|\mathbf{J}^1|),\mathbf{I}^2\in\mathcal{J}_1^{n_i}(|\mathbf{J}^2|),$$
$$\left.r=|U(\mathbf{I}^1\cup\mathbf{I}^2)|,\ (t_1,\ldots,t_r)=U(\mathbf{I}^1\cup\mathbf{I}^2),\ z_{i,t_s}\in\{1,\ldots,K\}\right\},$$

and $B_{m_1\cdots m_d}^i(\psi)$, $d>2$, is defined similarly.

We require the following conditions for the remainder of this chapter.

Condition 1. The transition probabilities satisfy $P_{k\ell}>0$ for all $k,\ell=1,\ldots,K$.

Condition 2. The function $\log f(y_{it} \mid z_{it}, u_i, \psi)$ has M continuous derivatives, $M\geq 4$, with respect to u_i and $\{\xi_k\}$ at the true value of these parameters.

Condition 3. For all $m_1,\ldots,m_d\leq M$, $d<\infty$, and for all i,

$$\sum_{i=1}^{\infty}\frac{1}{i^2}B_{m_1\cdots m_d}^i(\psi)<\infty.$$

Condition 4. There exists a fixed value $n<\infty$ such that $n_i\leq n$ for all $i=1,2,\ldots$.

REMARK. Conditions 1–3 are essentially those assumed by BRR, with a slight change in Condition 2 to account for the incorporation of the random effects in our model, and in Condition 3 to bound terms involving $B^i_{m_1 \cdots m_d}(\psi)$ in a specific way. Conditions 1, 2, and 4 are easy to verify. Theorem 6.2 below provides sufficient conditions for Condition 3 when the conditional distribution for the observed data is one of several common choices.

Before stating our main theorem, we present some preliminary results. Our first, which we prove in Appendix B.2, extends BRR's Theorem 2.1 to establish bounds on the expectation of the product of derivatives of $\log f(\mathbf{y}_i \mid u_i, \psi)$. We define

$$\mathbf{D}^i_m \equiv \mathbf{D}_m \log f(\mathbf{y}_i \mid u_i, \psi)\Big|_{u_i=0}.$$

Theorem 6.1 *Under Conditions 1 and 2, for all i, and all $m_1, \ldots, m_d \leq M$, $d < \infty$,*

$$\mathrm{E}\left|\mathbf{D}^i_{m_1} \times \cdots \times \mathbf{D}^i_{m_d}\right| \leq A_{m_1 \cdots m_d} m_1! \times \cdots \times m_d!\ n_i^d B^i_{m_1 \cdots m_d}(\psi),$$

where $A_{m_1 \cdots m_d}$ is a finite constant depending only on the transition probabilities.

Given the complicated form of the expression $B^i_{m_1 \cdots m_d}(\psi)$, it is of interest to identify models where Condition 3 holds. This builds on the work of BRR, who simply assume that $B^i_{m_1}(\psi) < \infty$. We also require conditions such that $\sum_{i=1}^{\infty} \frac{1}{i^2} B^i_{m_1 \cdots m_d}(\psi) < \infty$ for finite values of d. Theorem 6.2, which is proved in Appendix B.3, shows that this property holds for four common distributions for the conditional model for the observed data. The same line of argument also allows the investigation of the properties of $B^i_{m_1 \cdots m_d}(\psi)$ for other choices of distribution for the conditional model.

Theorem 6.2 *Assume that Conditions 1 and 4 hold. In addition, assume that ξ belongs to a bounded parameter space. Then for all i, and all $m_1, \ldots, m_d < \infty$, $d < \infty$,*

$$\sum_{i=1}^{\infty} \frac{1}{i^2} B^i_{m_1 \cdots m_d}(\psi) < \infty$$

when $f(y_{it}|z_{it}, u_i, \psi)$ is in the Poisson, binomial, normal, or gamma family of distributions.

We now present a series of technical results that will be required in the proof of our principal theorem, Theorem 6.3. These results will allow us to apply the law of large numbers and the central limit theorem in the same manner as Jacqmin-Gadda & Commenges (1995), and hence to determine the asymptotic distribution of our test statistic.

Lemma 6.1 *Under Conditions 1–4, for all* $k, \ell = 1, \ldots, p$,

$$\mathrm{E}\left|\frac{\partial S_i(\xi)}{\partial \xi_k}\right| < \infty \quad \textit{for} \quad i = 1, \ldots, N \tag{6.6}$$

$$\mathrm{E}\left|\frac{\partial^2 S_i(\xi)}{\partial \xi_k \partial \xi_\ell}\right| < \infty \quad \textit{for} \quad i = 1, \ldots, N \tag{6.7}$$

$$\sum_{i=1}^{\infty} \frac{1}{i^2} \operatorname{Var}\left\{\frac{\partial S_i(\xi)}{\partial \xi_k}\right\} < \infty \tag{6.8}$$

$$\sum_{i=1}^{\infty} \frac{1}{i^2} \operatorname{Var}\left\{\frac{\partial^2 S_i(\xi)}{\partial \xi_k \partial \xi_\ell}\right\} < \infty \tag{6.9}$$

Proof. Consider (6.6). We have that

$$\begin{aligned}
\mathrm{E}\left|\frac{\partial S_i(\xi)}{\partial \xi_k}\right| &= \frac{1}{2}\mathrm{E}\left|\frac{\partial}{\partial \xi_k}\left\{\frac{\partial^2}{\partial u_i^2}\log f(\mathbf{y}_i \mid u_i, \psi)\Big|_{u_i=0}\right\} + \frac{\partial}{\partial \xi_k}\left[\frac{\partial}{\partial u_i}\log f(\mathbf{y}_i \mid u_i, \psi)\Big|_{u_i=0}\right]^2\right| \\
&= \frac{1}{2}\mathrm{E}\left|\mathbf{D}_3^i + 2\,\frac{\partial^2}{\partial \xi_k \partial u_i}\log f(\mathbf{y}_i \mid u_i, \psi)\Big|_{u_i=0} \cdot \frac{\partial}{\partial u_i}\log f(\mathbf{y}_i \mid u_i, \psi)\Big|_{u_i=0}\right| \\
&= \frac{1}{2}\mathrm{E}\left|\mathbf{D}_3^i + 2\mathbf{D}_2^i\mathbf{D}_1^i\right|.
\end{aligned}$$

Thus, by Theorem 6.1 and Condition 3, we see that $\mathrm{E}\left|\frac{\partial S_i(\xi)}{\partial \xi_k}\right| < \infty$ for all i and k.

In the same way, we can show that $\mathrm{E}\left|\frac{\partial^2}{\partial \xi_k \partial \xi_\ell} S_i(\xi)\right|$, $\operatorname{Var}\left\{\frac{\partial}{\partial \xi_k} S_i(\xi)\right\}$, and $\operatorname{Var}\left\{\frac{\partial^2}{\partial \xi_k \partial \xi_\ell} S_i(\xi)\right\}$ are expectations of sums of products of $\{\mathbf{D}_m^i\}$, $m \leq 4$, and hence are finite for all i, k, and ℓ by Theorem 6.1. Now, let L_k^i and $M_{k\ell}^i$ be finite constants such that $\operatorname{Var}\left\{\frac{\partial}{\partial \xi_k} S_i(\xi)\right\} < L_k^i$ and $\operatorname{Var}\left\{\frac{\partial^2}{\partial \xi_k \xi_\ell} S_i(\xi)\right\} < M_{k\ell}^i$, $i = 1, \ldots, N$, $k, \ell = 1, \ldots, K$. For fixed values of k and ℓ, these moments vary with i only because n_i may be different for each i. Since n_i can take on only values from the finite set $\{1, \ldots, n\}$, we see that $\operatorname{Var}\left\{\frac{\partial}{\partial \xi_k} S_i(\xi)\right\}$ and $\operatorname{Var}\left\{\frac{\partial^2}{\partial \xi_k \xi_\ell} S_i(\xi)\right\}$ can take on at most n distinct values. Thus, there exist finite constants L_k and $M_{k\ell}$ (independent of N) such that for all $i = 1, \ldots, N$,

$$\begin{aligned}
\operatorname{Var}\left\{\frac{\partial S_i(\xi)}{\partial \xi_k}\right\} &< L_k \\
\text{and } \operatorname{Var}\left\{\frac{\partial^2 S_i(\xi)}{\partial \xi_k \xi_\ell}\right\} &< M_{k\ell}.
\end{aligned}$$

Thus, it is clear that (6.8) and (6.9) hold for all k and ℓ. □

In the proof of the next results, we will require a law of large numbers for non-identically distributed random variables. We will use the law that if $X_1, X_2, \ldots$ are independent with mean zero, $\{c_i\}$ are non-negative constants with $c_i \to \infty$, and $\sum_{i=1}^{\infty} \mathrm{E}[X_i^2]/c_i^2 < \infty$, then

$$\frac{X_1 + \cdots + X_N}{c_N} \to 0 \text{ a.s.}$$

(See, e.g., Breiman 1968, Theorem 3.27.) For our purposes, we will take $c_i \equiv i$. We will henceforth refer to this law as (**L1**).

Lemma 6.2 *Assume that Conditions 1–4 hold, and let $\hat{\xi}$ be a MLE of ξ under the null hypothesis that $D = 0$. Then $\hat{\xi}$ converges in probability to ξ as $N \to \infty$.*

Proof. Let $f_0(\mathbf{y}_i; \xi)$ be the density of patient i's observations under the null hypothesis, i.e.

$$f_0(\mathbf{y}_i; \xi) = \sum_{\mathbf{z}_i} \left\{ \pi_{z_{i1}} f(y_{i1} \mid z_{i1}, u_i = 0, \xi) \prod_{t=1}^{n_i} P_{z_{i,t-1}, z_{it}} f(y_{it} \mid z_{it}, u_i = 0, \xi) \right\}.$$

This is just the distribution of a stationary HMM. Now, using the method of Bradley & Gart (1962), the following requirements are sufficient to ensure the properties (6.2)-(6.4):

1. For almost all $\mathbf{y}_i$ and for all ξ,

$$\frac{\partial}{\partial \xi_j} \log f_0(\mathbf{y}_i; \xi), \quad \frac{\partial^2}{\partial \xi_j \partial \xi_k} \log f_0(\mathbf{y}_i; \xi), \quad \text{and} \quad \frac{\partial^3}{\partial \xi_j \partial \xi_k \partial \xi_\ell} \log f_0(\mathbf{y}_i; \xi)$$

   exist for all $j, k, \ell = 1, \ldots, p$, and for all $i = 1, \ldots, N$.

2. The following results hold for all ξ and all $j, k, \ell = 1, \ldots, p$ as $N \to \infty$:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \xi_j} \log f_0(\mathbf{y}_i; \xi) = o_p(1) \tag{6.10}$$

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{\partial^2}{\partial \xi_j \partial \xi_k} \log f_0(\mathbf{y}_i; \xi) - E\left[\frac{\partial^2}{\partial \xi_j \partial \xi_k} \log f_0(\mathbf{y}_i; \xi) \right] \right\} = o_p(1) \tag{6.11}$$

$$\lim_{N \to \infty} \mathrm{P}\left(\left| \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^3}{\partial \xi_j \partial \xi_k \partial \xi_\ell} \log f_0(\mathbf{y}_i; \xi) \right| > C \right) = 0, \tag{6.12}$$

   where C is some finite constant.

From Conditions 1 and 2, it is clear that the first requirement is satisfied. For the second requirement, we use (**L1**) to show all three results. This law will, in fact, give us convergence a.s., which is a stronger result than necessary.

In particular, for (6.10), the fact that $\mathrm{E}\,|\mathbf{D}_1^i| < \infty$ allows us to interchange the expectation and derivative to compute $\mathrm{E}\left[\frac{\partial}{\partial \xi_j} \log f_0(\mathbf{y}_i; \xi)\right] = 0$ for all i. Conditions 1–4 and Theorem 6.1 imply that $\mathrm{E}[(\mathbf{D}_1^i)^2]$ has a finite bound, independent of i. Denote this bound by C. Then

$$\sum_{i=1}^{\infty} \frac{1}{i^2} \mathrm{E}\left[\left(\frac{\partial}{\partial \xi_j} \log f_0(\mathbf{y}_i; \xi) \right)^2 \right] < C \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty,$$

so the conditions of (**L1**) are satisfied, and (6.10) holds. Similarly, (**L1**) applies in the case of (6.11), since the expected values of the summands are zero, and $\mathrm{E}[(\mathbf{D}_2^i)^2] - (\mathrm{E}[\mathbf{D}_2^i])^2$ has a finite bound, independent of i. For (6.12), we first consider the expression

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\partial^3}{\partial\xi_j\partial\xi_k\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi) - \mathrm{E}\left[\frac{\partial^3}{\partial\xi_j\partial\xi_k\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi)\right]\right\}. \tag{6.13}$$

Clearly, the expected values of the summands are zero. Since $\mathrm{E}[(\mathbf{D}_3^i)^2] - (\mathrm{E}[\mathbf{D}_3^i])^2$ has a finite bound, independent of i, (**L1**) implies that (6.13) converges to zero a.s. Since, for fixed values of j, k, and ℓ, the terms $\mathrm{E}\left[\frac{\partial^3}{\partial\xi_j\partial\xi_k\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi)\right]$, $i = 1, 2, \ldots$ take on only a finite number of values (as argued in the proof of Lemma 6.1), the second sum in (6.13) is bounded. Therefore, the first term must also be bounded a.s. These convergence results guarantee that the arguments of Bradley & Gart (1962) apply, and thus we have verified the consistency of $\hat{\xi}$. □

Lemma 6.3 *Assume that Conditions 1–4 hold, and let $\hat{\xi}$ be a MLE of ξ under the null hypothesis that $D = 0$. Then $\sqrt{N}(\hat{\xi} - \xi)$ has the form*

$$\sqrt{N}(\hat{\xi} - \xi) = \sqrt{N}\ \mathbf{I}_{\xi\xi(N)}^{-1}\sum_{i=1}^{N}\mathbf{U}_i + o_p(1), \tag{6.14}$$

where each $\mathbf{U}_i$ is a random vector with kth element $\frac{\partial}{\partial\xi_k}\log f_0(\mathbf{y}_i;\xi)$, $k = 1,\ldots,p$, and variance-covariance matrix $\mathbf{I}_{\xi\xi i}$. We write $\mathbf{I}_{\xi\xi(N)} \equiv \sum_{i=1}^{N}\mathbf{I}_{\xi\xi i}$.

Proof. Let $\mathbf{H}_{(N)}(\xi)$ be the matrix with entries $\sum_{i=1}^{N}\frac{\partial^2}{\partial\xi_k\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi)$. Under Conditions 1 and 2, we can do a first-order Taylor series expansion of $\sum_{i=1}^{N}\frac{\partial}{\partial\xi_k}\log f_0(\mathbf{y}_i;\hat{\xi})$ about ξ, for $k = 1,\ldots,p$, yielding

$$\hat{\xi} - \xi = [-\mathbf{H}_{(N)}(\xi^\dagger)]^{-1}\sum_{i=1}^{N}\mathbf{U}_i,$$

where $|\xi^\dagger - \xi| < |\hat{\xi} - \xi|$. By Lemma 6.2, $\hat{\xi} = \xi + o_p(1)$, so that $\xi^\dagger = \xi + o_p(1)$ as well. By the continuity of $\frac{\partial^2}{\partial\xi_k\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi)$,

$$\hat{\xi} - \xi = \left[-\frac{1}{N}\mathbf{H}_{(N)}(\xi) + o_p(1)\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{U}_i\right]. \tag{6.15}$$

As shown in the proof of (6.11),

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\partial^2}{\partial\xi_k\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi) - \mathrm{E}\left[\frac{\partial^2}{\partial\xi_k\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi)\right]\right\} \to 0 \quad a.s.$$

By Conditions 3 and 4 and Theorem 6.1, we can interchange the expectation and derivative to obtain

$$\mathrm{E}\left[\frac{\partial^2}{\partial\xi_k\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi)\right] = -\mathrm{E}\left[\frac{\partial}{\partial\xi_k}\log f_0(\mathbf{y}_i;\xi)\ \frac{\partial}{\partial\xi_\ell}\log f_0(\mathbf{y}_i;\xi)\right],$$

so

$$\frac{1}{N}\left(\mathbf{H}_{(N)}(\xi) + \mathbf{I}_{\xi\xi(N)}\right) = o_p(1).$$

Substituting this result into (6.15), we obtain

$$\begin{aligned}\sqrt{N}(\hat{\xi} - \xi) &= \sqrt{N}\left[\frac{1}{N}\mathbf{I}_{\xi\xi(N)} + o_p(1)\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{U}_i\right] \\ &= \sqrt{N}\,\mathbf{I}_{\xi\xi(N)}^{-1}\sum_{i=1}^{N}\mathbf{U}_i + o_p(1)\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{U}_i. \qquad (6.16)\end{aligned}$$

We now demonstrate that the second term of (6.16) is $o_p(1)$. It suffices to show that $\sum_{i=1}^{N}\mathbf{U}_i/\sqrt{N} = O_p(1)$. Let V_{ij} be the jth element of $\mathbf{U}_i$, and fix a value of $\epsilon > 0$ and $C > 0$. By Chebyshev's inequality,

$$\begin{aligned}\mathrm{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}V_{ij}\right| \geq C\right) &\leq \frac{\mathrm{Var}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}V_{ij}\right]}{C^2} \\ &= \frac{\sum_{i=1}^{N}\mathrm{Var}[V_{ij}]}{NC^2}.\end{aligned}$$

As argued in the proof of (6.10), $\mathrm{E}[V_{ij}] = 0$, and $\mathrm{Var}[V_{ij}]$ has a finite bound independent of i, say γ. Then,

$$\begin{aligned}\mathrm{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}V_{ij}\right| \geq C\right) &\leq \frac{N\gamma}{NC^2} \\ &= \frac{\gamma}{C^2}.\end{aligned}$$

Since C is arbitrary, we can choose a value large enough that

$$\mathrm{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}V_{ij}\right| \geq C\right) \leq \epsilon.$$

This statement implies that $\sum_{i=1}^{N}\mathbf{U}_i/\sqrt{N} = O_p(1)$, as required. □

We will need the following quantities for our next theorem. Define $S_{(N)}(\xi) \equiv \sum_{i=1}^{N}S_i(\xi)$ as in (6.5). Let $I_i \equiv \mathrm{Var}[S_i(\xi)]$, and let $I_{(N)} \equiv \sum_{i=1}^{N}I_i$. Let $\mathbf{J}_{(N)}$ and $\mathbf{K}_{(N)}$ be column vectors defined by $\mathbf{J}_{(N)} \equiv \sum_{i=1}^{N}\mathrm{E}\left[\frac{\partial}{\partial\xi}S_i(\xi)\right]$ and $\mathbf{K}_{(N)} \equiv \sum_{i=1}^{N}\mathrm{E}\left[S_i(\xi)\mathbf{U}_i\right]$. Finally, define

$$I_{c(N)} \equiv I_{(N)} + \mathbf{J}_{(N)}^T\mathbf{I}_{\xi\xi(N)}^{-1}\mathbf{J}_{(N)} + 2\mathbf{K}_{(N)}^T\mathbf{I}_{\xi\xi(N)}^{-1}\mathbf{J}_{(N)}. \qquad (6.17)$$

Theorem 6.3 *Using the above definitions, under Conditions 1–4,*

$$S_{(N)}(\hat{\xi})/I_{c(N)}^{\frac{1}{2}} \qquad (6.18)$$

has an asymptotically standard normal distribution as $N \to \infty$.

Proof. Following Jacqmin-Gadda & Commenges (1995), we expand $S_{(N)}(\hat{\xi})$ in a Taylor series about ξ to obtain

$$S_{(N)}(\hat{\xi}) = \sum_{i=1}^{N} S_i(\xi) + N(\hat{\xi} - \xi)^T \left\{ N^{-1} \sum_{i=1}^{N} \frac{\partial}{\partial \xi} S_i(\xi) + \mathbf{e}_N \right\},$$

where

$$\mathbf{e}_N = \frac{1}{2} N^{-1} \sum_{i=1}^{N} \frac{\partial^2}{\partial \xi \partial \xi^T} S_i(\xi^\dagger) \cdot (\hat{\xi} - \xi),$$

and $|\xi^\dagger - \xi| < |\hat{\xi} - \xi|$.

Using (6.7), (6.9), and (**L1**), we have that

$$\mathbf{e}_N = \frac{1}{2} \left\{ N^{-1} \sum_{i=1}^{N} \mathrm{E}\left[\frac{\partial^2}{\partial \xi \partial \xi^T} S_i(\xi^\dagger) \right] + o_p(1) \right\} (\hat{\xi} - \xi).$$

By Lemma 6.2, $\hat{\xi}$, and hence $\xi^\dagger$, is consistent. In addition, by (6.7) and the arguments in the proof of Lemma 6.1, $N^{-1} \sum_{i=1}^{N} \mathrm{E}\left[\frac{\partial^2}{\partial \xi \partial \xi^T} S_i(\xi^\dagger) \right]$ is bounded away from infinity. So, $\mathbf{e}_N = o_p(1)$.

Likewise, we can use (6.6), (6.8), and (**L1**) to show that

$$N^{-1} \sum_{i=1}^{N} \frac{\partial}{\partial \xi} S_i(\xi) = N^{-1} \mathbf{J}_{(N)} + o_p(1).$$

Using Lemma 6.3, we now have that

$$\begin{aligned}
S_{(N)}(\hat{\xi}) &= \sum_{i=1}^{N} S_i(\xi) + \sqrt{N} \left[\sqrt{N} \sum_{i=1}^{N} \mathbf{U}_i^T \mathbf{I}_{\xi\xi(N)}^{-1} + o_p(1) \right] \left(N^{-1} \mathbf{J}_{(N)} + o_p(1) \right) \\
&= \sum_{i=1}^{N} S_i(\xi) + \sum_{i=1}^{N} \mathbf{U}_i^T \mathbf{I}_{\xi\xi(N)}^{-1} \mathbf{J}_{(N)} + o_p(1) N \sum_{i=1}^{N} \mathbf{U}_i^T \mathbf{I}_{\xi\xi(N)}^{-1} + o_p(1) N^{-\frac{1}{2}} \mathbf{J}_{(N)} + o_p(1) \sqrt{N} \\
&= \sum_{i=1}^{N} S_i(\xi) + \sum_{i=1}^{N} \mathbf{U}_i^T \mathbf{I}_{\xi\xi(N)}^{-1} \mathbf{J}_{(N)} + o_p(1) \sqrt{N} \left(\frac{\sum_{i=1}^{N} \mathbf{U}_i^T}{\sqrt{N}} \right) \left(N \mathbf{I}_{\xi\xi(N)}^{-1} \right) \\
&\quad + o_p(1) \sqrt{N} \left(N^{-1} \mathbf{J}_{(N)} \right) + o_p(1) \sqrt{N}.
\end{aligned}$$

By (6.6), (6.7), and Condition 4, we have that $N^{-1} \mathbf{J}_{(N)}$ is bounded away from infinity. As argued in the proof of Lemma 6.3, $\sum_{i=1}^{N} \mathbf{U}_i^T / \sqrt{N} = O_p(1)$. Thus, to show that

$$S_{(N)}(\hat{\xi}) = \sum_{i=1}^{N} \left\{ S_i(\xi) + \mathbf{U}_i^T \mathbf{I}_{\xi\xi(N)}^{-1} \mathbf{J}_{(N)} + o_p(1) \sqrt{N} \right\}, \tag{6.19}$$

we need only prove that $N \mathbf{I}_{\xi\xi(N)}^{-1} = O(1)$.

Using the idea discussed in the proof of Lemma 6.1, the matrices $\{\mathbf{I}_{\xi\xi i}\}_{i=1}^{N}$ belong to a set of at most n matrices. Denote this set by $\{\mathcal{I}_j\}_{j=1}^{n}$, where $\mathcal{I}_j$ is the matrix $\mathbf{I}_{\xi\xi i}$ for $n_i = j$.

Let $\{d_j\}$ be the set of minimum eigenvalues of $\{\mathcal{I}_j\}$, and let $\{d'_j\}$ be the set of maximum eigenvalues of $\{\mathcal{I}_j\}$. Define $d = \min\{d_j\}$ and define $d' = \max\{d'_j\}$. Since the matrices $\{\mathcal{I}_j\}$ are positive definite, d and d' are strictly positive. Let $c_j^{(N)}$ be the number of times that $n_i = j$, $i = 1, \ldots, N$. Then

$$\mathbf{I}_{\xi\xi(N)} = \sum_{j=1}^{n} c_j^{(N)} \mathcal{I}_j.$$

The minimum eigenvalue of $\mathbf{I}_{\xi\xi(N)}$ is given by

$$\min \frac{\mathbf{v}'\mathbf{I}_{\xi\xi(N)}\mathbf{v}}{\mathbf{v}'\mathbf{v}},$$

where this minimum is computed over all p-dimensional vectors $\mathbf{v}$. Thus

$$\begin{aligned}
\min \frac{\mathbf{v}'\mathbf{I}_{\xi\xi(N)}\mathbf{v}}{\mathbf{v}'\mathbf{v}} &= \min \frac{\sum_{j=1}^{n} c_j^{(N)} \mathbf{v}'\mathcal{I}_j\mathbf{v}}{\mathbf{v}'\mathbf{v}} \\
&\geq \sum_{j=1}^{n} c_j^{(N)} \min \frac{\mathbf{v}'\mathcal{I}_j\mathbf{v}}{\mathbf{v}'\mathbf{v}} \\
&\geq \sum_{j=1}^{n} c_j^{(N)} d \\
&= Nd
\end{aligned}$$

By a similar argument, the maximum eigenvalue of $\mathbf{I}_{\xi\xi(N)}$ is at most Nd'. Thus, the eigenvalues of $N\mathbf{I}_{\xi\xi(N)}^{-1}$ must lie in the interval $[1/d', 1/d]$. This is sufficient to guarantee that $N\mathbf{I}_{\xi\xi(N)}^{-1} = O(1)$, and hence that (6.19) holds.

Now, let $R_{i(N)} = S_i(\xi) + \mathbf{U}_i^T \mathbf{I}_{\xi\xi(N)}^{-1} \mathbf{J}_{(N)}$. As demonstrated in the proof of (6.10), $\mathrm{E}[\mathbf{U}_i] = 0$. Similarly, by Condition 3 and Theorem 6.1, we can show that $\mathrm{E}[S_i(\xi)] = 0$. So, $\mathrm{E}[R_{i(N)}] = 0$. Defining $\sigma_i^2 \equiv \mathrm{Var}\left[R_{i(N)}\right]$, Jacqmin-Gadda & Commenges (1995) prove that $s_N^2 \equiv \sum_{i=1}^{N} \sigma_i^2 = I_{c(N)}$. We now show that $\left(\sum_{i=1}^{N} R_{i(N)}\right)/s_N$ is asympotically distributed as $N(0,1)$, using Lyapounov's condition. In particular, we show that $\sigma_i^2 < \infty$ for all i, and that

$$\lim_{N\to\infty} \sum_{i=1}^{N} \frac{1}{s_N^{2+\delta}} \mathrm{E}[|R_{i(N)}|^{2+\delta}] = 0 \tag{6.20}$$

for some $\delta > 0$. Now

$$\begin{aligned}
\sigma_i^2 &= \mathrm{E}[S_i(\xi) + \mathbf{U}_i^T \mathbf{I}_{\xi\xi(N)}^{-1} \mathbf{J}_{(N)}]^2 \\
&= \mathrm{E}[S_i(\xi)]^2 + \mathbf{J}_{(N)}^T \mathbf{I}_{\xi\xi(N)}^{-1} \mathrm{E}[\mathbf{U}_i\mathbf{U}_i^T] \mathbf{I}_{\xi\xi(N)}^{-1} \mathbf{J}_{(N)} + 2\mathrm{E}[S_i(\xi)\mathbf{U}_i^T] \mathbf{I}_{\xi\xi(N)}^{-1} \mathbf{J}_{(N)}.
\end{aligned}$$

Using the same argument as in the proof of Lemma 6.1, σ_i^2 can take on at most n possible values. So, $\sigma^2 \equiv \min \sigma_i^2 > 0$. Similarly, using Conditions 3 and 4 and Theorem 6.1, $\mathrm{E}[|R_{i(N)}|^3] \leq \gamma$, where γ is a finite constant. So, letting $\delta = 1$ in (6.20),

$$\lim_{N\to\infty} \sum_{i=1}^{N} \frac{1}{s_N^{2+\delta}} \mathrm{E}[|R_{i(N)}|^{2+\delta}] \leq \lim_{N\to\infty} \frac{\sum_{i=1}^{N} \gamma}{(N\sigma^2)^{\frac{3}{2}}}$$

$$= \lim_{N\to\infty} \frac{\gamma}{\sqrt{N}\sigma^3}$$
$$= 0,$$

verifying that $\sum_{i=1}^{N} R_{i(N)}/s_N$ is asymptotically distributed as $N(0,1)$. Since $s_N^2 = I_{c(N)}$, we see that

$$\frac{S_{(N)}(\hat{\xi})}{I_{c(N)}^{\frac{1}{2}}} = \frac{\sum_{i=1}^{N}\left\{R_{i(N)} + o_p(1)\sqrt{N}\right\}}{s_N} = \frac{\sum_{i=1}^{N} R_{i(N)}}{s_N} + o_p(1),$$

where the last equality follows from the fact that $\sqrt{N}/s_N \leq 1/\sigma < \infty$. Then, by Slutsky's theorem, $S_{(N)}(\hat{\xi})/I_{c(N)}^{\frac{1}{2}}$ is also asymptotically distributed as $N(0,1)$. □

We conclude this section with a discussion of other models to which this type of test may apply. Adding time-independent covariates to the conditional model for the observed data is the easiest of these; by assuming that the coefficients are bounded, all of the results we have presented are valid in this case as well. Likewise, incorporating a patient-specific random effect in the model for the hidden processes should be fairly simple, as long the associated parameters are bounded, and the processes $\{Z_{it} \mid u_i\}_{t=1}^{n_i}$, $i = 1, \ldots, N$ are homogeneous and stationary.

It should be reasonably straightforward to extend our theory to the case where there are multiple, patient-specific random effects, $\mathbf{u}_i$, in either the conditional model for the observed data or the model for the hidden process or both, and we wish to simultaneously test the significance of all of the variance components. In this case, $S_i(\xi)$ would be a vector obtained by taking derivatives of $\log f(\mathbf{y}_i; \mathbf{u}_i, \psi)$ with respect to each variance component, and then evaluating each variance component at 0. Consequently, we would require a multivariate central limit theorem in the derivation of the asymptotic distribution of $\sum_{i=1}^{N} S_i(\xi)$. Similar theory should apply to the case where we have a large number of independent clusters of processes, where the number of processes in each cluster is bounded. For instance, in Example 2 of Section 5.2, we assumed both patient- and centre-specific random effects. Presumably, a multivariate version of our test could be used to assess the significance of both variance components, assuming that the number of patients in each centre is bounded, and that the number of centres is large.

Some models, however, would present more substantial challenges. For instance, BRR's theory applies only to homogeneous, stationary HMMs. It is not immediately apparent how to extend this theory to a more general setting. Even more difficult is the case where the random effects are crossed, so that observations on different patients (or clusters) are not necessarily independent. Our approach relies heavily on the assumption of independence among patients; for example the law of large numbers and central limit theorem that we use apply only to independent observations. Finally, in the case where there are multiple

random effects, it is not clear how to use our procedure to test the significance of a subset of the variance components. In this case, the test statistic (6.5) would involve a (possibly multidimensional) integral, and the results of BRR would not apply. Further research is required to address the issue of hypothesis testing in these contexts.

6.4 Applications

In this section we illustrate the use of HMMs with random effects in applied settings, both in terms of fitting the model and in conducting the test of the significance of the variance components. First, we address the issue of estimating the quantity $I_{c(N)}$ in (6.18). We then analyze two different data sets: MRI data from a group of MS patients, and faecal coliform count data originally presented by Turner *et al.* (1998).

6.4.1 Computing the Test Statistic

The quantity (6.18) contains unknown parameters, and hence is not a test statistic. In particular, $I_{c(N)}$ (defined by 6.17) is a function of parameter values which are not known *a priori*. However, as long as we have a consistent estimate of $I_{c(N)}$, $\hat{I}_{c(N)}$, Conditions 1–4 and Slutsky's theorem guarantee that $S_{(N)}(\hat{\xi})/\hat{I}_{c(N)}^{\frac{1}{2}}$ has an asymptotically standard normal distribution as $N \to \infty$.

In the GLMM setting, $I_{c(N)}$ can be expressed analytically in terms of ξ. In this case, the usual practice is to estimate $I_{c(N)}$ by replacing ξ with $\hat{\xi}$ (e.g. Dean 1992; Jacqmin-Gadda & Commenges 1995; Lin 1997).

In contrast, in the HMM context, $I_{c(N)}$ does not have a closed form, and it is unclear how to evaluate this function at $\hat{\xi}$. For this reason, we use a different method to obtain $\hat{I}_{c(N)}$: the parametric bootstrap (e.g. Efron & Tibshirani 1993). Specifically, under the null hypothesis that the variance components are 0, by Lemma 6.2 and Conditions 1 and 2, $f_0(\mathbf{y}_i;\hat{\xi})$ is a consistent estimate of $f_0(\mathbf{y}_i;\xi)$. Using this fact, we estimate the unknown expected values $I_{(N)}$, $\mathbf{J}_{(N)}$, and $\mathbf{K}_{(N)}$ in $I_{c(N)}$ by generating V samples from $f_0(\mathbf{y}_i;\hat{\xi})$ and computing the mean of appropriate functions of these. For example, to estimate $\mathbf{J}_{(N)}$, for each $v = 1,\ldots,V$ and each $i = 1,\ldots,N$, we generate a sample of n_i observations from $f_0(\mathbf{y}_i;\hat{\xi})$. Denote the vth sample by $\mathbf{y}_i^{(v)}$, and the associated value of $S_i(\xi)$ by $S_i(\mathbf{y}_i^{(v)};\xi)$. We then approximate $\frac{\partial}{\partial\xi}S_i(\mathbf{y}_i^{(v)};\hat{\xi})$ by taking differences of $S_i(\mathbf{y}_i^{(v)};\hat{\xi})$ in the neighbourhood of $\hat{\xi}$. Repeating this

procedure V times, we compute our estimate as

$$\hat{\mathbf{J}}_{(N)} = \frac{1}{V}\sum_{v=1}^{V}\frac{\partial}{\partial\xi}S_{(N)}(\mathbf{y}_i^{(v)};\hat{\xi}).$$

Rather than using the parametric bootstrap to estimate the quantity $\mathbf{I}_{\xi\xi(N)}$ in $I_{c(N)}$, we instead use the matrix of second derivatives of the log-likelihood evaluated at $\hat{\xi}$. This matrix is a by-product of the quasi-Newton minimization routine, and hence is readily available.

Sometimes this method leads to a negative value for $\hat{I}_{c(N)}$. We discard these cases, which presumably results in an estimate of $I_{c(N)}$ that is positively biased. However, the fact that this bias is positive makes the test more conservative, and hence is not a concern.

It is difficult to assess the accuracy of this estimation procedure in our setting. However, we have investigated the simpler case where there is only one observed lesion count on each patient and we wish to test for overdispersion relative to the Poisson distribution (e.g. Dean 1992). We are able to compare the estimates of $I_{c(N)}$ obtained using the parametric bootstrap with $V = 100$ with those obtained by replacing ξ with $\hat{\xi}$ in $I_{c(N)}$ (since, in this case, $I_{c(N)}$ can be computed analytically). The estimates were very similar for the data sets we considered. Thus, the parametric bootstrap seems to be well-behaved, at least in this simpler context.

If $n_i \equiv n$ for all i and there are no covariates, the vectors of patients' observations are iid. In this case, two other options are available for estimating $I_{c(N)}$: the nonparametric bootstrap and the jackknife (Efron & Tibshirani 1993). As with the parametric bootstrap, these methods require the generation of realizations of the random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$. The former involves resampling with replacement from the observed vectors of data; the latter focuses on the samples that omit the ith patient's observations, $i = 1, \ldots, N$, one at a time. In our simulation study in Section 6.5 we compare the parametric and nonparametric bootstrap estimates of $I_{c(N)}$ for some of the cases that we consider.

6.4.2 MS/MRI Data

In Section 2.4, we discussed the model for individual MS patients developed by Albert *et al.* (1994). These authors comment that MS is a very heterogeneous disease, in the sense that the degree and behaviour of the disease is expected to vary considerably among patients. Although modelling each patient's data separately certainly allows for between-patient differences, the cost of adding so many parameters to the model is increased uncertainty about all parameter estimates.

In this section, we propose using random effects as a means of capturing between-patient

differences while maintaining a parsimonious model. We fit our model to the Vancouver PRISMS data on 13 placebo patients described in Section 4.3.2.

Let Y_{it} be the lesion count for patient i at time t, and let Z_{it} be the associated hidden state. We fit Model I of Chapter 5 with one, patient-specific random effect, u_i in the conditional model for the observed data, and no covariates. In particular, we assume that given $Z_{it} = k$ and u_i, Y_{it} is distributed as Poisson(μ_{itk}) with

$$\log \mu_{itk} = \tau_k + u_i,$$

where τ_k is the fixed effect of being in state k, and $\{u_i\}$ are iid, each with a $N(0, \sigma^2)$ distribution. Furthermore, we assume that the transition probabilities are homogeneous, non-zero, and common to all patients, and that $\{Z_{it}\}_{t=1}^{n_i}$ is stationary for all i. Our results in Section 3.4 suggested that two hidden states are appropriate for this type of data, and thus we will use $K = 2$ here.

For the purpose of comparison, we also fit the same model with no random effect. In other words, we consider the case where $Y_{it} \mid Z_{it} = k$ is distributed as Poisson(μ_{itk}) with

$$\log \mu_{itk} = \tau_k.$$

We refer to this model as the fixed model, and to the model with a random effect as the mixed model.

The advantage of the mixed model is that we allow different mean lesion counts for different patients with the addition of only one extra unknown parameter (σ). Although we have assumed a restricted form for the distribution of these means without any prior information about its appropriateness, we could test this assumption by fitting the fixed model to each patient individually, and then estimating the distribution of the values of τ_1 and τ_2 that we obtain.

To avoid the need to use constrained optimization methods, we reparameterize the model so that each parameter has as its range the whole real line. These transformations are provided in Table 6.1. We then fit both models by directly maximizing the likelihood. In the case of the mixed model, we use the Gauss-Hermite quadrature formula to approximate the integral that appears in the likelihood. We would expect this formula to provide a high level of accuracy in this case, where we are integrating with respect to the normal distribution. Table 6.1 gives the parameter estimates and standard errors that result from fitting both models.

By including the random effect, we observe what appears to be quite a large increase in the log-likelihood. In addition, we see substantial changes in the estimates of some of the fixed effects, τ_2 and P_{11}^*, in particular.

Table 6.1: Parameter estimates and standard errors for Vancouver PRISMS data

		Mixed Model		Fixed Model	
Parameter	Transformation	Estimate	S.E.	Estimate	S.E.
τ_1	NA	−0.807	0.289	−0.907	0.115
τ_2	NA	1.344	0.173	1.657	0.107
P_{11}^*	$\log\left(\frac{P_{11}}{1-P_{11}}\right)$	2.876	0.594	3.700	0.456
P_{21}^*	$\log\left(\frac{P_{21}}{1-P_{21}}\right)$	−1.097	0.517	−1.488	0.578
σ^*	$\log\sigma$	0.752	0.188	NA	NA
	$\log\mathcal{L}$	−249.57		−268.41	

More formally, the mixed model satisfies the conditions of Theorem 6.3, so we can use our variance component test to test the hypothesis that $\sigma = 0$. This test is equivalent to a comparison of the fixed and mixed models. For these data, we compute $\sum_{i=1}^{13} S_i(\hat{\xi}) = 215.1$, and use the parametric bootstrap to obtain $\hat{I}_{c(13)} = 3396.0$. The value of our test statistic is then $215.1/\sqrt{3396.0} = 3.69$, which results in a p-value of < 0.001 when compared to the standard normal distribution. Thus, we have strong evidence in favour of the mixed model.

6.4.3 Faecal Coliform Data

Another application of HMMs with random effects is to the faecal coliform data first analyzed by Turner *et al.* (1998). In this study, sea water samples were collected at seven sites near Sydney, Australia, over a four year period. At each site, samples were taken from four depths (0, 20, 40, and 60 m) and the associated coliform counts recorded.

Turner *et al.* (1998) analyze these data using a Poisson HMM with depth and site as covariates in the conditional model for the observed data. Observations from each depth and site combination are treated as independent time series. The hidden states (0 or 1) are presumed to correspond to whether the sample was taken from above or below the thermocline. These authors treat site as a fixed effect, but certainly in some contexts this effect might be more appropriately treated as random. For example, repeated measurements at each site (i.e. at the four depths) may be correlated. Moreover, it may be desirable to think of the sites as a random sample, in which case we would want to generalize our results to all "sites" in the area.

We thus re-analyze the data by incorporating a random site effect, u_i, using the framework of Model I. Let Y_{ijt} be the coliform count at site i, depth j, and time t. Given the hidden state $Z_{ijt} = k$ and the site effect, u_i, we model the distribution of Y_{ijt} as Poisson(μ_{ijtk}),

Table 6.2: Parameter estimates and standard errors for faecal coliform data

		Mixed Model		Fixed Model	
Parameter	Transformation	Estimate	S.E.	Estimate	S.E.
τ_0	NA	-2.182	0.135	-2.209	0.129
τ_1	NA	1.515	0.055	1.537	0.034
b_1	NA	0.027	0.065	0.012	0.061
b_2	NA	-0.062	0.063	-0.054	0.060
b_3	NA	0.018	0.056	0.016	0.055
P_{00}^*	$\log\left(\frac{P_{00}}{1-P_{00}}\right)$	1.976	0.149	1.993	0.148
P_{10}^*	$\log\left(\frac{P_{10}}{1-P_{10}}\right)$	-0.194	0.227	-0.218	0.223
σ^*	$\log\sigma$	-2.246	0.561	NA	NA
	$\log\mathcal{L}$	-1533.3		-1534.4	

where

$$\log\mu_{ijtk} = \tau_k + b_j + u_i.$$

Following Turner *et al.* (1998), we assume that $k = 0$ or 1, i.e. that $K = 2$. To prevent over-parameterization of the model, we impose the constraint $\sum_{j=1}^{4} b_j = 0$. We model the site effects, $\{u_i\}$, as independent, each with a $N(0, \sigma^2)$ distribution. Finally, we assume that the transition probabilities are homogeneous, non-zero, and common to all patients, and that $\{Z_{ijt}\}_{t=1}^{n_i}$ is stationary for all i, j.

We also fit a model with no site effect, i.e. we propose that $Y_{ijt} \mid Z_{ijt} = k$ is distributed as Poisson(μ_{ijtk}), where

$$\log\mu_{ijtk} = \tau_k + b_j.$$

We again refer to these models as the fixed and mixed models, respectively.

The results of these analyses are presented in Table 6.2. In this case, it appears that the random site effect is not necessary. The log-likelihoods and the parameter estimates for the fixed and mixed models are both very similar, indicating that there is likely very little site-to-site variation.

This example differs somewhat from the paradigm presented in Section 6.3 in that there are four sequences (one for each depth) associated with each site. In particular, letting $\mathbf{Y}_{ij}$ represent the observations from site i, depth j, (6.5) becomes

$$S_i(\xi) = \frac{1}{2}\left\{\sum_{j=1}^{4}\frac{\partial^2}{\partial u_i^2}\log f(\mathbf{y}_{ij} \mid u_i, \psi)\bigg|_{u_i=0} + \left[\sum_{j=1}^{4}\frac{\partial}{\partial u_i}\log f(\mathbf{y}_{ij} \mid u_i, \psi)\bigg|_{u_i=0}\right]^2\right\}.$$

However, assuming that the depth effects are bounded, it is easy to show that the mixed

model satisfies the conditions of Theorem 6.3. Our variance component test is thus statistically valid for assessing the difference in fit between the mixed and fixed models.

We compute $\sum_{i=1}^{7} S_i(\hat{\xi}) = 390.3$, and estimate $\hat{I}_{c(7)} = 79240.5$ using the parametric bootstrap. These estimates give a value of $390.3/\sqrt{79240.5} = 1.39$ for our test statistic, and a corresponding p-value of 0.166. Therefore, as expected, we do not reject the null hypothesis that the fixed model adequately represents these data. Interestingly, in an analysis of a subset of these data using fixed site effects, Turner *et al.* (1998) find evidence of site-to-site variation. This difference in conclusions leads naturally to the question of the sensitivity of our test, an issue which we address in Section 6.5.

As an aside, this example is also a nice illustration of the ease with which HMMs can accommodate missing values. In order to satisfy the condition that the observations be equally spaced in time, Turner *et al.* (1998) round the observation times to the nearest week, and use missing values for weeks with no associated observation. For instance, say there are s missing values between times t_0 and t at site i. Then, in the piece of the likelihood corresponding to the observation at t, $A^{it}_{k\ell} = P_{k\ell}\, f(y_{it} \mid Z_{it} = \ell, u_i)$ (see (5.4)), we simply replace $P_{k\ell}$ with the $(s+1)$-step transition probability, $P_{k\ell}(s+1)$.

6.5 Performance of the Variance Component Test

In Section 6.4, we applied our variance component test to two data sets with unknown generating mechanisms. It is also of interest to apply the test to data generated from known models so that we can assess its performance. In particular, we wish to make statements about the probability of Type I and Type II error for selected cases.

As in Section 3.5, we use the ideas of experimental design to choose the parameter values for the cases we consider. For simplicity, we assume that each process has the same number of observations (n). There are many factors that could potentially impact the performance of the test, including n, the number of processes (N), the number of hidden states (K), the conditional distribution for the observed data, the size of the variance component (σ^2), and the spacing of the components and the proportion of time spent in each hidden state (see Section 3.5). It is expected that of these factors, N, K, and σ^2 are the most influential. Thus, in our study, we examine only this subset. We look at the levels $N = 10, 30, 50$, $K = 1, 2, 3$, and $\sigma = 0, 1.11, 1.22$, which reflect "small", "medium", and "large" values of these factors. Since the applications in this thesis have focused on lesion counts observed during MS/MRI clinical trials, we consider only the model (5.5) with $f(y_{it} \mid Z_{it} = k, u_i, \psi)$ corresponding to the Poisson(μ_{itk}) distribution, $\log \mu_{itk} = \tau_k + u_i$, and fix $n = 30$ (a typical

Table 6.3: Parameter values used in the simulation study

K^0	State Effects	Transition Probabilities
1	$\tau = [0]$	$P = [1]$
2	$\tau = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$	$P = \begin{bmatrix} 0.69 & 0.31 \\ 0.31 & 0.69 \end{bmatrix}$
3	$\tau = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$	$P = \begin{bmatrix} 0.71 & 0.19 & 0.10 \\ 0.19 & 0.10 & 0.71 \\ 0.10 & 0.71 & 0.19 \end{bmatrix}$

value in this setting). Furthermore, we investigate only the case of well-spaced components with an equal proportion of observations from each. The values of $\{\tau_k\}$ and of the transition probabilities that we use are provided in Table 6.3.

For each model, we generate 400 data sets. Then, for each data set, we compute the test statistic (using the parametric bootstrap with $V = 100$ to estimate $I_{c(N)}$) and record whether the null hypothesis is rejected at the 5% level. For some models, we also estimate $I_{c(N)}$ using the nonparametric bootstrap and 100 of the data sets. In these cases, we resample with replacement 100 times from the simulated data set and compute $\hat{I}_{c(N)}$ using these samples. We then recompute the test statistic using this value of $\hat{I}_{c(N)}$, and again record whether the null hypothesis is rejected at the 5% level. The results of our simulations are given in Table 6.4.

In the case of the parametric bootstrap, Table 6.4 shows that the observed rejection rate when $\sigma = 0$ is less than 5% for all values of N and for $K = 1, 2$. When $K = 3$, this rate is a little higher, ranging between 6 and 12%. We conclude that, overall, the test seems to control for Type I error reasonably well, even in finite samples. In terms of the power of the test, it appears that σ and N are quite influential. For the "large" values of these factors, the test may have at least 88% power. The power seems to be quite low for small sample sizes, though may be as much as 49% when $\sigma = 1.22$ and $N = 10$. Similarly, the power appears to be relatively low when $\sigma = 1.11$, but may be up to 29% when $N = 50$. With respect to the value of K, the power seems to decrease somewhat with increasing K, but the test does not seem as sensitive to K as to σ and N.

As expected, since the parametric and nonparametric bootstrap methods both give consistent estimates of $I_{c(N)}$ under the null hypothesis, both methods lead to good control of

Table 6.4: Results of the simulation study

			Percentage Rejected	
K^0	N	σ	Parametric Bootstrap	Nonparametric Bootstrap
1	10	0	4.25	18
1	30	0	3.75	NA
1	50	0	4.25	6
1	10	1.11	10.25	6
1	30	1.11	19.00	NA
1	50	1.11	29.0	4
1	10	1.22	49.25	5
1	30	1.22	87.50	NA
1	50	1.22	98.25	81
2	10	0	3.25	19
2	30	0	3.50	NA
2	50	0	2.75	7
2	10	1.11	10.50	5
2	30	1.11	13.50	NA
2	50	1.11	24.50	3
2	10	1.22	30.50	1
2	30	1.22	75.25	NA
2	50	1.22	90.25	32
3	10	0	6.00	NA
3	30	0	12.25	NA
3	50	0	9.00	NA
3	10	1.11	9.50	NA
3	30	1.11	22.50	NA
3	50	1.11	23.75	NA
3	10	1.22	23.75	NA
3	30	1.22	69.50	NA
3	50	1.22	87.75	NA

the Type I error when N is large. However, it is not clear that the use of the nonparametric bootstrap results in a valid test for smaller values of N. In particular, the rejection rates when $\sigma = 0$ and $N = 10$ are up to 19% in this case. The power of the test when we use the nonparametric bootstrap appears to be very low in general (except perhaps when $K = 1$, $\sigma = 1.22$, and $N = 50$), especially relative to the power when we use the parametric bootstrap. Thus, based on these preliminary results, we recommend using the parametric rather than the nonparametric bootstrap to estimate $I_{c(N)}$. Further investigation is, nonetheless, warranted, especially in the case where the model has been specified incorrectly.

Chapter 7

Future Work

In our final chapter, we review the contents of this thesis, and discuss possible extensions to our work. We present these ideas in the context of the design and analysis of MS/MRI clinical trials, one of our main areas of interest for future research.

In Chapter 2, through our exploration of some extensions to the model of Albert *et al.* (1994), we illustrated some issues surrounding the selection of a HMM. Although we presented our discussion in the context of a single time series, these issues are relevant to multiple time series as well. In particular, the application of a HMM requires decisions concerning the type of model (homogeneous or non-homogeneous, stationary or non-stationary), the conditional distribution of the observed data, the number of lags in the hidden Markov chain, and the number of hidden states. These features of the model must be carefully considered when developing a HMM for a given data set. Chapter 2 also reveals that, if the model structure proposed by Albert *et al.* (1994) is indeed appropriate for relapsing-remitting MS/MRI data, further research regarding the asymptotic properties of non-homogeneous HMMs will be necessary.

In Chapter 3, we presented a method for consistently estimating the number of hidden states in a single, stationary HMM. As outlined in Section 3.6, there are a number of issues related to this method that are worthy of further study, such as the proposed two-stage estimation procedure, the choice of distance function and c_n, and the form of the penalty term. In addition, estimating the number of hidden states in a HMM for multiple processes is still an open question.

In Chapter 4, we developed a graphical method for investigating the GOF of a single, stationary HMM. We demonstrated that, in the limit, this method will detect deviations from the proposed model with probability 1. Our GOF plots will provide a simple, useful tool for assessing the quality of the fit of models for the relapsing-remitting MS/MRI data.

Outstanding topics for research include the formal assessment of the variability in the plots, and GOF techniques for HMMs for multiple processes.

In Chapter 5, we proposed a general class of HMMs for multiple processes. We showed how both covariates and random effects can be used to account for inter-patient variability while maintaining a parsimonious model. Several extensions to these models are possible. First, we could allow ϕ to vary among patients, or among treatment arms, in (5.2). In most mixture models and HMMs, the mixing occurs over the means of the components. Nonetheless, other possibilities exist, including mixing over the variances of the components. The cost of this generalization would presumably be loss of efficiency in estimating the other model parameters. Second, methods of estimation other than maximum likelihood may be available. For example, in the GLMM setting, penalized and marginal quasi-likelihood (Breslow & Clayton 1993), maximum hierarchical likelihood (Lee & Nelder 1996), and Monte Carlo Newton-Raphson (McCulloch 1997) have been used. In the HMM context, these procedures may also give good asymptotic properties and be easier to implement than maximum likelihood estimation.

Finally, in Chapter 6, we discussed the asymptotic properties of the maximum likelihood estimates of the models presented in Chapter 5. Formal proof of the consistency and asymptotic normality of these estimates (e.g. by verification of the regularity conditions of Bradley & Gart 1962), and hence justification for using standard hypothesis tests, is still pending. These results will be required for power calculations in the context of the design of experiments. Our primary contribution in this chapter, however, was the development of a test for the variance component in our models. We studied the performance of our test for finite samples in Section 6.5, and showed that the test seems to control for Type I error quite well, even in small samples. The power of the test seems reasonable for moderate-sized samples and variance components, and does not vary substantially with the number of hidden states. Lin (1997) and Hall & Præstgaard (2001) proved optimality properties of this test in the GLMM setting, and these properties may apply in our context as well. This test will be important in the design of clinical trials, since if the model without random effects is adequate, then the computation of the required sample sizes will be substantially simplified.

Our preliminary results suggest that HMMs may be useful models for capturing the behaviour of relapsing-remitting MS/MRI lesion count data. However, other models may provide an equally good – or better – representation of these data. For example, GLMMs are frequently employed in the analysis of longitudinal data. Some features that are common to HMMs and GLMMs are immediately apparent: both postulate the existence of a latent (unobserved) process, both assume the conditional independence of the observed data given the values of this latent process, and both control the correlation structure of the observed

data through specification of the latent process. Despite the similarities between the two models, little is known about the connections between their theoretical properties, or about the issues involved in choosing between models in applications. Thus, one area of interest for future research is the establishment of a formal link between HMMs and GLMMs.

Assuming that we have identified a reasonable class of models for the MS/MRI data, two natural questions arise: how do we incorporate a treatment effect in a realistic, but interpretable, way, and which design is best for clinical trials in this setting? With respect to the first question, a treatment could be beneficial in two ways: either by reducing the mean lesion count, or by extending the time between relapses. Choosing between these perspectives, and deciding on the particular form of the treatment effect, are issues of critical importance. The second question relates to the selection of a sample size (both number of patients and number of scans per patient) and frequency of sampling that would yield a desired level of power. Ideally, we would be able to plan for the following types of studies:

1. Studies of untreated patients where lesion counts can be considered stationary (e.g. in Phase II trials, which are relatively short-term, and where MRI responses tend to be used as primary outcomes)

2. Studies of untreated patients where lesion counts may not be stationary (e.g. when patients are selected for their high level of disease activity at the commencement of the study, or in longer-term studies)

3. Studies of treated patients, whose lesion counts, we would hope, would not be stationary.

The work in this thesis provides a starting point for addressing these questions.

Appendix A

The EM Algorithm

This appendix details the implementation of the EM algorithm for three different models. This algorithm consists of an Expectation (E-) step followed by a Maximization (M-) step. In the E-step, we form the "complete" log-likelihood, which is the log-likelihood we would have if we were able to observe the hidden process and random effects as well as realizations of the process $\{Y_t\}$. We then take the expectation of the complete log-likelihood conditional on $\{Y_t\}$, and evaluate the resulting expression at the initial parameter estimates. In the M-step, we maximize this expectation over the space of the unknown parameters to get updated parameter estimates. We then use these new estimates in place of the initial parameter estimates in the next iteration. This procedure is repeated until convergence of the parameter estimates is achieved.

A.1 EM Algorithm for HMMs for Single Processes

In this section, we outline the steps of the EM algorithm for a single, homogeneous HMM with unknown initial probabilities, and with observations taking on only a finite number of values. The parameter estimates given at each iteration of the EM algorithm take on a special form in this context, and, for this reason, are usually referred to as the Forward-Backward Algorithm. These estimates will converge to the maximum likelihood estimates (Baum *et al.* 1970); in other words, the parameter estimates will, in the limit, maximize (2.2).

We use the additional notation $\gamma_{y,k} = \mathrm{P}(Y_t = y \mid Z_t = k)$ for $t = 1, \ldots, n$, $k = 1, \ldots, K$, and $y \in \Omega$, for some finite set $\Omega = \{y_1, \ldots, y_d\}$.

E-Step

Thinking of the hidden states as "missing", we can write the complete likelihood, $\mathcal{L}_c(\psi)$, as

$$\begin{aligned}\mathcal{L}_c(\psi) &= f(\mathbf{y} \mid \mathbf{z}, \psi) f(\mathbf{z}; \psi) \\ &= \prod_{t=1}^{n} f(y_t \mid z_t, \psi) \cdot \pi_{z_1} \prod_{t=2}^{n} P_{z_{t-1}, z_t} \\ &= \prod_{t=1}^{n} \gamma_{y_t, z_t} \cdot \pi_{z_1} \prod_{t=2}^{n} P_{z_{t-1}, z_t}.\end{aligned}$$

So

$$\log \mathcal{L}_c(\psi) = \sum_{t=1}^{n} \log \gamma_{y_t, z_t} + \log \pi_{z_1} + \sum_{t=2}^{n} \log P_{z_{t-1}, z_t}.$$

Let $\psi^p = (\gamma^p_{y_1,1}, \ldots, \gamma^p_{y_d,K}, P^p_{1,1}, \ldots, P^p_{K,K}, \pi_1, \ldots, \pi_K)$ be the estimates of all unknown parameters at iteration p. Then the E-step is

$$\begin{aligned}\mathrm{E}[\log \mathcal{L}_c(\psi) \mid \mathbf{Y}, \psi^p] &= \sum_{t=1}^{n} \mathrm{E}[\log \gamma_{y_t, Z_t} \mid \mathbf{Y}, \psi^p] + \mathrm{E}[\log \pi_{Z_1} \mid \mathbf{Y}, \psi^p] + \sum_{t=2}^{n} \mathrm{E}[\log P_{Z_{t-1}, Z_t} \mid \mathbf{Y}, \psi^p] \\ &= \sum_{t=1}^{n} \sum_{z_t=1}^{K} f(z_t | \mathbf{y}, \psi^p) \log \gamma_{y_t, z_t} \qquad \text{(A.1)} \\ &\quad + \sum_{z_1=1}^{K} f(z_1 | \mathbf{y}, \psi^p) \log \pi_{z_1} \qquad \text{(A.2)} \\ &\quad + \sum_{t=2}^{n} \sum_{z_{t-1}=1}^{K} \sum_{z_t=1}^{K} f(z_{t-1}, z_t | \mathbf{y}, \psi^p) \log P_{z_{t-1}, z_t}. \qquad \text{(A.3)}\end{aligned}$$

M-Step

We now need to maximize this expectation over all of the unknown parameters in the model. Since $\{\gamma_{y,k}\}$, $\{\pi_k\}$, and $\{P_{k\ell}\}$ appear in separate terms, we can maximize each term individually.

Let $\mathbf{Y}_s^t = (Y_s, \ldots, Y_t)$. Define the *forward probabilities* at the pth iteration as

$$W_t^p(k) = \mathrm{P}(\mathbf{Y}_1^t = \mathbf{y}_1^t, Z_t = k \mid \psi^p)$$

and the *backward probabilities* as

$$X_t^p(k) = \mathrm{P}(\mathbf{Y}_{t+1}^n = \mathbf{y}_{t+1}^n \mid Z_t = k, \psi^p).$$

These probabilities can be computed recursively as

$$W_{t+1}^p(k) = \left(\sum_{\ell=1}^{K} W_t^p(\ell) P_{\ell k}^p \right) \gamma_{y_{t+1}, k}^p \qquad \text{(A.4)}$$

and

$$X_t^p(k) = \sum_{\ell=1}^{K} \gamma_{y_{t+1}, \ell}^p X_{t+1}^p(\ell) P_{k\ell}^p \qquad \text{(A.5)}$$

with $W_1^p(k) = \pi_k^p \gamma_{y_1,k}^p$ and $X_n^p(k) = 1$ (by convention).

First we will consider the maximization of (A.2), and hence the derivation of π_k^{p+1}, $k = 1, \ldots, K$. Since we have the constraint $\sum_{k=1}^K \pi_k = 1$, we will use the method of Lagrange multipliers and maximize the function

$$g_1 = \sum_{z_1=1}^{K} f(z_1 \mid \mathbf{y}, \psi^p) \log \pi_{z_1} - \lambda(\sum_{k=1}^{K} \pi_k - 1).$$

Then

$$\frac{\partial g_1}{\partial \pi_k} = \frac{1}{\pi_k} \mathrm{P}(Z_1 = k \mid \mathbf{Y}, \psi^p) - \lambda$$

and

$$\frac{\partial g_1}{\partial \lambda} = -\sum_{k=1}^{K} \pi_k + 1.$$

Setting these derivatives equal to 0 then gives the equations

$$\begin{aligned} \pi_k^{p+1} &= \frac{\mathrm{P}(Z_1 = k \mid \mathbf{Y}, \psi^p)}{\lambda^{p+1}} \\ \sum_{k=1}^{K} \pi_k^{p+1} &= 1, \end{aligned}$$

which we can solve to get

$$\pi_k^{p+1} = \frac{W_1^p(k) X_1^p(k)}{\sum_{\ell=1}^{K} W_1^p(\ell) X_1^p(\ell)}. \tag{A.6}$$

Next consider the maximization of (A.3) and the derivation of $P_{k\ell}^{p+1}$. We have the constraint that $\sum_{\ell=1}^K P_{k\ell} = 1$ for each k. Thus, we see that we need to maximize the function

$$g_2 = \sum_{t=2}^{n} \sum_{z_{t-1}=1}^{K} \sum_{z_t=1}^{K} f(z_{t-1}, z_t \mid \mathbf{y}, \psi^p) \log P_{z_{t-1}, z_t} - \sum_{k=1}^{K} \lambda_k \left(\sum_{\ell=1}^{K} P_{k\ell} - 1 \right).$$

Then

$$\frac{\partial g_2}{\partial P_{k\ell}} = \frac{1}{P_{k\ell}} \sum_{t=2}^{n} \mathrm{P}(Z_{t-1} = k, Z_t = \ell \mid \mathbf{Y}, \psi^p) - \lambda_k$$

and

$$\frac{\partial g_2}{\partial \lambda_k} = -\sum_{\ell=1}^{K} P_{k\ell} + 1.$$

Setting these derivatives to equal to 0 then gives the equations

$$\begin{aligned} P_{k\ell}^{p+1} &= \frac{\sum_{t=2}^{n} \mathrm{P}(Z_{t-1} = k, Z_t = \ell \mid \mathbf{Y}, \psi^p)}{\lambda_k^{p+1}} \\ \sum_{\ell=1}^{K} P_{k\ell}^{p+1} &= 1, \end{aligned}$$

which we can solve to get

$$P_{k\ell}^{p+1} = \frac{P_{k\ell}^{p} \sum_{t=2}^{n} W_{t-1}^{p}(k) X_{t}^{p}(\ell) \gamma_{y_t,\ell}^{p}}{\sum_{t=2}^{n} W_{t-1}^{p}(k) X_{t-1}^{p}(k)}. \tag{A.7}$$

Finally we will maximize (A.1) and derive $\gamma_{y,k}^{p+1}$. We have the constraint that $\sum_y \gamma_{y,k} = 1$ for each k. Thus the function to be maximized is

$$g_3 = \sum_{t=1}^{n} \sum_{z_t=1}^{K} f(z_t \mid \mathbf{y}, \psi^p) \log \gamma_{y_t,z_t} - \sum_{k=1}^{K} \lambda_k \left(\sum_{y} \gamma_{y,k} - 1 \right).$$

Then

$$\frac{\partial g_3}{\partial \gamma_{y,k}} = \sum_{\substack{t=1 \\ s.t.\ y_t=y}}^{n} \frac{1}{\gamma_{y,k}} \mathrm{P}(Z_t = k \mid \mathbf{Y}, \psi^p) - \lambda_k$$

and

$$\frac{\partial g_3}{\partial \lambda_k} = -\sum_{y} \gamma_{y,k} + 1.$$

Setting these derivatives equal to 0 then gives the equations

$$\begin{aligned} \gamma_{y,k}^{p+1} &= \frac{\sum_{\substack{t=1 \\ s.t.\ y_t=y}}^{n} \mathrm{P}(Z_t = k \mid \mathbf{Y}, \psi^p)}{\lambda_k^{p+1}} \\ \sum_{y} \gamma_{y,k}^{p+1} &= 1, \end{aligned}$$

which we can solve to get

$$\gamma_{y,k}^{p+1} = \frac{\sum_{\substack{t=1 \\ s.t.\ y_t=y}}^{n} W_t^p(k) X_t^p(k)}{\sum_{t=1}^{n} W_t^p(k) X_t^p(k)}. \tag{A.8}$$

A.2 EM Algorithm for HMMs with Random Effects in the Observed Process

In this section, we give the steps of the EM algorithm required to estimate the parameters of the model (5.4) in the case where the random effects are patient-specific. We assume that $\{\pi_k\}$ are unknown parameters to be estimated.

The convergence properties of the sequence of estimators given by the EM algorithm have not been studied in the context of HMMs for multiple processes, but Wu (1983) provides

sufficient conditions in the general case. One of these conditions is that the true parameters be in the interior of the parameter space. In other words, we require that the variance components be non-zero, and that $0 < P_{k\ell} < 1$ for each k and l where $P_{k\ell}$ is unknown.

E-Step

Recall that, for the model (5.4), the hidden states are assumed to be independent of the random effects. Now, thinking of both the hidden states and the random effects as "missing", the complete likelihood for this model is

$$\begin{aligned}\mathcal{L}_c(\psi) &= f(\mathbf{y} \mid \mathbf{z}, \mathbf{u}, \psi) f(\mathbf{z}; \psi) f(\mathbf{u}; \psi) \\ &= \prod_{i=1}^{N} \left\{ \prod_{t=1}^{n_i} f(y_{it} \mid z_{it}, \mathbf{u}_i, \psi) \cdot \pi_{z_{i1}} \prod_{t=2}^{n_i} P_{z_{i,t-1}, z_{it}} \cdot f(\mathbf{u}_i; \psi) \right\}\end{aligned}$$

So the contribution to the complete log-likelihood from patient i, $\log \mathcal{L}_c^i(\psi)$, is

$$\log \mathcal{L}_c^i(\psi) = \sum_{t=1}^{n_i} \log f(y_{it} \mid z_{it}, \mathbf{u}_i, \psi) + \log \pi_{z_{i1}} + \sum_{t=2}^{n_i} \log P_{z_{i,t-1}, z_{it}} + \log f(\mathbf{u}_i; \psi).$$

Using the fact that observations on different patients are independent (so that Z_{it} and $\mathbf{u}_i$ are independent of $\mathbf{Y}_j$ for $j \neq i$), the E-step is

$$\begin{aligned}&\mathrm{E}[\log \mathcal{L}_c(\psi) \mid \mathbf{Y}, \psi^p] \\ &= \sum_{i=1}^{N} \mathrm{E}[\log \mathcal{L}_c^i(\psi) \mid \mathbf{Y}_i, \psi^p] \\ &= \sum_{i=1}^{N} \sum_{t=1}^{n_i} \sum_{k=1}^{K} \int \log f(y_{it} \mid Z_{it} = k, \mathbf{u}_i, \psi)\ \mathrm{P}(Z_{it} = k | \mathbf{y}_i, \psi^p) f(\mathbf{u}_i | \mathbf{y}_i, \psi^p)\ d\mathbf{u}_i \qquad \text{(A.9)} \\ &\quad + \sum_{i=1}^{N} \sum_{k=1}^{K} \log \pi_k\ \mathrm{P}(Z_{i1} = k | \mathbf{y}_i, \psi^p) \qquad \text{(A.10)} \\ &\quad + \sum_{i=1}^{N} \sum_{t=2}^{n_i} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \log P_{k,\ell}\ \mathrm{P}(Z_{i,t-1} = k, Z_{it} = \ell | \mathbf{y}_i, \psi^p) \qquad \text{(A.11)} \\ &\quad + \sum_{i=1}^{N} \int \log f(\mathbf{u}_i; \psi) f(\mathbf{u}_i \mid \mathbf{y}_i, \psi^p)\ d\mathbf{u}_i. \qquad \text{(A.12)}\end{aligned}$$

M-Step

Again, since each unknown parameter appears in exactly one of the terms (A.9)-(A.12), we can maximize each of these terms individually.

Now define the forward probabilities for patient i at the pth iteration as

$$W_{it}^p(k, \mathbf{u}_i) = f(y_{i1}, \ldots, y_{it} \mid Z_{it} = k, \mathbf{u}_i, \psi^p) \mathrm{P}(Z_{it} = k \mid \psi^p)$$

and the backward probabilities as

$$X_{it}^p(k, \mathbf{u}_i) = f(y_{i,t+1}, \ldots, y_{i,n_i} \mid Z_{it} = k, \mathbf{u}_i, \psi^p).$$

With a slight change of notation, the recursions (A.4) and (A.5) also apply to these new definitions:

$$W_{i,t+1}^p(k, \mathbf{u}_i) = \left(\sum_{\ell=1}^K W_{it}^p(\ell, \mathbf{u}_i) P_{\ell k}^p \right) f(y_{i,t+1} \mid Z_{it} = k, \mathbf{u}_i, \psi^p)$$

and

$$X_{it}^p(k, \mathbf{u}_i) = \sum_{\ell=1}^K f(y_{i,t+1} \mid Z_{i,t+1} = \ell, \mathbf{u}_i, \psi^p) X_{i,t+1}^p(\ell, \mathbf{u}_i) P_{k\ell}^p$$

with $W_{i1}^p(k, \mathbf{u}_i) = \pi_k^p \; f(y_{i1} \mid Z_{i1} = k, \mathbf{u}_i, \psi^p)$ and $X_{in}^p(k, \mathbf{u}_i) = 1$ (again, by convention).

Applying the same method used to derive (A.6), we can now show that the maximum value of (A.10) occurs at

$$\pi_k^{p+1} = \frac{1}{N} \sum_{i=1}^N \frac{\int W_{i1}^p(k, \mathbf{u}_i) X_{i1}^p(k, \mathbf{u}_i) f(\mathbf{u}_i; \psi^p) \; d\mathbf{u}_i}{\sum_{\ell=1}^K \int W_{i1}^p(\ell, \mathbf{u}_i) X_{i1}^p(\ell, \mathbf{u}_i) f(\mathbf{u}_i; \psi^p) \; d\mathbf{u}_i}. \tag{A.13}$$

Similarly, applying the method used to derive (A.7), it is clear that the maximum value of (A.11) occurs at

$$P_{k\ell}^{p+1} = \frac{P_{k\ell}^p \sum_{i=1}^N \sum_{t=2}^{n_i} \int W_{i,t-1}^p(k, \mathbf{u}_i) X_{it}^p(\ell, \mathbf{u}_i) f(y_{it} \mid Z_{it} = \ell, \mathbf{u}_i, \psi^p) f(\mathbf{u}_i; \psi^p) \; d\mathbf{u}_i}{\sum_{i=1}^N \sum_{t=2}^{n_i} \int W_{i,t-1}^p(k, \mathbf{u}_i) X_{i,t-1}^p(k, \mathbf{u}_i) f(\mathbf{u}_i; \psi^p) \; d\mathbf{u}_i}.$$

If $\{Z_{it}\}$ is not stationary, numerical maximization will be required in order to compute $P_{k\ell}^{p+1}$. In general, the terms (A.9) and (A.12) must also be maximized numerically, using, for example, a Gaussian quadrature technique. However, we can simplify the computations by noting that (A.9) can be written as

$$\sum_{i=1}^N \frac{\sum_{t=1}^{n_i} \sum_{k=1}^K \int \log f(y_{it} \mid Z_{it} = k, \mathbf{u}_i, \psi) \; W_{it}^p(k, \mathbf{u}_i) \; X_{it}^p(k, \mathbf{u}_i) \; f(\mathbf{u}_i; \psi^p) \; d\mathbf{u}_i}{\sum_{\ell=1}^K \int W_{i1}^p(\ell, \mathbf{u}_i) \; X_{i1}^p(\ell, \mathbf{u}_i) \; f(\mathbf{u}_i; \psi^p) \; d\mathbf{u}_i} \tag{A.14}$$

and (A.12) as

$$\sum_{i=1}^N \frac{\int \log f(\mathbf{u}_i; \psi) \sum_{\ell=1}^K W_{i1}^p(\ell, \mathbf{u}_i) \; X_{i1}^p(\ell, \mathbf{u}_i) \; f(\mathbf{u}_i; \psi^p) \; d\mathbf{u}_i}{\sum_{\ell=1}^K \int W_{i1}^p(\ell, \mathbf{u}_i) \; X_{i1}^p(\ell, \mathbf{u}_i) \; f(\mathbf{u}_i; \psi^p) \; d\mathbf{u}_i}. \tag{A.15}$$

Since $f(y_{it} \mid z_{it}, \mathbf{u}_i, \psi)$ is in the exponential family, $\log f(y_{it} \mid z_{it}, \mathbf{u}_i, \psi)$ will have a nice form. Likewise, if $f(\mathbf{u}_i; \psi)$ is in the exponential family, $\log f(\mathbf{u}_i; \psi)$ will also have a nice form. Thus, for certain choices of these functions, we would expect that the estimates of the parameters associated with these distributions would be quite easy to compute.

Thus, we see that, aside from the question of integration, the EM algorithm for the model with patient-specific random effects in the conditional model for the observed data is not much different from the case where the algorithm is applied to a single HMM. If the random effects are not patient-specific, this algorithm can still be used, but the expressions (A.9)-(A.12) will be more complicated since we will need to take the expectations conditional on the full data set, $\mathbf{Y}$, rather on each patient's data individually.

A.3 EM Algorithm for HMMs with Random Effects in the Hidden Process

In this section, we give the steps of the EM algorithm required to estimate the parameters of the model (5.13) assuming again that the random effects are patient-specific. We further assume that $\pi_k^i \equiv \pi_k$, i.e. that the initial probabilities are fixed, unknown parameters common to all patients.

As mentioned in Section A.2, the convergence properties of the EM algorithm are unknown in the context of HMMs for multiple processes. Wu (1983), however, provides sufficient conditions for convergence in a very general setting. In particular, the variance components must be strictly positive.

E-Step

For this model, thinking of the hidden states and the random effects as "missing" data, the complete likelihood is

$$\begin{aligned} \mathcal{L}_c(\psi) &= f(\mathbf{y} \mid \mathbf{u}, \mathbf{z}, \psi) f(\mathbf{z} \mid \mathbf{u}, \psi) f(\mathbf{u}; \psi) \\ &= \prod_i \left\{ \prod_{t=1}^{n_i} f(y_{it} \mid z_{it}, \mathbf{u}_i, \psi) \cdot \pi_{z_{i1}} \prod_{t=2}^{n_i} f(z_{it} \mid z_{i,t-1}, \mathbf{u}_i, \psi) \cdot f(\mathbf{u}_i; \psi) \right\}. \end{aligned}$$

So

$$\log \mathcal{L}_c^i(\psi) = \sum_{t=1}^{n_i} \log f(y_{it} \mid z_{it}, \mathbf{u}_i, \psi) + \log \pi_{z_{i1}} + \sum_{t=2}^{n_i} \log f(z_{it} \mid z_{i,t-1}, \mathbf{u}_i, \psi) + \log f(\mathbf{u}_i; \psi).$$

Then, using the assumption that the random effects are patient-specific,

$$\mathrm{E}[\log \mathcal{L}_c(\psi) \mid \mathbf{Y}, \psi^p]$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \mathrm{E}[\log \mathcal{L}_c^i(\psi) \mid \mathbf{Y}_i, \psi^p] \\
&= \sum_{i=1}^{N}\sum_{t=1}^{n_i}\sum_{k=1}^{K} \int \log f(y_{it} \mid Z_{it}=k, \mathbf{u}_i, \psi)\ \mathrm{P}(Z_{it}=k|\mathbf{y}_i, \mathbf{u}_i, \psi^p) f(\mathbf{u}_i|\mathbf{y}_i, \psi^p)\ d\mathbf{u}_i \quad \text{(A.16)} \\
&\quad + \sum_{i=1}^{N}\sum_{k=1}^{K} \log \pi_k\ \mathrm{P}(Z_{i1}=k|\mathbf{y}_i, \psi^p) \quad \text{(A.17)} \\
&\quad + \sum_{i=1}^{N}\sum_{t=2}^{n_i}\sum_{k=1}^{K}\sum_{\ell=1}^{K} \int \log \mathrm{P}(Z_{it}=\ell \mid Z_{i,t-1}=k, \mathbf{u}_i, \psi) \\
&\qquad \cdot \mathrm{P}(Z_{i,t-1}=k, Z_{it}=\ell|\mathbf{y}_i, \mathbf{u}_i, \psi^p) f(\mathbf{u}_i \mid \mathbf{y}_i, \psi^p)\ d\mathbf{u}_i \quad \text{(A.18)} \\
&\quad + \sum_{i=1}^{N} \int \log f(\mathbf{u}_i; \psi) f(\mathbf{u}_i \mid \mathbf{y}_i, \psi^p)\ d\mathbf{u}_i. \quad \text{(A.19)}
\end{aligned}
$$

M-Step

Again, we see that each unknown parameter appears in exactly one of the terms (A.16)-(A.19), so we can maximize each of these terms individually.

For this model, define the forward probabilities for patient i as

$$W_{it}^p(k, \mathbf{u}_i) = f(y_{i1}, \ldots, y_{it} \mid Z_{it}=k, \mathbf{u}_i, \psi^p)\mathrm{P}(Z_{it}=k \mid \mathbf{u}_i, \psi^p)$$

and the backward probabilities as

$$X_{it}^p(k, \mathbf{u}_i) = f(y_{i,t+1}, \ldots, y_{i,n_i} \mid Z_{it}=k, \mathbf{u}_i, \psi^p).$$

We then have the recursions

$$W_{i,t+1}^p(k, \mathbf{u}_i) = \left(\sum_{\ell=1}^{K} W_{it}^p(\ell, \mathbf{u}_i)\mathrm{P}(Z_{it}=k \mid Z_{i,t-1}=\ell, \mathbf{u}_i, \psi^p)\right) f(y_{i,t+1} \mid Z_{it}=k, \mathbf{u}_i, \psi^p)$$

and

$$X_{it}^p(k, \mathbf{u}_i) = \sum_{\ell=1}^{K} f(y_{i,t+1} \mid Z_{i,t+1}=\ell, \mathbf{u}_i, \psi^p) X_{i,t+1}^p(\ell, \mathbf{u}_i)\mathrm{P}(Z_{it}=\ell \mid Z_{i,t-1}=k, \mathbf{u}_i, \psi^p)$$

with $W_{i1}^p(k, \mathbf{u}_i) = \pi_k^p\ f(y_{i1} \mid Z_{i1}=k, \mathbf{u}_i, \psi^p)$ and $X_{in}^p(k, \mathbf{u}_i) = 1$ (again, by convention).

Since the parameters $\{\pi_k\}$ are fixed, the maximum value of (A.10) occurs at (A.13), as for Model I (but using the above definitions of $W_{it}^p(k, \mathbf{u}_i)$ and $X_{it}^p(k, \mathbf{u}_i)$). However, the evaluation of the integrals will be much more complicated in this case, for the reasons discussed in Section 5.4. Similarly, the expressions (A.16) and (A.19) are equivalent to (A.14) and (A.15), respectively, but will be much more difficult to compute.

For this model, we will also need to numerically maximize (A.18). However, by writing this expression in terms of the forward and backward probabilities we can obtain the simpler

form

$$\sum_{i=1}^{N} \frac{\sum_{t=2}^{n_i} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \int \log \mathrm{P}(Z_{it} = \ell \mid Z_{i,t-1} = k, \mathbf{u}_i; \psi)\ q_{it}^{p}(k, \ell, \mathbf{u}_i)\ f(\mathbf{u}_i; \psi^p)\ d\mathbf{u}_i}{\sum_{\ell=1}^{K} \int W_{i1}^{p}(\ell, \mathbf{u}_i)\ X_{i1}^{p}(\ell, \mathbf{u}_i)\ f(\mathbf{u}_i; \psi^p)\ d\mathbf{u}_i},$$

where

$$q_{it}^{p}(k, \ell, \mathbf{u}_i) = \mathrm{P}(Z_{it} = \ell \mid Z_{i,t-1} = k, \mathbf{u}_i; \psi^p)\ W_{i,t-1}^{p}(k, \mathbf{u}_i)\ X_{it}^{p}(\ell, \mathbf{u}_i)\ f(y_{it} \mid Z_{it} = \ell, \mathbf{u}_i; \psi^p).$$

Again, evaluating these integrals may prove difficult.

In summary, we see that the EM algorithm may be of only limited use in the estimation of the parameters of Model II because of the complex nature of the integrals involved.

Appendix B

Proofs

B.1 Proof of Lemma 3.1

Here we prove Lemma 3.1. For ease of exposition, we assume $m = 1$. However, the theory is valid for all finite values of k.

Since the measures μ_n are tight (as a result of Condition 2), there exists a subsequence $\{G_{n_j}\}$ such that G_{n_j} converges weakly to G, where G is a distribution function (Billingsley 1995, Theorem 29.3).

We have that

$$d_{KS}\{F(y,G),F(y,G_0)\} \leq d_{KS}\{F(y,G),F(y,G_{n_j})\} + d_{KS}\{F(y,G_{n_j}),F(y,G_0)\}. \quad \text{(B.1)}$$

By our hypothesis, the second term on the right-hand side is $o(1)$. We claim that the first term is also $o(1)$. Taking limits in (B.1) will imply that $d_{KS}\{F(y,G),F(y,G_0)\} = 0$, and hence, by Condition 5, that $G = G_0$. Therefore, G_{n_j} converges weakly to G_0. Since the measures μ_n are tight, and since every subsequence of $\{G_n\}$ that converges weakly at all converges weakly to G_0, we may conclude that G_n converges weakly to G_0 (Billingsley 1995, Theorem 29.3).

We now show that $d_{KS}\{F(y,G),F(y,G_{n_j})\} = o(1)$. By Condition 4, for all ϵ there exists $A > 0$ such that for all $(\theta,\phi) \in \Theta$, $H(A;\theta,\phi) - H(-A;\theta,\phi) \geq 1-\epsilon$. In particular, for $y \leq -A$, $H(y;\theta,\phi) \leq \epsilon$. So, for all $y \leq -A$,

$$F(y,G_{n_j}) = \int_\Theta H(y;\theta,\phi)dG_{n_j}(\theta,\phi) \leq \epsilon \int_\Theta dG_{n_j}(\theta,\phi) = \epsilon. \quad \text{(B.2)}$$

Likewise, for $y \leq -A$, $F(y,G) \leq \epsilon$. Similarly, for $y \geq A$, $1 - F(y,G_{n_j})$ and $1 - F(y,G)$ are also bounded by ϵ.

Since G_{n_j} converges weakly to G and $H(y;\theta,\phi)$ is continuous in θ and ϕ, we have that

$$F(y,G) - F(y,G_{n_j}) = \int H(y;\theta,\phi)d\{G(\theta,\phi) - G_{n_j}(\theta,\phi)\} \to 0 \tag{B.3}$$

for each y (Billingsley 1995, Theorem 29.1).

We claim that $F(y,G_{n_j})$ converges not only pointwise but uniformly to $F(y,G)$ on the interval $[-A,A]$. It is enough to show that $F(y_j,G_{n_j})$ converges to $F(y,G)$ for all sequences $y_j \downarrow y$, with $y_j, y \in [-A,A]$. (See Strichartz 1995, Theorem 7.3.5, with a slight modification to account for the fact that $\{F(\cdot,G_{n_j})\}$ are right-continuous rather than continuous.)

Fix ϵ, y, and a sequence $\{y_j\}$ such that $y_j \downarrow y$. Define $F_j(y) \equiv F(y,G_{n_j})$ and $F(y) \equiv F(y,G)$. Now

$$|F_j(y_j) - F(y)| \leq |F_j(y_j) - F_j(y)| + |F_j(y) - F(y)|. \tag{B.4}$$

By the right-continuity of F, there exists $\delta > 0$ such that

$$|F(y+\delta) - F(y)| \leq \epsilon. \tag{B.5}$$

By (B.3), there exists N such that for $j \geq N$,

$$|F_j(y) - F(y)| \leq \epsilon \tag{B.6}$$

and there exists N_δ such that for $j \geq N_\delta$,

$$|F_j(y+\delta) - F(y+\delta)| \leq \epsilon. \tag{B.7}$$

Since $y_j \downarrow y$ and F_j is non-decreasing, there exists N_1 such that for $j \geq N_1$,

$$|F_j(y_j) - F_j(y)| \leq |F_j(y+\delta) - F_j(y)|. \tag{B.8}$$

Combining results (B.4)–(B.8), we have that for $j \geq \max(N, N_\delta, N_1)$,

$$\begin{aligned}
&|F_j(y_j) - F(y)| \\
&\quad\leq |F_j(y+\delta) - F_j(y)| + |F_j(y) - F(y)| \\
&\quad\leq |F_j(y+\delta) - F(y+\delta)| + |F(y+\delta) - F(y)| + |F(y) - F_j(y)| + |F_j(y) - F(y)| \\
&\quad\leq 4\epsilon.
\end{aligned}$$

Thus, F_j converges uniformly to F on the interval $[-A,A]$, implying that for all ϵ there exists N^* such that for all $j \geq N^*$ and for all $y \in [-A,A]$, $|F(y,G) - F(y,G_{n_j})| \leq \epsilon$. Using (B.2), we see that this result holds for all y, and hence

$$d_{KS}\{F(y,G), F(x,G_{n_j})\} = o(1),$$

as desired. □

Although Lemma 3.1 is stated in terms of deterministic mixing distributions, we in fact require that for a sequence $\{\hat{G}_n\}$ of estimated mixing distributions, $d_{KS}\{F(y,\hat{G}_n), F(y,G_0)\} = o(1)$ a.s. implies that $\hat{G}$ converges weakly to G_0 a.s. However, it is easy to see that this requirement is met when Conditions 2–5 are satisfied. Denote by $\hat{G}_n(\omega)$ the terms of the estimated sequence of mixing distributions associated with the sample point ω. Since $(\hat{\theta}_j^0, \hat{\phi}^0) \in \Theta$ for all j, Lemma 3.1 applies to each sequence $\{\hat{G}_n(\omega)\}$ for which $d_{KS}\{F(y,\hat{G}_n(\omega)), F(y,G_0)\} = o(1)$.

B.2 Proof of Theorem 6.1

To develop the necessary bounds, we begin by examining terms of the form

$$\mathrm{E}\left|\mathbf{D}_{m_1}^i \mathbf{D}_{m_2}^i\right|$$

where $m_1, m_2 \leq M$. The extension to the case where we have a product of more than two such derivatives is straightforward. Our proof relies on the fact that $\mathbf{D}_m^i$ is a derivative of $\log f(\mathbf{y}_i \mid u_i, \psi)$, and that we compute the expectation of products of these derivatives with respect to the null distribution, $f(\mathbf{y}_i \mid u_i = 0, \psi)$. For our model, $f(\mathbf{y}_i \mid u_i, \psi)$ is the likelihood of a stationary HMM. In this way, our problem effectively reduces to that considered by BRR.

To avoid having to work with mixed partial derivatives, we will use the claim of BRR that, without loss of generality, ξ can be taken to be unidimensional. (Recall that $\psi = (\xi, D)$, so that ξ does not include parameters associated with the random effect.) Also following BRR, any operation between two sequences is understood to be performed termwise. For example, for sequences $\mathbf{I} = (I_1, \ldots, I_m)$ and $\mathbf{J} = (J_1, \ldots, J_m)$, and functions $\{f_i\}_{i=1}^m$, we would have

$$\prod \mathbf{J}! = \prod_{j=1}^m J_j! \quad \text{and} \quad \frac{f_{\mathbf{I}}}{\mathbf{J}!} = \left(\frac{f_{I_1}}{J_1!}, \ldots, \frac{f_{I_m}}{J_m!}\right).$$

Define $\mathbf{Y}_{i,s}^t = (Y_{is}, \ldots, Y_{it})$. We let $h_i(1) \equiv \log f(Y_{i1}, Z_{i1} | u_i, \psi)$, and, for $t > 1$, we define $h_i(t) \equiv \log f(Y_{it}, Z_{it} | \mathbf{Y}_{i,1}^{t-1}, \mathbf{Z}_{i,1}^{t-1}, u_i, \psi)$. As a result of these definitions, we have that $\log f(\mathbf{Y}_i, \mathbf{Z}_i \mid u_i) = \sum_{t=1}^{n_i} h_i(t)$. Let $h_i^{(m)}(t)$ represent the mth derivative of $h_i(t)$ with respect to ξ, evaluated at $u_i = 0$. By Conditions 1 and 2, these derivatives exist for $m \leq M$.

Using the notation of BRR, we denote the cumulant of a random vector $\mathbf{X} = (X_1, \ldots, X_m)$ by

$$\Gamma(\mathbf{X}) \equiv \Gamma(X_1, \ldots, X_m) \equiv \frac{1}{\iota^m} \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_m} \log\left(\mathrm{E}\left[e^{\iota(x_1 X_1 + \cdots x_m X_m)}\right]\right)\Big|_{x_1 = \cdots = x_m = 0},$$

where $\iota = \sqrt{-1}$. The cumulant of the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ will be denoted by $\Gamma^{\mathbf{Y}}(\mathbf{X})$. Finally, let $\chi'(X_1) = X_1$, and let $\chi(X_1) = \mathrm{E}[X_1]$. We define the centred moment function (Saulis & Statulevičious 1991), χ, recursively by

$$\begin{aligned}\chi'(X_1, \ldots, X_m) &= X_1(\chi'(X_2, \ldots, X_m) - \chi(X_2, \ldots, X_m)) \\ \chi(X_1, \ldots, X_m) &= \mathrm{E}[\chi'(X_1, \ldots, X_m)],\end{aligned}$$

$m = 2, 3, \ldots$. We denote the centred moment of the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ by $\chi^{\mathbf{Y}}(\mathbf{X})$.

In their Proposition 3.1, BRR show how to write the derivatives of the log-likelihood of a single, stationary HMM as a linear combination of cumulants. It turns out that their result also holds for derivatives of $\log f(\mathbf{y}_i \mid u_i, \psi)$ if we replace the distribution of each random variable in their Equation 4 with its distribution conditional on u_i. In particular,

$$\begin{aligned}\left|\mathbf{D}^i_m\right| &= \left|\sum_{\mathbf{J}\in\mathcal{J}^+(m)} \frac{m!}{|\mathbf{J}|!}\Gamma^{\mathbf{Y}_i,u_i}\left(\frac{\sum_{t=1}^{n_i} h_i^{(\mathbf{J})}(t)}{\mathbf{J}!}\right)\right| \\ &\leq \sum_{\mathbf{J}\in\mathcal{J}^+(m)} \frac{m!}{|\mathbf{J}|!}\left|\Gamma^{\mathbf{Y}_i,u_i}\left(\frac{\sum_{t=1}^{n_i} h_i^{(\mathbf{J})}(t)}{\mathbf{J}!}\right)\right|. \qquad \text{(B.9)}\end{aligned}$$

BRR provide bounds on terms of the form $\left|\Gamma^{\mathbf{Y}_i,u_i}\left(\sum_{t=1}^{n_i} h_i^{(\mathbf{J})}(t)\right)\right|$. Specifically, from their Lemma 3.3(i),

$$\begin{aligned}\left|\Gamma^{\mathbf{Y}_i,u_i}\left(\sum_{t=1}^{n_i} h_i^{(\mathbf{J})}(t)\right)\right| \leq \sum_{t=1}^{n_i} \sum_{\substack{\mathbf{I}\in\mathcal{J}_1^{n_i}(|\mathbf{J}|) \\ \min \mathbf{I}^1 = t}} \sum_{v=1}^{|\mathbf{J}|} \sum_{\biguplus K_q = \{1,\ldots,|\mathbf{J}|\}} M_v(K_1, \ldots, K_v) \\ \cdot \prod_{q=1}^{v} \left|\chi^{\mathbf{Y}_i,u_i}\left(h_i^{(\mathbf{J}(K_q))}(\mathbf{I}(K_q))\right)\right| \qquad \text{(B.10)}\end{aligned}$$

where $\biguplus K_q = \{1, \ldots, |\mathbf{J}|\}$ denotes the set of all v-block partitions of the set $\{1, \ldots, |\mathbf{J}|\}$, M_v are non-negative combinatorial constants, $\mathbf{J}(K_q) = (J_{K_{q_1}}, J_{K_{q_2}}, \ldots)$, and $\mathbf{I}(K_q) = (I_{K_{q_1}}, I_{K_{q_2}}, \ldots)$. Then, letting $\Delta(\mathbf{I}) = \max\{I_j\} - \min\{I_j\}$, Equation 11 of BRR gives the following relationship:

$$\prod_{q=1}^{v} \left|\chi^{\mathbf{Y}_i,u_i}\left(h_i^{(\mathbf{J}(K_q))}(\mathbf{I}(K_q))\right)\right| \leq 2^{|\mathbf{J}|-v}\rho^{\sum_{q=1}^{v}\Delta(\mathbf{I}(K_q))} \prod_{j=1}^{|\mathbf{J}|} C^i_{J_j}(Y_{I_j}, \psi).$$

where $\rho = 1 - \min\left\{\min_{k,\ell} P_{k\ell}, \min_{k,\ell} P^*_{k\ell}\right\}$, with $P^*_{k\ell} = \pi_\ell P_{\ell k}/\pi_k$. Under Condition 1, BRR show that $0 < \rho < 1$.

We can now use these results to establish bounds on $\mathrm{E}\left|\mathbf{D}^i_{m_1}\mathbf{D}^i_{m_2}\right|$. From the multilinearity of the cumulant function and (B.9), we can write

$$\mathrm{E}\left|\mathbf{D}^i_{m_1}\mathbf{D}^i_{m_2}\right| \leq \mathrm{E}\left\{\sum_{\mathbf{J}^1\in\mathcal{J}^+(m_1)} \frac{m_1!}{|\mathbf{J}^1|!\prod \mathbf{J}^1!}\left|\Gamma^{\mathbf{Y}_i,u_i}\left(\sum_{t=1}^{n_i} h_i^{(\mathbf{J}^1)}(t)\right)\right|\right.$$

$$\cdot \sum_{\mathbf{J}^2 \in \mathcal{J}^+(m_2)} \frac{m_2!}{|\mathbf{J}^2|! \prod \mathbf{J}^2!} \left| \Gamma^{\mathbf{Y}_i, u_i} \left(\sum_{t=1}^{n_i} h_i^{(\mathbf{J}^2)}(t) \right) \right| \Bigg\}. \tag{B.11}$$

Using the bound (B.10), the right-hand side of (B.11) becomes a linear combination of terms of the form

$$\mathrm{E} \left\{ \prod_{q_1=1}^{v_1} \left| \chi^{\mathbf{Y}_i, u_i} \left(h_i^{(\mathbf{J}^1(K_{q_1}^1))}(\mathbf{I}^1(K_{q_1}^1)) \right) \right| \cdot \prod_{q_2=1}^{v_2} \left| \chi^{\mathbf{Y}_i, u_i} \left(h_i^{(\mathbf{J}^2(K_{q_2}^2))}(\mathbf{I}^2(K_{q_2}^2)) \right) \right| \right\}.$$

Let the values $t_1 < t_2 < \cdots < t_r$ denote the distinct elements of the union of $\mathbf{I}_1$ and $\mathbf{I}_2$. Using the technique in Equation 12 of BRR and the definition of $B^i_{m_1 m_2}(\psi)$,

$$\begin{aligned}
& \mathrm{E} \left\{ \prod_{q_1=1}^{v_1} \left| \chi^{\mathbf{Y}_i, u_i} \left(h_i^{(\mathbf{J}^1(K_{q_1}^1))}(\mathbf{I}^1(K_{q_1}^1)) \right) \right| \cdot \prod_{q_2=1}^{v_2} \left| \chi^{\mathbf{Y}_i, u_i} \left(h_i^{(\mathbf{J}^2(K_{q_2}^2))}(\mathbf{I}^2(K_{q_2}^2)) \right) \right| \right\} \\
\leq\ & 2^{|\mathbf{J}^1|-1} \rho^{\sum_{q_1=1}^{v_1} \Delta(\mathbf{I}_1(K_{q_1}^1))} \cdot 2^{|\mathbf{J}^2|-1} \rho^{\sum_{q_2=1}^{v_2} \Delta(\mathbf{I}_2(K_{q_2}^2))} \\
& \cdot \max_z \prod_{s=1}^{r} \mathrm{E} \left\{ \prod_{\substack{j_1 \in \{1,\ldots,|\mathbf{J}^1|\}: \\ t_s = I_{j_1}^1}} C^i_{J_{j_1}^1}(Y_{i,t_s}, \psi) \cdot \prod_{\substack{j_2 \in \{1,\ldots,|\mathbf{J}^2|\}: \\ t_s = I_{j_2}^2}} C^i_{J_{j_2}^2}(Y_{i,t_s}, \psi) \,\middle|\, Z_{i,t_s} = z \right\} \\
=\ & 2^{|\mathbf{J}^1|-1} \rho^{\sum_{q_1=1}^{v_1} \Delta(\mathbf{I}_1(K_{q_1}^1))} \cdot 2^{|\mathbf{J}^2|-1} \rho^{\sum_{q_2=1}^{v_2} \Delta(\mathbf{I}_2(K_{q_2}^2))} \cdot \prod \mathbf{J}^1! \prod \mathbf{J}^2! \\
& \cdot \max_z \prod_{s=1}^{r} \mathrm{E} \left\{ \prod_{\substack{j_1 \in \{1,\ldots,|\mathbf{J}^1|\}: \\ t_s = I_{j_1}^1}} \frac{C^i_{J_{j_1}^1}(Y_{i,t_s}, \psi)}{J_{j_1}^1!} \cdot \prod_{\substack{j_2 \in \{1,\ldots,|\mathbf{J}^2|\}: \\ t_s = I_{j_2}^2}} \frac{C^i_{J_{j_2}^2}(Y_{i,t_s}, \psi)}{J_{j_2}^2!} \,\middle|\, Z_{i,t_s} = z \right\} \\
\leq\ & 2^{|\mathbf{J}^1|-1} \rho^{\sum_{q_1=1}^{v_1} \Delta(\mathbf{I}_1(K_{q_1}^1))} \cdot 2^{|\mathbf{J}^2|-1} \rho^{\sum_{q_2=1}^{v_2} \Delta(\mathbf{I}_2(K_{q_2}^2))} \cdot \prod \mathbf{J}^1! \prod \mathbf{J}^2! \cdot B^i_{m_1 m_2}(\psi).
\end{aligned}$$

Then, from (B.10) and Lemma 3.3(ii) of BRR, it is clear that

$$\begin{aligned}
& \mathrm{E} \left\{ \left| \Gamma^{\mathbf{Y}_i, u_i} \left(\sum_{t=1}^{n_i} h_i^{(\mathbf{J}^1)}(t) \right) \right| \left| \Gamma^{\mathbf{Y}_i, u_i} \left(\sum_{t=1}^{n_i} h_i^{(\mathbf{J}^2)}(t) \right) \right| \right\} \\
\leq\ & n_i |\mathbf{J}^1|! \left(\frac{8}{1-\rho} \right)^{|\mathbf{J}^1|-1} n_i |\mathbf{J}^2|! \left(\frac{8}{1-\rho} \right)^{|\mathbf{J}^2|-1} \prod \mathbf{J}^1! \prod \mathbf{J}^2! \, B^i_{m_1 m_2}(\psi)
\end{aligned}$$

Substituting into (B.11) and using Lemma 4.1(ii) of BRR, we conclude that

$$\mathrm{E} \left| \mathbf{D}^i_{m_1} \mathbf{D}^i_{m_2} \right| \leq n_i^2 m_1! m_2! B^i_{m_1 m_2}(\psi) \left(1 + \frac{8}{1-\rho} \right)^{m_1+m_2-2}$$

and, more generally, that

$$\mathrm{E} \left| \mathbf{D}^i_{m_1} \times \cdots \times \mathbf{D}^i_{m_d} \right| \leq n_i^d m_1! \times \cdots \times m_d! B^i_{m_1 \cdots m_d}(\psi) \left(1 + \frac{8}{1-\rho} \right)^{\sum_{j=1}^{d} m_j - d}.$$

This proves our theorem. □

B.3 Proof of Theorem 6.2

To prove Theorem 6.2, it suffices to show that $B^i_{m_1 \cdots m_d}(\psi)$ is finite and independent of i.

To this end, we first study the random variable $C^i_m(Y_{it}, \psi)$. As an aside, BRR's definition of $C^i_m(y_{it}, \psi)$ also involves a supremum over all values of the parameters in the neighbourhood of the true values of $\{\xi_k\}$. These authors require this generalization since they apply their results to problems relating to the MLEs, but it is unnecessary for our purposes.

Under Condition 1, $0 < P_{k\ell} < 1$, and hence $0 < \pi_k < 1$, for all k, ℓ. We then have

$$\max_{\mathbf{D}_m} \max_{k,\ell} \left\{ \left| \mathbf{D}_m \log P_{k\ell} \right| + \left| \mathbf{D}_m \log \pi_k \right| \right\} < \infty$$

for all m. Furthermore, this quantity is independent of i and $\{Y_{it}\}$. Thus, for our purposes, we ignore these terms, and focus our attention on the term

$$\max_{\mathbf{D}_m} \max_{k} \left| \mathbf{D}_m \log f(y_{it} \mid Z_{it} = k, u_i, \psi) \Big|_{u_i=0} \right|$$

for $m \leq M$.

When $a(\phi)$ is a constant (e.g. when $f(y_{it} \mid z_{it}, u_i, \psi)$ is in the Poisson or binomial family), we need only be concerned with derivatives with respect to $\{\tau_k\}$ and u_i. Since for our model $\eta_{itk} = \tau_k + u_i$, we can equivalently consider derivatives with respect to η_{itk}. We have that

$$\frac{\partial}{\partial \eta_{itk}} \log f(y_{it} \mid Z_{it} = k, u_i, \psi) \Big|_{u_i=0} = \left. \frac{y_{it} - \frac{\partial}{\partial \eta_{itk}} c(\eta_{itk})}{a(\phi)} \right|_{u_i=0} \tag{B.12}$$

and

$$\frac{\partial^m}{\partial^m \eta_{itk}} \log f(y_{it} \mid Z_{it} = k, u_i, \psi) \Big|_{u_i=0} = \left. -\frac{\frac{\partial^m}{\partial^m \eta_{itk}} c(\eta_{itk})}{a(\phi)} \right|_{u_i=0}, \tag{B.13}$$

$2 \leq m \leq M$. When $f(y_{it} \mid z_{it}, u_i, \psi)$ is in the Poisson family, $c(\eta) = e^\eta$, and when $f(y_{it} \mid z_{it}, u_i, \psi)$ is in the binomial family, $c(\eta) = \log(1 + e^\eta)$. Since, by assumption, $\{\tau_k\}$ are bounded, in both of these cases, all derivatives of $c(\eta_{itk})$ (and hence expectations of products of derivatives of the form (B.13)) will be bounded and independent of i. When derivatives of the form (B.12) appear as factors in the product of interest, we note that the absolute moments of $Y_{it} \mid Z_{it}, u_i$ are finite and independent of i in both the Poisson and binomial families. Thus, $B^i_{m_1 \cdots m_d}(\psi)$ is also finite and independent of i, which proves Theorem 6.2 for these cases.

When $a(\phi)$ is not constant, we need to consider partial derivatives with respect to ϕ as well as $\{\tau_k\}$ and u_i. When these functions include at least one derivative with respect to $\{\tau_k\}$ or u_i, we have

$$\frac{\partial^{\ell+1}}{\partial \eta_{itk} \partial^\ell \phi} \log f(y_{it} \mid Z_{it} = k, u_i, \psi) \Big|_{u_i=0} = \left(y_{it} - \frac{\partial}{\partial \eta_{itk}} c(\eta_{itk}) \Big|_{u_i=0} \right) \frac{\partial^\ell}{\partial^\ell \phi} a^{-1}(\phi), \tag{B.14}$$

$\ell = 0, 1, \ldots, M-1$, and

$$\frac{\partial^{\ell+m}}{\partial^m \eta_{itk} \partial^\ell \phi} \log f(y_{it} \mid Z_{it} = k, u_i, \psi)\Big|_{u_i=0} = -\frac{\partial^m}{\partial^m \eta_{itk}} c(\eta_{itk})\Big|_{u_i=0} \frac{\partial^\ell}{\partial^\ell \phi} a^{-1}(\phi), \tag{B.15}$$

$\ell + m \leq M$, $m \geq 2$. Since $\{\tau_k\}$ and ϕ are bounded, as long as derivatives of $\frac{1}{a(\phi)}$ and $c(\eta_{itk})$ are well-behaved and the absolute moments of $Y_{it} \mid Z_{it}, u_i$ are finite and independent of i, the expected value of products of such terms are finite and independent of i. In particular, when $f(y_{it} \mid z_{it}, u_i)$ is in the normal family with parameterization

$$\log f(y_{it} \mid Z_{it} = k, u_i, \psi) = \frac{\eta_{itk} y_{it} - \frac{1}{2}\eta_{itk}^2}{\phi} - \frac{y_{it}^2}{2\phi} - \frac{1}{2}\log(2\pi\phi), \tag{B.16}$$

then all absolute moments are finite, and derivatives of $\frac{1}{a(\phi)} \equiv \frac{1}{\phi}$ and $c(\eta_{itk}) \equiv \frac{1}{2}\eta_{itk}^2$ are bounded since $\{\tau_k\}$ and ϕ are bounded. Likewise, when $f(y_{it} \mid z_{it}, u_i)$ is in the gamma family with parameterization

$$\log f(y_{it} \mid Z_{it} = k, u_i, \psi) = \frac{\eta_{itk} y_{it} + \log(-\eta_{itk})}{\frac{1}{\phi}} + \phi \log \phi + (\phi - 1)\log y_{it} - \log \Gamma(\phi), \tag{B.17}$$

then again, we see that all absolute moments are finite, and that derivatives of $\frac{1}{a(\phi)} \equiv \phi$ and $c(\eta_{itk}) \equiv -\log(-\eta_{itk})$ are bounded since $\{\tau_k\}$ and ϕ are bounded.

Derivatives with respect to ϕ only have the form

$$\frac{\partial^m}{\partial^m \phi} \log f(y_{it} \mid Z_{it} = k, u_i, \psi)\Big|_{u_i=0} = [y_{it}\tau_k - c(\tau_k)] \frac{\partial^m}{\partial^m \phi} a^{-1}(\phi) + \frac{\partial^m}{\partial^m \phi} d(y_{it}, \phi).$$

Verifying that the expected value of products of such terms are finite and independent of i is more complicated since $\frac{\partial^m}{\partial^m \phi} d(y_{it}, \phi)$ is a function of y_{it}. However, we can do so for the special cases we consider. For example, when $f(y_{it} \mid z_{it}, u_i, \psi)$ is the normal distribution parameterized as in (B.16),

$$\begin{aligned} &\frac{\partial^m}{\partial^m \phi} \log f(y_{it} \mid Z_{it} = k, u_i, \psi)\Big|_{u_i=0} \\ &= \left\{\left[\tau_k y_{it} - \frac{1}{2}\tau_k^2\right] - \frac{1}{2}y_{it}^2\right\}(-1)^m m! \phi^{-m-1} - \frac{1}{2}(-1)^{m-1}(m-1)!\phi^{-m} \end{aligned} \tag{B.18}$$

Since ϕ and $\{\tau_k\}$ are bounded and the absolute moments of the normal distribution are finite, the expected value of products of terms such as (B.14), (B.15), and (B.18) are finite and independent of i.

Similarly, in the case of the gamma distribution parameterized as in (B.17),

$$\begin{aligned} &\frac{\partial^m}{\partial^m \phi} \log f(y_{it} \mid Z_{it} = k, u_i, \psi)\Big|_{u_i=0} \\ &= \begin{cases} \tau_k y_{it} + \log(-\tau_k) + \log \phi + 1 + \log y_{it} - \frac{\partial}{\partial \phi}\log \Gamma(\phi), & m = 1 \\ (-1)^m (m-1)! \phi^{-m+1} - \frac{\partial^m}{\partial^m \phi}\log \Gamma(\phi), & m \geq 2 \end{cases} \end{aligned} \tag{B.19}$$

Since ϕ and $\{\tau_k\}$ are bounded, using the facts that $\log y_{it} < y_{it}$ and that the absolute moments of the gamma distribution are finite, the expected value of products of terms such as (B.14), (B.15), and (B.19) are also finite and independent of i.

Thus, we have shown that $B^i_{m_1 \cdots m_d}(\psi)$ is finite and independent of i in the case of the normal and gamma distributions. This completes our proof of Theorem 6.2. $\square$

Bibliography

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

[2] Albert, P.S., McFarland, H.F., Smith, M.E., and Frank, J.A. (1994). Time series for modelling counts from a relapsing-remitting disease: application to modelling disease activity in multiple sclerosis. *Statistics in Medicine*, **13**, 453–466.

[3] Alzaid, A.A. and Al-Osh, M.A. (1993). Some autoregressive moving average processes with generalized Poisson marginal distributions. *Annals of the Institute of Statistical Mathematics*, **45**, 223–232.

[4] Baras, J.S. and Finesso, L. (1992). Consistent estimation of the order of hidden Markov chains. In *Stochastic Theory and Adaptive Control (Lawrence, KS, 1991)*, 26–39. Springer-Verlag, Berlin.

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

[6] Bickel, P.J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics*, **26**, 1614–1635.

[7] Bickel, P.J., Ritov, Y., and Rydén, T. (2002). Hidden Markov model likelihoods and their derivatives behave like IID ones. *Annales de L'Institut Henri Poincaré - Probabilités et Statistiques*, **38**, 825–846.

[8] Billingsley, P. (1995). *Probability and Measure.* John Wiley & Sons, New York.

[9] Blais, M., MacGibbon, B., and Roy, R. (2000). Limit theorems for regression models of times series of counts. *Statistics & Probability Letters*, **46**, 161–168.

[10] Bradley, R.A. and Gart. J.J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika*, **49**, 205–214.

[11] Breiman, L. (1968). *Probability.* Addison-Wesley Publishing Company, Reading, Massachusetts.

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

[13] Chen, J. and Kalbfleisch, J.D. (1996). Penalized minimum-distance estimates in finite mixture models. *The Canadian Journal of Statistics*, **24**, 167–175.

[14] Chen, M.-H. and Ibrahim, J.G. (2000). Bayesian predictive inference for time series count data. *Biometrics*, **56**, 678-685.

[15] Chesher, A. (1984). Testing for neglected heterogeneity. *Econometrica*, **52**, 865–872.

[16] Chung, S.H., Moore, J.B., Xia, L., Premkumar, L.S., and Gage, P.W. (1990). Characterization of single channel currents using digital signal processing techniques based on hidden Markov models. *Philosophical Transactions of the Royal Society of London B*, **329**, 265–285.

[17] Dean, C.B. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, **87**, 451–457.

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

[19] Doukhan, P. (1994). *Mixing: Properties and Examples.* Springer-Verlag, New York.

[20] Dortet-Bernadet, V. (2001). Choix de modèle pour des chaînes de Markov cachées. *Comptes Rendus des Séances de l'Académie des Sciences. Série I*, **332**, 469–472.

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

[22] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

[23] Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, **10**, 254–272.

[24] Feng, Z. and McCulloch, C.E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics & Probability Letters*, **13**, 325–332.

[25] Giudici, P., Rydén, T., and Vandekerkhove, P. (2000). Likelihood-ratio tests for hidden Markov models. *Biometrics*, **56**, 742–747.

[26] Hall, D.B. and Præstgaard, J.T. (2001) Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika*, **88**, 739–751.

[27] Hettmansperger, T.P. and Thomas, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society B, 62*, 811–825.

[28] Hughes, J.P. and Guttorp, P. (1994). A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resources Research*, **30**, 1535–1546.

[29] Humphreys, K. (1997). Classification error adjustments for female labour force transitions using a latent Markov chain with random effects. In *Applications of Latent Class and Latent Trait Models in the Social Sciences* (ed. J. Rost and R. Langeheine), 370–380. Waxmann, Münster.

[30] Humphreys, K. (1998). The latent Markov chain with multivariate random effects: an evaluation of instruments measuring labour market status in the British Household Panel Study. *Sociological Methods and Research*, **26**, 269–299.

[31] Ibragimov, I.A. and Linnik, Y.V. (1971). *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff Publishing, The Netherlands.

[32] Jacqmin-Gadda, H. and Commenges, D. (1995). Tests of homogeneity for generalized linear models. *Journal of the American Statistical Association*, **90**, 1237–1246.

[33] Juang, B.H. and Rabiner, L.R. (1991). Hidden Markov models for speech recognition. *Technometrics*, **33**, 251–272.

[34] Kieffer, J.C. (1993). Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Transactions on Information Theory*, **39**, 893–902.

[35] Krogh, A. (1998). An introduction to hidden Markov models for biological sequences. In *Computational Methods in Molecular Biology* (ed. S.L. Salzberg, D.B. Searls, and S. Kasif), 45–63. Elsevier, Amsterdam.

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

[37] Leroux, B.G. (1992a). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, **40**, 127–143.

[38] Leroux, B.G. (1992b). Consistent estimation of a mixing distribution. *Annals of Statistics*, **20**, 1350–1360.

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

[40] Levinson, S.E., Rabiner, L.R., and Sondhi, M.M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, **62**, 1035–1074.

[41] Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications*, **65**, 69–80.

[42] Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84**, 309–326.

[43] Lin, Z. and Lu, C. (1996). *Limit Theory for Mixing Dependent Random Variables.* Kluwer Academic Publishers, Boston.

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

[45] Liu, C.-C. and Narayan, P. (1994). Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures. *IEEE Transactions on Information Theory*, **40**, 1167–1180.

[46] Lockhart, R.A. and Stephens, M.A. (1998). The probability plot: tests of fit based on the correlation coefficient. In *Handbook of Statistics 17*, 453–473. Elsevier Science, Amsterdam.

[47] Lystig, T.C. (2001). *Evaluation of Hidden Markov Models.* Ph.D. thesis, University of Washington, Seattle.

[48] MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov Models and Other Models for Discrete-Valued Time Series.* Chapman & Hall, London.

[49] MacKay, R.J. (2002). Estimating the order of a hidden Markov model. *The Canadian Journal of Statistics*, **30**, 573–589.

[50] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

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

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

[53] McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, **20**, 822–835.

[54] Nash, J.C. (1979). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation.* John Wiley & Sons, Inc., New York.

[55] Ould–Saïd, E. (1994). Loi du log itéré pour la fonction de répartition empirique dans le cas multidimensionel et α-mélangeant. *Comptes Rendus des Séances de l'Académie des Sciences. Série I*, **318**, 759–763.

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

[57] Poskitt, D.S. and Chung, S.-H. (1996). Markov chain models, time series analysis and extreme value theory. *Advances in Applied Probability*, **28**, 405–425.

[58] Prakasa Rao, B.L.S. (1992). *Identifiability in Stochastic Models: Characterization of Probability Distributions.* Academic Press, Inc., London.

[59] PRISMS (Prevention of Relapses and Disability by Interferon β-1a Subcutaneously in Multiple Sclerosis) Study Group (1998). Randomised double-blind placebo-controlled study of interferon β-1a in relapsing/remitting multiple sclerosis. *The Lancet*, **352**, 1498–1504.

[60] Raubertas, R.F. (1992). The envelope probability plot as a goodness-of-fit test. *Communications in Statistics – Simulation and Computation*, **21**, 189–202.

[61] Robert, C.P., Rydén, T., and Titterington, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society Series B*, **62**, 57–75.

[62] Rydén, T. (1995). Estimating the order of hidden Markov models. *Statistics*, **26**, 345–354.

[63] Rynkiewicz, J. (2001). Estimation of hybrid HMM/MLP models. In *Proceedings of the European Symposium on Artificial Neural Networks.* Bruges, Belgium, 383–389.

[64] Saulis, L. and Statulevičius, V.A. (1991). *Limit Theorems for Large Deviations.* Kluwer, Dordrecht.

[65] Self, S.G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

[66] Seltman, H.J. (2002). Hidden Markov models for analysis of biological rhythm data. In *Case Studies in Bayesian Statistics V* (ed. C. Gatsonis, R.E. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli, and M. West), 398–406. Springer, New York.

[67] Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.

[68] Strichartz, R.S. (1995). *The Way of Analysis.* Jones and Barlett Publishers, Boston.

[69] Teicher, H. (1967). Identifiability of mixtures of product measures. *Annals of Mathematical Statistics*, **38**, 1300–1302.

[70] Turner, T.R., Cameron, M.A., and Thomson, P.J. (1998). Hidden Markov chains in generalized linear models. *The Canadian Journal of Statistics*, **26**, 107–125.

[71] Wang, P. and Puterman, M.L. (1999). Markov Poisson regression models for discrete time series. *Journal of Applied Statistics*, **26**, 855–869.

[72] White, L.B., Mahony, R., and Brushe, G.D. (2000). Lumpable hidden Markov models – model reduction and reduced complexity filtering. *IEEE Transactions on Automatic Control*, **45**, 2297–2306.

[73] Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, **11**, 95–103.

[74] Zucchini, W. and Guttorp, P. (1991). A hidden Markov model for space-time precipitation. *Water Resources Research*, **27**, 1917–1923.