A Penalized Quasi-Likelihood Approach for Estimating the Number of States in a Hidden Markov Model

Annaliza McGillivray

Department of Mathematics and Statistics

McGill University

Montreal, Quebec

August 2012

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Master of Science

Abstract

In statistical applications of hidden Markov models (HMMs), one may have no knowledge of the number of hidden states (or order) of the model needed to be able to accurately represent the underlying process of the data. The problem of estimating the number of states of the HMM is thus a task of major importance. We begin with a literature review of the major developments in the problem of order estimation for HMMs. We then propose a new penalized quasi-likelihood method for estimating the number of hidden states, which makes use of the fact that the marginal distribution of the HMM observations is a finite mixture model. Starting with a HMM with a large number of states, the method obtains a model of lower order by clustering and merging similar states of the model through two penalty functions. We study some of the asymptotic properties of the proposed method and present a numerical procedure for its implementation. The performance of the new method is assessed via extensive simulation studies for normal and Poisson HMMs. The new method is more computationally efficient than existing methods, such as AIC and BIC, as the order of the model is determined in a single optimization. We conclude with applications of the method to two real data sets.

Résumé

Dans les applications des chaînes de Markov cachées (CMC), il se peut que les statisticiens n'aient pas l'information sur le nombre d'états (ou ordre) nécessaires pour représenter le processus. Le problème d'estimer le nombre d'états du CMC est ainsi une tâche d'importance majeure. Nous commençons avec une revue de littérature des développements majeurs dans le problème d'estimation de l'ordre d'un CMC. Nous proposons alors une nouvelle méthode de la quasi-vraisemblance pénalisée pour estimer l'ordre dans des CMC. Cette méthode utilise le fait que la distribution marginale des observations CMC est un mélange fini. La méthode débute avec un CMC avec un grand nombre d'états et obtient un modèle d'ordre inférieur en regroupant et fusionnant les états à l'aide de deux fonctions de pénalité. Nous étudions certaines propriétés asymptotiques de la méthode proposée et présentons une procédure numérique pour sa mise en œuvre. La performance est évaluée via des simulations extensives. La nouvelle méthode est plus efficace qu'autres méthodes, comme CIA et CIB, comme l'ordre du modèle est déterminé dans une seule optimisation. Nous concluons avec l'application de la méthode à deux vrais jeux de données.

Acknowledgements

I would like to thank my supervisor, Professor Abbas Khalili, for his guidance and support. I am also grateful to him for providing me with financial support to attend the 40th Annual Meeting of the Statistical Society of Canada in Guelph, Ontario as well as the 2012 Joint Statistical Meetings in San Diego, California.

I must also express my sincere gratitude to the professors at McGill for making my graduate studies such an enriching experience. Lastly, I would like to thank FQRNT and the Department of Mathematics and Statistics for their generous financial support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A hidden Markov model (HMM) is a doubly stochastic process, consisting of an observed process and an unobserved or "hidden" process that is used for modeling dependent data. It can be viewed as a generalization of a finite mixture model, where the hidden states are assumed to be Markov-dependent, rather than independent. More specifically, the observations are conditionally independent given the hidden states with the conditional distribution of the observation at time t depending only on the hidden state at this time. HMMs thus provide a method for dealing with unobserved sources of heterogeneity. However, unlike observations arising from a finite mixture model, HMM observations are dependent.

HMMs have been widely applied in fields such as engineering, biology, medicine and finance. For example, Churchill (1989) used HMMs to analyze DNA sequences due to their ability to capture the different patterns of base composition and dependence between adjacent bases on a DNA molecule. Levinson, Rabiner and Sondhi (1983) used HMMs for the purposes of prediction in speech recognition. Rydén, Teräsvirta and Åsbrink (1998) fit zero-mean normal state-dependent distributions to series of log-returns of daily values of the Standard & Poor's (S&P) 500 index.

HMMs are often used for the analysis of overdispersed series of count data. As we will see in Chapter 3, Leroux and Puterman (1992) fit a series of Poisson HMMs to a data set of movement counts by fetal lambs observed through ultrasound. The distribution of each observation is assumed to depend on the lamb's physiological state. Albert (1991) proposed a two-state Poisson HMM for a series of daily counts of epileptic seizures in one patient. He found that the model was able to adequately capture the apparent switching between states of high and low seizure frequency over time.

While in some applications, such as the modeling of epileptic seizure counts found in Albert (1991), the number of states (or order) of the HMM to be fitted is clear from

the background of the problem, our scientific knowledge may not always be enough to determine the number of hidden states. This is the case for the application to the lamb data presented in Leroux and Puterman (1992). The order estimation problem in HMMs is thus the focus of this thesis.

In Chapter 2, we formally define a HMM and highlight some of its key properties. We discuss model identifiability, maximum likelihood estimation via the expectation-maximization (EM) algorithm (Dempster et al., 1977) and direct numerical maximization as well as the asymptotic properties of the maximum likelihood estimators (MLEs).

In Chapter 3, we address the statistical task of estimating the number of states in a HMM. The order estimation methods that have been proposed thus far in the HMM context have utilized either the full-model likelihood or a so-called quasi-likelihood, which is based on the finite mixture marginal distributions. We first provide a literature review of these existing methods in Sections 3.1-3.3, including the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978), hypothesis testing-based methods as well as penalized minimum-distance approaches. We also discuss a Bayesian approach in Section 3.4. We then present our proposed method in Section 3.5, which is a penalized quasi-likelihood approach that is an extension of the modified smoothly clipped absolute deviation (MSCAD) method of Chen and Khalili (2008) for estimating the number of components in a finite mixture model. Some asymptotic properties of the penalized quasi-likelihood estimator of the order are discussed as well as the implementation of the proposed method and the selection of the tuning parameters used in the penalty functions. In Section 3.6, we evaluate the performance of the proposed method against AIC and BIC, based on both the full-model likelihood and the quasi-likelihood. The method is found to be an appealing alternative to the information criteria, especially when the true order of the model is high. We then demonstrate the use of the method with an analysis of two famous data sets.

We conclude with Chapter 4, where we summarize the work in this thesis and discuss future work relating to the proposed method. In particular, we highlight the need for the development of new procedures for choosing the tuning parameters used in the penalty functions. We also discuss possible extensions of the proposed method.

Chapter 2

Hidden Markov Models

In this chapter, we provide the mathematical definition of a HMM and highlight some of its key properties. The primary purpose of this chapter is to discuss statistical inference for HMMs when the number of hidden states is known.

2.1 Basic Definition

A hidden Markov model (HMM) is a doubly stochastic process (Y_t, Z_t), where (Y_t) is an observed process and (Z_t) is an unobserved process, which satisfy the following two properties.

- The unobserved process (Z_t) follows a Markov chain with discrete state space $\{1, 2, ..., K\}$, transition probabilities

$$p_{ij}^{(t)} = P(Z_t = j \mid Z_{t-1} = i) \text{ for } t = 2, ..., n \text{ and } i, j = 1, 2, ..., K,$$

  and vector of initial probabilities $\boldsymbol{\pi}^{(1)} = (\pi_1^{(1)}, \pi_2^{(1)}, ..., \pi_K^{(1)})$, where $\pi_k^{(1)} = P(Z_1 = k)$ for $k = 1, 2, ..., K$.

- Given (Z_t), the sequence of random variables (Y_t) are conditionally independent with the conditional distribution of Y_t depending only on Z_t.

For parametric HMMs, the conditional density function of $Y_t \mid Z_t = k$ is denoted by $f(y_t; \theta_k)$, belonging to some parametric family $\{f(\,\cdot\,; \theta) : \theta \in \Theta\}$ for $\Theta \subseteq \mathbb{R}^d$, $d \geq 1$. The conditional distributions of Y_t given Z_t are called state-dependent distributions. Note that while the results in this chapter are applicable to the case where the state-dependent parameters θ are multi-dimensional, we will restrict ourselves to the one-dimensional setting in Chapter 3.

We let $\mathbf{\Phi} = (\pi_1, \pi_2, ..., \pi_K, \mathbb{P}, \theta_1, \theta_2, ..., \theta_K)$ denote the vector of all parameters in the K-state HMM for $K \geq 1$, which belongs to the parameter space

$$\mathbf{\Omega} = \left\{ \mathbf{\Phi} : \sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0, \theta_k \in \Theta, \sum_{j=1}^{K} p_{ij} = 1, 0 \leq p_{ij} \leq 1, i, j, k = 1, 2, ..., K \right\}.$$

Note that for $K = 1$, (Y_t) are independent and identically distributed from $f(y; \theta)$.

In this chapter, we make the following three assumptions.

Assumption 1. The transition probabilities are homogenous; that is,

$$p_{ij} = P(Z_t = j \mid Z_{t-1} = i) = P(Z_2 = j \mid Z_1 = i)$$

for all $i, j = 1, 2, ..., K$.

Assumption 2. The unobserved Markov process (Z_t) is stationary.

Assumption 3. The number of states K is known and finite.

Remark. Assumption 1 is assumed throughout most of the HMM literature. Results on the asymptotic properties of non-homogeneous HMMs have yet to be established. Assumption 2 implies that the random variables (Y_t) are identically distributed, which, as pointed out by MacKay (2003), is a property that sometimes allows existing theory for independent and identically distributed random variables to be extended to the HMM context.

Under the stationarity assumption, for $k = 1, 2, ..., K$, $\pi_k = P(Z_t = k)$ for all $t = 1, 2, ..., n$ and the vector of π_k's is uniquely determined from the transition matrix $\mathbb{P}$ through the equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$. Furthermore, under the stationarity assumption, the marginal distribution of Y_t is a finite mixture with density given by

$$f(y_t; \mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k f(y_t; \theta_k),$$

where $\mathbf{\Psi} = (\pi_1, \pi_2, ..., \pi_K, \theta_1, \theta_2, ..., \theta_K)$. In contrast to the case of a finite mixture model, $Y_1, Y_2, ..., Y_n$ are marginally dependent. Thus, a HMM can be viewed as a generalization of a finite mixture model.

2.1.1 HMM Moments

Consider a K-state HMM (Y_t, Z_t), where Y_t is observed and Z_t is a stationary unobserved Markov chain with transition matrix $\mathbb{P}$ and stationary distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K)$,

and having (univariate) state-dependent distributions $f(y;\theta_k)$.

Let μ_k and σ_k^2 denote the mean and variance of the state-dependent distribution, $D = \text{diag}\,(\mu_1, \mu_2, ..., \mu_K)$, and $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_K)$.

Then

$$\begin{aligned}
\mathbb{E}(Y_t) &= \sum_{k=1}^{K} \mathbb{E}(Y_t \mid Z_t = k) P(Z_t = k) = \sum_{k=1}^{K} \pi_k \mu_k = \boldsymbol{\pi}\boldsymbol{\mu}^T, \\
\mathbb{E}(Y_t^2) &= \sum_{k=1}^{K} \mathbb{E}(Y_t^2 \mid Z_t = k) P(Z_t = k) = \sum_{k=1}^{K} \pi_k (\sigma_k^2 + \mu_k^2), \\
Var(Y_t) &= \sum_{k=1}^{K} \pi_k (\sigma_k^2 + \mu_k^2) - (\boldsymbol{\pi}\boldsymbol{\mu}^T)^2.
\end{aligned}$$

To obtain the autocorrelation function of a K-state HMM, we need to evaluate

$$\begin{aligned}
\mathbb{E}(Y_t Y_{t+h}) &= \sum_{k,l=1}^{K} \mathbb{E}(Y_t Y_{t+h} \mid Z_t = k, Z_{t+h} = l) P(Z_t = k, Z_{t+h} = l) \\
&= \sum_{k,l=1}^{K} \mathbb{E}(Y_t \mid Z_t = k) \mathbb{E}(Y_{t+h} \mid Z_{t+h} = l) p_{kl}^{(h)} \pi_k \\
&= \sum_{k,l=1}^{K} \pi_k \mu_k p_{kl}^{(h)} \mu_l \\
&= \sum_{k=1}^{K} \pi_k \mu_k \left(\sum_{l=1}^{K} p_{kl}^{(h)} \mu_l \right) \\
&= \boldsymbol{\pi} D \mathbb{P}^h \boldsymbol{\mu}^T
\end{aligned}$$

for positive integer h. Therefore, the autocorrelation function of a K-state HMM is given by

$$\begin{aligned}
\rho(h) &= Corr(Y_t, Y_{t+h}) \\
&= \frac{\mathbb{E}(Y_t Y_{t+h}) - \{\mathbb{E}(Y_t)\}^2}{Var(Y_t)} \\
&= \frac{\boldsymbol{\pi} D \mathbb{P}^h \boldsymbol{\mu}^T - (\boldsymbol{\pi}\boldsymbol{\mu}^T)^2}{\sum_{k=1}^{K} \pi_k (\sigma_k^2 + \mu_k^2) - (\boldsymbol{\pi}\boldsymbol{\mu}^T)^2}.
\end{aligned}$$

2.2 Identifiability

Before discussing parameter estimation in HMMs, we first address the issue of model identifiability. In general, a parametric model is said to be identifiable if different values of the parameter generate different probability distributions. As in Rydén (1995), we assume that the true order K_0 is minimal, that is, there does not exist a parameter under the model with $K < K_0$ states that induces the same probability distribution for (Y_t) as the parameter under the model with K_0 states. Model identifiability is important since parameters that are not identifiable cannot be consistently estimated. In what follows, we present the conditions that we assume to ensure the identifiability of a HMM.

Condition 1. The transition probability matrix of (Z_t) is ergodic (that is, irreducible and aperiodic).

Condition 2. The family of finite mixtures of $\{f(y;\theta) : \theta \in \Theta\}$ is identifiable, that is, equality of the density functions

$$\sum_{k=1}^{K} \pi_k f(y;\theta_k) = \sum_{k=1}^{K'} \pi'_k f(y;\theta'_k)$$

implies that $K = K'$, $\pi_k = \pi'_k$ and $\theta_k = \theta'_k$ for each $k = 1, 2, ..., K$ up to a permutation of the labels of the hidden states, where $\theta_k, \theta'_k \in \Theta$ and $0 < \pi_k, \pi'_k < 1$, with $\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K'} \pi'_k = 1$.

Condition 3. The parameters of the state-dependent distributions θ_k are distinct.

Remark. Condition 1 ensures the existence and uniqueness of the stationary distribution of the hidden process (Z_t), and that the true mixing proportions π_{0k} are positive for all $k = 1, 2, ..., K$. It also implies that the observed process (Y_t) is stationary and ergodic (Bickel et al., 1998 and references therein). Condition 2 is satisfied for various families of distributions, such as normal, Poisson, binomial and exponential mixtures. Condition 3 ensures that the distributions of the hidden states are distinct.

2.3 The Likelihood of a Hidden Markov Model

Unlike the case of a finite mixture model, the likelihood of a hidden Markov model is not simply the product of the marginal distributions of the observations. It does, however, factor neatly into a product of matrices. Given the number of states K, we will formulate

the likelihood for $\boldsymbol{\Phi}$ from the observed data $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. The observed data $\boldsymbol{y}$ will be referred to as the *incomplete data* and $(\boldsymbol{y}, \boldsymbol{z})$ the *complete data*, where $\boldsymbol{z} = (z_1, z_2, ..., z_n)$ are the hidden states. We present the likelihood in the case of discrete observations. The joint probability mass function of $(\boldsymbol{Y}, \boldsymbol{Z})$ is given by

$$\begin{aligned} P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z}) &= P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) P(\boldsymbol{Z} = \boldsymbol{z}) \\ &= P(Z_1 = z_1) \prod_{t=2}^{n} P(Z_t = z_t \mid Z_{t-1} = z_{t-1}) \prod_{t=1}^{n} P(Y_t = y_t \mid Z_t = z_t) \\ &= \pi_{z_1} \prod_{t=2}^{n} p_{z_{t-1}, z_t} \prod_{t=1}^{n} f(y_t; \theta_{z_t}). \end{aligned}$$

Thus, the likelihood function is

$$\begin{aligned} \mathcal{L}_n(\boldsymbol{\Phi}; \boldsymbol{y}) &= P(\boldsymbol{Y} = \boldsymbol{y}) \\ &= \sum_{z_1, ..., z_n = 1}^{K} P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z}) \\ &= \sum_{z_1, ..., z_n = 1}^{K} \pi_{z_1} f(y_1; \theta_{z_1}) p_{z_1, z_2} f(y_2; \theta_{z_2}) \cdots p_{z_{n-1}, z_n} f(y_n; \theta_{z_n}) \\ &= \sum_{z_1 = 1}^{K} \pi_{z_1} f(y_1; \theta_{z_1}) \sum_{z_2 = 1}^{K} p_{z_1, z_2} f(y_2; \theta_{z_2}) \cdots \sum_{z_n = 1}^{K} p_{z_{n-1}, z_n} f(y_n; \theta_{z_n}). \end{aligned}$$

The computation of the likelihood is on the order of nK^n since it involves the summation of K^n terms, each being a product of $2n$ factors. Now if we let $B(y_t)$ denote the $K \times K$ diagonal matrix with k^{th} diagonal element $f(y_t; \theta_k)$, then as shown in Zucchini and MacDonald (2009), the likelihood can be written as a product of matrices:

$$\mathcal{L}_n(\boldsymbol{\Phi}; \boldsymbol{y}) = \boldsymbol{\pi} B(y_1) \mathbb{P} B(y_2) \cdots \mathbb{P} B(y_n) \mathbf{1}', \tag{2.1}$$

where $\mathbf{1}$ is the K-dimensional row vector of ones.

2.3.1 Forward and Backward Probabilities

The evaluation of the likelihood can be made computationally less expensive through the use of the so-called forward and backward probabilities (Rabiner and Juang, 1986). In order to set up the likelihood computation in the form of an algorithm, Zucchini and MacDonald (2009) suggest defining the $1 \times K$ vector

$$\boldsymbol{\alpha}_t = \boldsymbol{\pi} B(y_1) \mathbb{P} B(y_2) \mathbb{P} B(y_3) \cdots \mathbb{P} B(y_t) = \boldsymbol{\pi} B(y_1) \prod_{s=2}^{t} \mathbb{P} B(y_s)$$

with the convention that the empty product is the identity matrix. The likelihood can thus be computed as $\mathcal{L}_n(\boldsymbol{\Phi}; \boldsymbol{y}) = \boldsymbol{\alpha}_n \mathbf{1}'$. To compute $\boldsymbol{\alpha}_n$, we will use the recursion

$$\boldsymbol{\alpha}_1 = \boldsymbol{\pi} B(y_1) \text{ and } \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} \mathbb{P} B(y_t) \text{ for } t = 2, ..., n.$$

In scalar form, we have for $j = 1, 2, ..., K$,

$$\alpha_1(j) = \pi_j f(y_1; \theta_j) \text{ and } \alpha_t(j) = \left\{ \sum_{i=1}^{K} \alpha_{t-1}(i)\, p_{ij} \right\} f(y_t; \theta_j) \text{ for } t = 2, ..., n.$$

In this form, the complexity of the likelihood computation is $O(nK^2)$, which is far more efficient than the $O(nK^n)$ complexity of the likelihood computation previously defined. For each $t = 1, 2, ..., n$, there are K elements of $\boldsymbol{\alpha}_t$ to be computed, each being a sum of K products.

In the following proposition, we will show that the components of $\boldsymbol{\alpha}_t$ are indeed probabilities, which are called forward probabilities.

Proposition 2.3.1. *For $t = 1, 2, ..., n$ and $j = 1, 2, ..., K$,*

$$\alpha_t(j) = P(Y_1 = y_1, ..., Y_t = y_t, Z_t = j).$$

Proof. See Appendix A.

We also define the vector of backward probabilities, $\boldsymbol{\beta}_t$, where for $i = 1, 2, ..., K$,

$$\beta_t(i) = \sum_{j=1}^{K} p_{ij} f(y_{t+1}; \theta_j) \beta_{t+1}(j) \text{ for } t = 1, ..., n-1 \text{ and } \beta_n(i) = 1.$$

The components of vector $\boldsymbol{\beta}_t$ can be computed with the matrices

$$\boldsymbol{\beta}_t' = \mathbb{P} B(y_{t+1}) \mathbb{P} B(y_{t+2}) \cdots \mathbb{P} B(y_n) \mathbf{1}' = \left\{ \prod_{s=t+1}^{n} \mathbb{P} B(y_s) \right\} \mathbf{1}'$$

for $t = 1, 2, ..., n$, where it follows that $\boldsymbol{\beta}_t' = \mathbb{P} B(y_{t+1}) \boldsymbol{\beta}_{t+1}'$ for $t = 1, 2, ..., n-1$ and $\boldsymbol{\beta}_n = \mathbf{1}$. We show that the backward probabilities are indeed probabilities in the following proposition.

Proposition 2.3.2. *For $t = 1, 2, ..., n-1$,*

$$\beta_t(i) = P(Y_{t+1} = y_{t+1}, ..., Y_n = y_n \mid Z_t = i).$$

Proof. See Appendix A.

While the likelihood may be computed from the forward probabilities alone, it may also be computed as

$$\mathcal{L}_n(\boldsymbol{\Phi}; \boldsymbol{y}) = \sum_{i=1}^{K} \alpha_t(i) \beta_t(i) = \sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_t(i) p_{ij} f(y_{t+1}; \theta_j) \beta_{t+1}(j),$$

which will be shown in Appendix A. Finally, we present two more results needed in the formulation of the EM algorithm in the next section.

Proposition 2.3.3. *For $t = 2, ..., n$ and $i, j = 1, 2, ..., K$,*

$$(1)\ P(Z_t = i \mid Y_1, ..., Y_n) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{K} \alpha_t(i)\beta_t(i)} \text{ and}$$

$$(2)\ P(Z_{t-1} = i, Z_t = j \mid Y_1, ..., Y_n) = \frac{\alpha_{t-1}(i)p_{ij}f(y_t;\theta_j)\beta_t(j)}{\sum_{i=1}^{K}\sum_{j=1}^{K} \alpha_{t-1}(i)p_{ij}f(y_t;\theta_j)\beta_t(j)}.$$

Proof. See Appendix A.

2.4 Maximum Likelihood Estimation

The most common approach for estimating the parameters in a HMM is to use maximum likelihood estimation. In general, suppose that we have a sample $y_1, y_2, ..., y_n$ drawn from some distribution with likelihood function $\mathcal{L}_n(\mathbf{\Phi}; \boldsymbol{y})$. The maximum likelihood estimator (MLE) of the parameter $\mathbf{\Phi}$ is defined as

$$\hat{\mathbf{\Phi}}_n = \arg\max_{\mathbf{\Phi} \in \mathbf{\Omega}} \mathcal{L}_n(\mathbf{\Phi}; \boldsymbol{y}).$$

In the case of a stationary HMM where the hidden process has a finite number of states, the strong consistency of the resulting MLE was proven by Leroux (1992a) and the asymptotic normality of the MLE was proven by Bickel et al. (1998) under mild conditions. Bickel et al. (1998) also showed that the observed information matrix converges in probability to the Fisher information matrix. In what follows, we present two methods for finding the maximum likelihood estimators of HMM parameters, namely the expectation-maximization (EM) algorithm (Dempster et al., 1977) and direct numerical maximization of the likelihood.

2.4.1 The EM Algorithm

We will now introduce the EM algorithm for the maximum likelihood fitting of a K-state hidden Markov model. In the HMM context, the EM algorithm is also known as the Baum-Welch algorithm (Baum et al., 1970). The EM algorithm is a two-step iterative procedure for estimating the MLE when dealing with latent variables. An EM iteration consists of an expectation step, known as the E-step, and a maximization step, known as the M-step.

To formulate the EM algorithm, we need the complete-data likelihood, given by

$$\mathcal{L}_n^C(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{z}) = P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z}) = \pi_{z_1} \prod_{t=2}^{n} p_{z_{t-1}, z_t} \prod_{t=1}^{n} f(y_t; \theta_{z_t}).$$

We also need to define the indicator variables

$$U_{tj} = \begin{cases} 1 & \text{if } Z_t = j \\ 0 & \text{otherwise} \end{cases}$$

for $t = 1, 2, ..., n$ as well as

$$V_{tij} = \begin{cases} 1 & \text{if } Z_{t-1} = i \text{ and } Z_t = j \\ 0 & \text{otherwise} \end{cases}$$

for $t = 2, ..., n$.

The complete-data likelihood then becomes

$$\mathcal{L}_n^C(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{z}) = \prod_{i=1}^{K} \pi_i^{u_{1i}} \prod_{t=2}^{n} \prod_{i=1}^{K} \prod_{j=1}^{K} p_{ij}^{v_{tij}} \prod_{t=1}^{n} \prod_{i=1}^{K} \{f(y_t; \theta_i)\}^{u_{ti}}$$

and thus the complete-data log-likelihood is given by

$$\ell_n^C(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{z}) = \sum_{i=1}^{K} u_{1i} \log \pi_i + \sum_{t=2}^{n} \sum_{i=1}^{K} \sum_{j=1}^{K} v_{tij} \log p_{ij} + \sum_{t=1}^{n} \sum_{i=1}^{K} u_{ti} \log f(y_t; \theta_i).$$

E-step: The E-step takes the conditional expectation of the complete-data log-likelihood $\ell_n^C(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{z})$ given the observed data $\boldsymbol{y}$ and the current parameter estimate $\boldsymbol{\Phi}^{(m)}$ to obtain

$$\begin{aligned} Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(m)}) &= \mathbb{E}\left[\ell_n^C(\boldsymbol{\Phi}; \boldsymbol{Y}, \boldsymbol{Z}) \mid \boldsymbol{Y}, \boldsymbol{\Phi}^{(m)}\right] \\ &= \sum_{i=1}^{K} \mathbb{E}\left[U_{1i} \mid \boldsymbol{Y}, \boldsymbol{\Phi}^{(m)}\right] \log \pi_i + \sum_{t=2}^{n} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{E}\left[V_{tij} \mid \boldsymbol{Y}, \boldsymbol{\Phi}^{(m)}\right] \log p_{ij} \\ &+ \sum_{t=1}^{n} \sum_{i=1}^{K} \mathbb{E}\left[U_{ti} \mid \boldsymbol{Y}, \boldsymbol{\Phi}^{(m)}\right] \log f(y_t; \theta_i). \end{aligned}$$

Now let us evaluate $\mathbb{E}\left[U_{ti} \mid \boldsymbol{Y}, \boldsymbol{\Phi}^{(m)}\right]$ and $\mathbb{E}\left[V_{tij} \mid \boldsymbol{Y}, \boldsymbol{\Phi}^{(m)}\right]$. We have

$$\hat{u}_{ti}^{(m)} = \mathbb{E}\left[U_{ti} \mid \boldsymbol{Y}, \boldsymbol{\Phi}^{(m)}\right] = P(Z_t = i \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\Phi}^{(m)}) = \frac{\alpha_t^{(m)}(i)\beta_t^{(m)}(i)}{\sum_{i=1}^{K} \alpha_t^{(m)}(i)\beta_t^{(m)}(i)}$$

and

$$\hat{v}_{tij}^{(m)} = \mathbb{E}\left[V_{tij} \mid \boldsymbol{Y}, \boldsymbol{\Phi}^{(m)}\right] = P(Z_{t-1} = i, Z_t = j \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\Phi}^{(m)})$$
$$= \frac{\alpha_{t-1}^{(m)}(i) p_{ij}^{(m)} f^{(m)}(y_t; \theta_j) \beta_t^{(m)}(j)}{\sum_{i=1}^{K}\sum_{j=1}^{K} \alpha_{t-1}^{(m)}(i) p_{ij}^{(m)} f^{(m)}(y_t; \theta_j) \beta_t^{(m)}(j)},$$

which were derived in Proposition 2.3.3.

Now the function $Q(\boldsymbol{\Phi};\, \boldsymbol{\Phi}^{(m)})$ becomes

$$Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(m)}) = \sum_{i=1}^{K} \hat{u}_{1i}^{(m)} \log \pi_i + \sum_{t=2}^{n}\sum_{i=1}^{K}\sum_{j=1}^{K} \hat{v}_{tij}^{(m)} \log p_{ij} + \sum_{t=1}^{n}\sum_{i=1}^{K} \hat{u}_{ti}^{(m)} \log f(y_t; \theta_i).$$

M-step: The M-step consists of the maximization of $Q(\boldsymbol{\Phi};\, \boldsymbol{\Phi}^{(m)})$ with respect to the parameter $\boldsymbol{\Phi}$, that is, it consists of finding

$$\boldsymbol{\Phi}^{(m+1)} = \arg\max_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(m)}).$$

This maximization neatly separates into three maximizations since the first summand of $Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(m)})$ depends only on the initial distribution $\boldsymbol{\pi}$, the second summand on the transition probabilities p_{ij}, and the last summand on the parameters of the state-dependent distributions θ_i.

Now we will perform the M-step to obtain the update equations for the initial distribution $\boldsymbol{\pi}$ and the transition probabilities p_{ij}. We have that for $m = 0, 1, 2, ...,$

$$\pi_i^{(m+1)} = \hat{u}_{1i}^{(m)} \text{ for } i = 1, 2, ..., K \text{ and}$$

$$p_{ij}^{(m+1)} = \frac{\sum_{t=2}^{n} \hat{v}_{tij}^{(m)}}{\sum_{j=1}^{K}\sum_{t=2}^{n} \hat{v}_{tij}^{(m)}} \text{ for } i, j = 1, 2, ..., K.$$

To obtain the update equations for the state-dependent parameters θ_i, we must solve

$$\frac{\partial Q(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(m)})}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left\{\sum_{t=1}^{n} \hat{u}_{ti}^{(m)} \log f(y_t; \theta_i)\right\} = 0$$

for $i = 1, 2, ..., K$.

Baum et al. (1970) show that the sequence of HMM parameter estimates $\{\mathbf{\Phi}^{(m)}\}$ have non-decreasing likelihood values, that is,

$$\mathcal{L}_n(\mathbf{\Phi}^{(m+1)}) \geq \mathcal{L}_n(\mathbf{\Phi}^{(m)}) \text{ for } m = 0, 1, 2, ...$$

Thus, since the sequence of likelihood values $\{\mathcal{L}_n(\mathbf{\Phi}^{(m)})\}$ is bounded above, it must converge. Starting from an initial value $\mathbf{\Phi}^{(0)}$, the E and M steps are iterated until

$$|\mathcal{L}_n(\mathbf{\Phi}^{(m+1)}) - \mathcal{L}_n(\mathbf{\Phi}^{(m)})| < \delta$$

for a pre-specified value $\delta > 0$. One may also use the convergence criterion

$$\|\mathbf{\Phi}^{(m+1)} - \mathbf{\Phi}^{(m)}\| < \epsilon$$

for a pre-specified value $\epsilon > 0$, where $\|\cdot\|$ is the Euclidean distance.

In the following two examples, we obtain the update equations for the parameters of K-state Poisson and normal HMMs.

Example 2.4.1. Now let us consider a Poisson HMM with means λ_i; that is, $f(y_t; \lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_t}}{y_t!}$. Then the update equation for λ_i is given by

$$\lambda_i^{(m+1)} = \frac{\sum_{t=1}^{n} y_t \hat{u}_{ti}^{(m)}}{\sum_{t=1}^{n} \hat{u}_{ti}^{(m)}} \text{ for } m = 0, 1, 2, ... \text{ and } i = 1, 2, ..., K.$$

Example 2.4.2. For a normal HMM with state-dependent parameters (μ_i, σ_i^2), the update equation for state mean μ_i is given by

$$\mu_i^{(m+1)} = \frac{\sum_{t=1}^{n} y_t \hat{u}_{ti}^{(m)}}{\sum_{t=1}^{n} \hat{u}_{ti}^{(m)}} \text{ for } i = 1, 2, ..., K$$

and the update equation for state variance σ_i^2 is

$$\sigma_i^{2\ (m+1)} = \frac{\sum_{t=1}^{n} (y_t - \mu_i^{(m)})^2 \hat{u}_{ti}^{(m)}}{\sum_{t=1}^{n} \hat{u}_{ti}^{(m)}} \text{ for } i = 1, 2, ..., K.$$

Note that in our presentation of the EM algorithm, no assumption of stationarity is made. The algorithm is used to estimate parameters of homogenous, but not necessarily

stationary, HMMs. When the additional constraint $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$ is imposed on the initial distribution $\boldsymbol{\pi}$ and the transition matrix $\mathbb{P}$, the M-step of the algorithm becomes more challenging.

2.4.2 The EM Algorithm for Stationary HMMs

Recall the objection function of the EM algorithm

$$Q(\boldsymbol{\Phi};\, \boldsymbol{\Phi}^{(m)}) = \sum_{j=1}^{K} \hat{u}_{1j}^{(m)} \log \pi_j + \sum_{t=2}^{n} \sum_{i=1}^{K} \sum_{j=1}^{K} \hat{v}_{tij}^{(m)} \log p_{ij} + \sum_{t=1}^{n} \sum_{j=1}^{K} \hat{u}_{tj}^{(m)} \log f(y_t; \theta_j).$$

Under the stationarity assumption, the first summand of $Q(\boldsymbol{\Phi};\, \boldsymbol{\Phi}^{(m)})$ also depends on the transition probabilities p_{ij} since $\pi_j = \sum_{i=1}^{K} \pi_i p_{ij}$ for $j = 1, 2, ..., K$. However, the maximization of the objective function subject to this constraint as well as the constraint $\sum_{j=1}^{K} p_{ij} = 1$ for $i = 1, 2, ..., K$ cannot be performed without resorting to numerical methods.

2.4.3 Direct Numerical Maximization

Another means of estimating the parameters of a HMM is through direct numerical maximization of the likelihood. To implement this method, the parameters of the Markov chain must first be transformed in order to deal with the parameter constraints. For any HMM, the rows of the transition matrix $\mathbb{P}$ must sum to 1 and all the parameters p_{ij} must be non-negative. To meet these constraints, we will reparametrize the transition matrix $\mathbb{P}$ by defining the matrix $\Gamma = \{\gamma_{ij}\}$ and then setting

$$p_{ij} = \begin{cases} \dfrac{1}{1 + \sum_{l=2}^{K} \exp(\gamma_{il})} & \text{if } i = j \\ \dfrac{\exp(\gamma_{ij})}{1 + \sum_{l=2}^{K} \exp(\gamma_{il})} & \text{if } i \neq j \end{cases}$$

for $i, j = 1, 2, ..., K$. If constraints are also imposed on the state-dependent parameters θ_k, we will let $\xi_k = g(\theta_k)$ denote the transformed θ_k for some one-to-one mapping $g : \Theta \to \mathbb{R}$. Thus, to estimate the constrained parameters $\mathbb{P}$ and $\boldsymbol{\theta}$, we first maximize the likelihood with respect to the unconstrained parameters Γ and $\boldsymbol{\xi}$. Then once the estimates of Γ and $\boldsymbol{\xi}$ are obtained, we transform them to estimates of $\mathbb{P}$ and $\boldsymbol{\theta}$.

When maximizing the likelihood numerically in order to estimate parameters, we run into the problem of numerical underflow. Recall from Section 2.3 that the likelihood

computation requires the calculation of the forward probabilities $\alpha_t(j)$ for $j = 1, 2, ..., K$. Since each $\alpha_t(j)$ is a product of the probabilities p_{ij}, $\alpha_{t-1}(i)$, and $f(y_t; \theta_j)$, it becomes progressively smaller as t increases and is eventually rounded to zero. In order to avoid this problem, Zucchini and MacDonald (2009) suggest a scaling of the likelihood. Since the EM algorithm also requires the computation of the forward and backward probabilities, it is not immune to numerical underflow and a scaling of the computation of the logarithms of these probabilities should also be used.

2.4.4 Estimation of the Initial Distribution

When no assumption of stationarity of the hidden Markov process is made, Cappé et al. (2005) point out that the initial distribution $\boldsymbol{\pi}$ cannot be estimated consistently since "there is only one random variable...(that is not even observed!) drawn from [the distribution]". In fact, as stated in Levinson, Rabiner and Sondhi (1983), at a maximum of the likelihood, the sequence of estimates $\{\boldsymbol{\pi}^{(m)}\}$ will approach one of the K unit vectors. Therefore, if the constraint $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$ is not placed on $\boldsymbol{\pi}$, the maximum likelihood estimate of the entire vector of parameters $\boldsymbol{\Phi} = (\boldsymbol{\pi}, \mathbb{P}, \theta)$ will be inconsistent. Nevertheless, as noted by Leroux (1992a), the consistency of the MLE of $(\mathbb{P}, \theta)$ does not depend on the initial distribution.

One approach for handling the initial distribution $\boldsymbol{\pi}$ would be to assume that it is known. However, this approach may not be reasonable if no prior information of the population being sampled is available.

A more widely used approach throughout the HMM literature would be to assume stationarity of the hidden Markov process. In this case, since the initial distribution $\boldsymbol{\pi}$ is dependent on the transition matrix $\mathbb{P}$ through the equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$, $\boldsymbol{\pi}$ is no longer a parameter to be estimated, but is rather a solution to a system of K linear equations. At each iteration of the maximization procedure, the updated initial distribution $\boldsymbol{\pi}^{(m+1)}$ can be found by solving the equation $\boldsymbol{\pi}^{(m+1)}(I_K - \mathbb{P}^{(m+1)} + O) = \mathbf{1}$, which is shown to be equivalent to the stationarity condition $\boldsymbol{\pi}^{(m+1)} = \boldsymbol{\pi}^{(m+1)}\mathbb{P}^{(m+1)}$ in Appendix A, and where I_K is the $K \times K$ identity matrix, $\mathbb{P}^{(m+1)}$ is the updated transition matrix, O is the $K \times K$ matrix of ones and $\mathbf{1}$ is the K-dimensional row vector of ones.

Therefore, when it comes to fitting a HMM to a given data set, assuming stationarity of the hidden Markov process provides the analyst the opportunity to obtain meaningful and interpretable results for the initial distribution. For illustrative purposes, we provide two examples, demonstrating the convergence of the sequence of estimates $\{\boldsymbol{\pi}^{(m)}\}$ to one of the K unit vectors when using the EM algorithm without the stationarity assumption.

We also provide the parameter estimates obtained using direct numerical maximization of the likelihood with the stationarity assumption.

Example 2.4.3. We consider a sample of 200 observations generated from the 2-state Poisson HMM fitted to the series of annual counts of major earthquakes from 1900 to 2006 provided in Zucchini and MacDonald (2009). The model is $\boldsymbol{\lambda} = (15, 26)$, $\boldsymbol{\pi} = (0.661, 0.339)$, with transition matrix

$$\mathbb{P} = \begin{pmatrix} 0.934 & 0.066 \\ 0.129 & 0.871 \end{pmatrix}.$$

With starting values $\boldsymbol{\lambda}^{(0)} = (12, 20)$, $\boldsymbol{\pi}^{(0)} = (0.50, 0.50)$ and

$$\mathbb{P}^{(0)} = \begin{pmatrix} 0.50 & 0.50 \\ 0.50 & 0.50 \end{pmatrix},$$

we obtain the results displayed in Table 2.1.

Parameter	Non-Stationary Model - EM Algorithm	Stationary Model - Direct Numerical Maximization
(λ_1, λ_2)	(14.497, 26.290)	(14.498, 26.305)
(π_1, π_2)	(0, 1)	(0.675, 0.325)
p_{12}	0.057	0.062
p_{21}	0.135	0.130
Negative Log-likelihood	603.788	604.727
No. of Iterations Until Convergence	19	22

Table 2.1: Results of the EM algorithm and direct numerical maximization, assuming stationarity, applied to a sample of 200 observations, generated from a 2-state Poisson HMM.

Example 2.4.4. Now we consider a sample of 200 observations generated from the 3-state Poisson HMM also fitted to the earthquake data provided in Zucchini and MacDonald (2009). The model is $\boldsymbol{\lambda} = (13, 19, 29)$, $\boldsymbol{\pi} = (0.3254, 0.4890, 0.1856)$, with transition matrix

$$\mathbb{P} = \begin{pmatrix} 0.9393 & 0.0321 & 0.0286 \\ 0.0404 & 0.9064 & 0.0532 \\ 0 & 0.1903 & 0.8097 \end{pmatrix}.$$

With starting values $\boldsymbol{\lambda}^{(0)} = (10, 15, 25)$, $\boldsymbol{\pi}^{(0)} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and

$$\mathbb{P}^{(0)} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix},$$

we obtain the results displayed in Table 2.2.

Parameter	Non-Stationary Model - EM Algorithm	Stationary Model - Direct Numerical Maximization
$(\lambda_1, \lambda_2, \lambda_3)$	(12.994, 18.486, 29.667)	(13.037, 18.508, 29.676)
(π_1, π_2, π_3)	(1, 0, 0)	(0.389, 0.466, 0.145)
p_{12}	0.079	0.069
p_{13}	0	0
p_{21}	0.051	0.058
p_{23}	0.039	0.038
p_{31}	0	0
p_{32}	0.118	0.122
Negative Log-likelihood	596.121	597.112
No. of Iterations Until Convergence	44	78

Table 2.2: Results of the EM algorithm and direct numerical maximization, assuming stationarity, applied to a sample of 200 observations, generated from a 3-state Poisson HMM.

2.4.5 A Comparison of the EM Algorithm and Direct Numerical Maximization

When hidden Markov models were first introduced, the EM algorithm was largely preferred by researchers and analysts for performing maximum likelihood estimation. There are likely three main reasons for its popularity. Firstly, for homogeneous HMMs, at each iteration of the EM algorithm, there are simple, closed-form expressions for the update equations of the parameter estimates. The code for the EM algorithm can also be easily altered for different parametric families of state-dependent distributions. Secondly, the EM does not require the supply of derivatives of HMM likelihoods, which are often difficult to compute (MacKay, 2003). Lastly, the EM algorithm deals with parameter constraints implicitly.

However, the EM algorithm is slow to converge. Furthermore, it does not readily produce standard errors of parameter estimates (Cappé et al., 2005). As a result, direct numerical maximization has recently been favoured for performing maximum likelihood estimation of HMMs since it is typically much more efficient (MacKay, 2003 and references therein).

In this case, parameter constraints must be dealt with explicitly either through transformations or using constrained optimization procedures. Built-in optimizers from the statistical software `R`, such as `nlm`, `optim` or `constrOptim`, do not even require the specification of derivatives. Once code for the transformation of parameters and computation of the likelihood is written, very little programming effort is required to code the maximization of the likelihood. Furthermore, the EM algorithm requires the computation of both the forward and backward probabilities for the update equations of the parameter estimates, while direct numerical maximization need only the forward probabilities for the computation of the likelihood.

In Altman and Petkau (2005), direct maximization of the likelihood based on a quasi-Newton routine applied to a data set of lesion counts in multiple sclerosis patients performed much faster than the EM algorithm in finding the MLEs. Turner (2008) also found that direct maximization by the Levenberg-Marquardt algorithm converged in fewer steps than the EM for two examples. Furthermore, the Levenberg-Marquardt algorithm provides an estimate of the covariance matrix of the parameter estimates and thus estimates of their precision.

Starting values of both iterative procedures are important since HMM likelihoods tend to have multiple local maxima. Both procedures may converge to local maxima, saddle points or the global maximum depending on the starting value $\mathbf{\Phi}^{(0)}$. One way of dealing with this dependence on the starting value is to run the numerical maximization procedure over a range of starting values in the parameter space $\mathbf{\Omega}$ and selecting the estimated parameter that has the highest likelihood. Direct numerical maximization is more sensitive to starting values than the EM. However, performing a grid search over a variety of possible starting values will increase the chances of identifying the global maximum.

2.5 Summary

In this chapter, we discussed key features of HMMs and compared two methods for performing maximum likelihood estimation in the HMM context, namely the EM algorithm (or Baum-Welch algorithm), and direct numerical maximization of the likelihood. While the EM algorithm has been traditionally applied in the HMM literature due to its relative simplicity, direct numerical maximization of the likelihood is also an appealing option for finding the MLEs of HMM parameters. In fact, for the case of a stationary HMM when the estimation of the initial distribution is of particular interest, direct maximization of the likelihood should be preferred since the optimization of the objective function in the M-step of the EM algorithm cannot be performed without resorting to numerical methods.

Chapter 3

Order Estimation in Hidden Markov Models

The methods discussed in Chapter 2 for estimating the parameters of a hidden Markov model assume that the number of states (or order) is known. However, when fitting a HMM to a given data set, a researcher or practitioner may have no knowledge of the number of states needed in the model in order to be able to adequately describe the data. For example, in Section 3.7 we consider the problem of estimating the number of physiological states of a fetal lamb, which is important when it comes to modeling its breathing and body movements during gestation. While more complex models are likely to provide more adequate fits to the data, they are not often favoured in applications since a large number of parameters can lead to high variance in parameter estimates and overfitting. Moreover, simple models are preferable over complex models in terms of model interpretability. Thus, order estimation is a task of major importance.

Several statistical methods have been proposed to estimate the number of states in a HMM. Since the likelihood increases as the number of states increases, maximum likelihood estimation is not a reasonable approach for this estimation problem. Information-theoretic approaches, such as the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978), are often used. Another method for estimating the number of states is by minimizing some distance measure between the empirical distribution function of the observed data and the fitted cumulative distribution function. One may also consider this problem from the point of view of hypothesis-testing on the order of a HMM.

While all of the methods are advantageous in certain aspects, it is important to note that none of the methods are optimal. Some methods might have computational advantages, while others might have a higher probability of selecting the most suitable model.

In this chapter, we propose a new order estimation procedure that extends to HMMs the MSCAD method of Chen and Khalili (2008) for estimating the number of components in a finite mixture. Starting with a HMM with a large number of states, the method seeks to cluster and then merge similar states of the model through two penalty functions. Our simulation results reveal that the proposed method is a viable alternative to existing order estimation procedures in the HMM context. The advantage of this method is that it does not involve the comparison of all the candidate models. With the selection of an appropriate tuning parameter, the method is able to obtain the order in a single optimization.

In what follows, we first review some of the existing methods in the HMM literature for order estimation.

3.1 Information-Based Methods

The two main information-theoretic approaches for estimating the true number of states are Akaike's information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978). AIC and BIC are both based on the penalization of the likelihood according to model complexity. While they differ only in the penalty term, their motivations have different origins. AIC selects the model which minimizes an estimate of the Kullback-Leibler distance between the true distribution and the distribution of the candidate models. BIC, on the other hand, was derived within a Bayesian framework. It selects the model which is *a posteriori* most probable among all candidate models.

For independent mixture distributions, the theoretical justification for the use of these information criteria is provided in Leroux (1992b), in which he proved that under mild regularity conditions the estimated number of components selected using AIC and BIC is at least as large as the true number K_0, asymptotically. This was shown for both finite and infinite mixtures. Keribin (2000) proved that under certain regularity conditions BIC provides a consistent estimator of the true number of components.

We will consider these information criteria in the HMM context. Recall from Section 2.3 that the likelihood function for a k-state HMM is given by

$$\mathcal{L}_n(\boldsymbol{\Phi}_k) = \sum_{z_1=1}^{k} \pi_{z_1} f(y_1;\theta_{z_1}) \sum_{z_2=1}^{k} p_{z_1,z_2} f(y_2;\theta_{z_2}) \cdots \sum_{z_n=1}^{k} p_{z_{n-1},z_n} f(y_n;\theta_{z_n}),$$

where $\boldsymbol{\Phi}_k = (\pi_1, ..., \pi_k, \mathbb{P}, \theta_1, ..., \theta_k) \in \boldsymbol{\Omega}_k$ and $\boldsymbol{\Omega}_k$ denotes the parameter space of a k-state HMM. The log-likelihood function is thus $\ell_n(\boldsymbol{\Phi}_k) = \log \mathcal{L}_n(\boldsymbol{\Phi}_k)$.

Assuming an upper bound K on the true number of states K_0, AIC selects the value of k that minimizes the criterion

$$AIC(k) = -2\,\ell_n(\hat{\mathbf{\Phi}}_k) + 2d_k, \tag{3.1}$$

over all $k = 1, 2, ..., K$, whereas BIC selects the value of k that minimizes the criterion

$$BIC(k) = -2\,\ell_n(\hat{\mathbf{\Phi}}_k) + d_k \log n, \tag{3.2}$$

over all $k = 1, 2, ..., K$, where d_k is the number of unknown parameters for the HMM of order k, n is the sample size and $\hat{\mathbf{\Phi}}_k$ is the maximum likelihood estimator of $\mathbf{\Phi}_k$ for the HMM of order k.

For both AIC and BIC, the greater the number of states, the more heavily the log-likelihood is penalized. In this way, these information criteria attempt to control the estimated number of states directly. In Equation (3.2), we see that the penalty term for BIC depends not only on the number of states k, but also on the sample size n. If $n \geq 8$, which holds in most situations, then $d_k \log(n) > 2d_k$ so that BIC will penalize complex models more heavily than AIC. Thus, AIC has the potential of overfitting. Note that for a k-state Poisson HMM, $d_k = k^2 + k - 1$ and for a k-state normal HMM, $d_k = k^2 + 2k - 1$. In the stationary case, these models will have $k - 1$ fewer parameters to be estimated since the initial distribution $\boldsymbol{\pi}$ can be determined from the transition matrix $\mathbb{P}$.

While theoretical justification for the use of AIC and BIC has been provided by Leroux (1998b) for independent mixtures, it has yet to be provided for HMMs (MacKay, 2003). Nevertheless, AIC and BIC have been considered by most authors applying HMMs, including Leroux and Puterman (1992), Rydén (1995), and Zucchini and MacDonald (2009). Rydén (1995) shows that a class of penalized likelihood estimators, including AIC and BIC, in the limit never underestimate the true order K_0.

Poskitt and Zhang (2005) use AIC and BIC based on the quasi-log-likelihood

$$\ell_n^Q(\mathbf{\Psi}_k) = \sum_{t=1}^{n} \log f(y_t; \mathbf{\Psi}_k) = \sum_{t=1}^{n} \log \left\{ \sum_{i=1}^{k} \pi_i f(y_t; \theta_i) \right\}$$

as methods for selecting the order of a HMM. In other words, they reduce the problem to selecting the number of components in the marginal mixture distribution of the observed process (Y_t) by replacing the maximized log-likelihood $\ell_n(\hat{\mathbf{\Phi}}_k)$ in Equations (3.1) and (3.2) by the maximized quasi-log-likelihood $\ell_n^Q(\hat{\mathbf{\Psi}}_k)$, where $\hat{\mathbf{\Psi}}_k$ is the maximum quasi-likelihood

estimator (MQLE) of $\mathbf{\Psi}_k$. They show that BIC based on quasi-likelihood inference provides a consistent estimator of K_0.

3.2 Hypothesis Testing-Based Methods

Tests of hypothesis on the order K of a HMM may also be viewed as order selection procedures. For testing $K = 1$ against $K \geq 2$, Gassiat and Keribin (2000) show that the likelihood ratio test (LRT) statistic tends to ∞ in probability as the number of observations increases. Note that a HMM with $K = 2$ states is the simplest nontrivial HMM since for $K = 1$, (Y_t) are independent and identically distributed from $f(y; \theta)$. Therefore, the test of $K = 2$ against $K \geq 3$ states is the basic testing problem for HMMs.

Since the marginal distribution of the observations of a stationary HMM is a finite mixture, the tests proposed in the literature thus far estimate the number of states in a HMM by determining the number of components in the marginal mixture distribution. Dannemann and Holzmann (2008) proposed testing the hypothesis $K = 2$ by extending to HMMs the modified likelihood ratio (MLR) test of Chen, Chen and Kalbfleisch (2004) for testing two states in a finite mixture.

The MLR test of Chen, Chen and Kalbfleisch (2004) had been proposed in order to overcome the complications of the asymptotic null distribution of the LR statistic in finite mixture models. A classic result of Wilks (1938) states that if standard regularity conditions hold, minus twice the logarithm of the LR statistic is asymptotically chi-squared distributed under the null hypothesis. However, in the finite mixture setting, the null hypothesis lies on the boundary of the parameter space rather than in the interior and the null distribution is not identifiable. Therefore, Wilks' result is not applicable since the standard regularity conditions are not satisfied.

In what follows, we outline the modified LRT for $K = 2$ against $K \geq 3$ states in a HMM, proposed by Dannemann and Holzmann (2008). First we introduce some notation. We write the marginal distribution of the HMM observations (Y_t) as

$$f(y_t; G) = \sum_{k=1}^{K} \pi_k f(y_t; \theta_k),$$

where $G(\theta)$, called the *mixing distribution*, is a discrete cumulative distribution function with a finite number of support points $\theta_1, ..., \theta_K \in \Theta$ and corresponding weights $\pi_1, ..., \pi_K$ that satisfy $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. Now let

$$\mathcal{M}_K = \left\{ G(\theta) = \sum_{k=1}^{K} \pi_k I(\theta_k \leq \theta) : \theta_1 \leq ... \leq \theta_K, \sum_{k=1}^{K} \pi_k = 1, \pi_k > 0 \right\}$$

denote the set of all finite mixing distributions with at most K support points, where $I(\cdot)$ is an indicator function. Also let $\mathcal{M} = \cup_{K \geq 2} \mathcal{M}_K$. Therefore, letting G_0 denote the true mixing distribution of the marginal distribution, Dannemann and Holzmann (2008) proposed a test for

$$H_0 : G_0 \in \mathcal{M}_2 \text{ against } H_1 : G_0 \in \mathcal{M} \backslash \mathcal{M}_2$$

Under H_0, the true mixing distribution of the marginal distribution is then

$$G_0(\theta) = \pi_{01} I(\theta_{01} \leq \theta) + (1 - \pi_{01}) I(\theta_{02} \leq \theta),$$

where $\theta_{01} < \theta_{02}$ are distinct points of the interior of Θ and $0 < \pi_{01} < 1$.

Dannemann and Holzmann (2008) then define the modified quasi-log-likelihood function as

$$\tilde{\ell}_n^{Q(K)}(G) = \sum_{t=1}^{n} \log f(y_t; G) + C_K \sum_{k=1}^{K} \log \pi_k,$$

where C_K is a positive constant. The penalty on the mixing proportions π_k prevents the estimates of π_k from being too close to 0. The amount of penalty is determined by the constant C_K, which is chosen to reflect the size of Θ. See Chen, Chen and Kalbfleisch (2004) and references therein for further details on the choice of C_K. As we will see in Section 3.5, the penalty on the mixing proportions also plays an important role in our new method for order selection. There we discuss the role of this penalty in our proposed method and provide a more in-depth discussion of the choice of C_K.

The test statistic for the modified quasi-likelihood ratio test for two components is given by

$$T_n^{mod} = 2 \left\{ \ell_n^{Q(K)}(\hat{G}_K) - \ell_n^{Q(2)}(\hat{G}_2) \right\},$$

where $\hat{G}_K$ results from the maximization of $\tilde{\ell}_n^{Q(K)}(\cdot)$ and is called a modified maximum quasi-likelihood estimate and $\ell_n^{Q(K)}(\cdot)$ is the unpenalized quasi-log-likelihood function.

Dannemann and Holzmann (2008) show that the modified LRT statistic T_n^{mod} has the same limit distribution as for independent mixtures. In particular, the asymptotic distribution of T_n^{mod} is that of the mixture

$$(\tfrac{1}{2} - \rho)\chi_0^2 + \tfrac{1}{2}\chi_1^2 + \rho\chi_2^2,$$

where ρ is specified in Dannemann and Holzmann (2008) and χ^2_0 is the distribution with unit mass at 0.

Dannemann and Holzmann (2008) investigated the finite-sample performance of the modified LRT for $K = 2$ against $K \geq 3$ for HMMs with state-dependent Poisson distributions as well as with state-dependent normal distributions. As expected from their asymptotic theory, the finite-sample performance is hardly affected when the transition matrices are altered and the stationary distribution is kept the same, provided the diagonal entries of the transition matrices are not too close to 0 or 1. Furthermore, their simulation studies show that one should only expect a slight loss of power when introducing dependence through the transition matrix $\mathbb{P}$.

3.3 Penalized Minimum-Distance Approaches

MacKay (2002) extends the penalized minimum-distance method of Chen and Kalbfleisch (1996) for estimating the number of hidden states in a HMM. She estimates K_0 by minimizing the penalized distance function

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

over all F, where F is a finite mixing distribution with K components, $\bar{F}_n$ is the empirical distribution function of Y_t, (c_n) is a sequence of positive constants, and d is the Kolmogorov-Smirnov distance:

$$d(F_1, F_2) = \sup_y |F_1(y) - F_2(y)|$$

for distribution functions F_1 and F_2. The method incorporates a penalty term which penalizes models with states that have small values of π_k. In this way, the method is indirectly controlling the number of states K.

As pointed out by MacKay (2002), locating the global minimum of the penalized distance function is often challenging since the objective function has many local minima. Running the algorithm for a wide variety of initial values may help in locating the global minimum, but at the expense of greater computational effort. Furthermore, with $c_n = Cn^{-1/2} \log n$ for some $C > 0$, the choice of C was found to have a considerable effect on the estimate of the true number of states K_0 in the simulations of MacKay (2002).

The performance of this method may improve if different distance functions are used. For example, the Cramér-von Mises distance

$$d(F_1, F_2) = \int \{F_1(y) - F_2(y)\}^2 \, d\{F_1(y) + F_2(y)\}$$

may also be considered.

3.4 A Bayesian Approach

A Bayesian approach for estimating the number of hidden states of a HMM compares marginal likelihoods. To decide between two competing models $K = k_1$ and $K = k_2$, one computes the ratio of the model probabilities:

$$\frac{p(K = k_2|\boldsymbol{y})}{p(K = k_1|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|K = k_2)}{p(\boldsymbol{y}|K = k_1)} \frac{p(K = k_2)}{p(K = k_1)},$$

where p can be a probability mass function or density function and $p(\boldsymbol{y}|K)$ is the marginal likelihood, which is also called the integrated likelihood. It is given by

$$p(\boldsymbol{y}|K) = \int p(\boldsymbol{\Phi}_K, \boldsymbol{y}|K) \, \mathrm{d}\boldsymbol{\Phi}_K = \int p(\boldsymbol{y}|K, \boldsymbol{\Phi}_K) p(\boldsymbol{\Phi}_K|K) \, \mathrm{d}\boldsymbol{\Phi}_K,$$

where $\boldsymbol{\Phi}_K$ is the parameter of the K-state HMM.

The term $\dfrac{p(\boldsymbol{y}|K = k_2)}{p(\boldsymbol{y}|K = k_1)}$ is called the Bayes factor, which represents the factor of increase in posterior probability for the model $K = k_2$ over that of model $K = k_1$. Therefore, to decide between J competing models $k_1, k_2, ..., k_J$, one selects the model with the highest marginal likelihood.

The marginal likelihood, however, is generally difficult to compute. Methods that use Markov Chain Monte Carlo (MCMC) output, such as the harmonic mean estimator (Zucchini and MacDonald, 2009 and references therein) or Chib's approach (1995), have been considered. For the lamb data application from Leroux and Puterman (1992), Frühwirth-Schnatter (2006) uses bridge sampling and importance sampling to calculate marginal likelihoods and estimate the number of hidden states.

3.5 A New Order Estimation Method

We now propose a new method for order estimation in hidden Markov models, which makes use of the fact that under the assumption of stationarity of the hidden Markov chain, the marginal distribution of the HMM observations is a finite mixture. The new method is an extension to the HMM context of the modified smoothly clipped absolute

deviation (MSCAD) procedure of Chen and Khalili (2008) for estimating the number of components in a finite mixture model. Their procedure is based on the maximization of a penalized likelihood, which incorporates the penalty function from the penalized minimum-distance method of Chen and Kalbfleisch (1996) as well as the smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001).

Recall that the marginal distribution of the HMM observations (Y_t) is given by

$$f(y_t; \mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k f(y_t; \theta_k),$$

where $\theta_k \in \Theta$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K)$ is the vector of stationary mixing proportions. We will assume that $\theta_1 \leq \theta_2 \leq ... \leq \theta_K$ and set $\eta_k = \theta_{k+1} - \theta_k$ for $k = 1, 2, ..., K-1$.

We let K_0 denote the true number of states. Our aim is to estimate K_0 using the following quasi-likelihood function

$$\mathcal{L}_n^Q(\mathbf{\Psi}) = \prod_{t=1}^{n} f(y_t; \mathbf{\Psi}),$$

or, equivalently, the quasi-log-likelihood function

$$\ell_n^Q(\mathbf{\Psi}) = \sum_{t=1}^{n} \log f(y_t; \mathbf{\Psi})$$

based on the finite mixture marginal distributions.

Inference in HMMs based on the marginal mixture distribution is not uncommon. Under mild regularity conditions, Lindgren (1978) showed that the maximum quasi-likelihood estimator is consistent and asymptotically normal. To initialize iterative procedures for obtaining the exact maximum likelihood estimates of a HMM, Leroux and Puterman (1992) suggest using the parameter estimates obtained from fitting a finite mixture model. Furthermore, as we saw in Sections 3.1 and 3.2, the marginal mixture distribution has not only been considered for estimation of HMM parameters, but also for estimation of the order. Dannemann and Holzmann (2008) proposed to do hypothesis-testing on the order via a modified quasi-likelihood ratio, and Poskitt and Zhang (2005) considered AIC and BIC based on the quasi-likehood.

One of the advantages of using the quasi-likelihood function is that it simplifies the optimization problem since the objective function does not involve the transition probabilities p_{ij}. Furthermore, there are computational gains that arise from using the quasi-likelihood. As we saw in Section 2.3, the complexity of the efficient evaluation of the likelihood is

$O(nK^2)$, whereas that of the quasi-likelihood is $O(nK)$.

By maximizing the quasi-log-likelihood, the resulting fitted model may overfit the data with two types of overfitting. For the first type of overfitting, the estimated values of π_k may be close to 0 and for the second type of overfitting, the state-dependent densities may be close to one another. To prevent these two types of overfitting, we maximize an objective function that incorporates a penalty on the mixing proportions π_k as well as a penalty on the differences in atoms η_k of the mixing distribution of the marginal distribution.

Our proposed method estimates the number of states by maximizing the penalized quasi-log-likelihood function

$$\tilde{\ell}_n^Q(\mathbf{\Psi}) = \ell_n^Q(\mathbf{\Psi}) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_n(\eta_k)$$

for some $K > K_0$. Here we assume that some information is available on an upper bound K on the true number of states. The constant $C_K > 0$ is chosen as in Chen, Chen and Kalbfleisch (2004). We provide a discussion of the selection of C_K in Section 3.5.3.

The first penalty was used in the finite mixture setting by a few authors, including Chen and Kalbfleisch (1996) in their penalized minimum-distance method for estimating the number of components in a finite mixture as well as Chen, Chen and Kalbfleisch (2004) in their modified likelihood ratio test. It forces the estimated values of π_k away from the boundary point 0 to prevent the first type of overfitting.

The second penalty is the SCAD penalty, developed by Fan and Li (2001), in the context of variable selection in regression. It is defined through its derivative

$$p_n'(\eta) = \gamma_n \sqrt{n} I\{\sqrt{n}|\eta| \le \gamma_n\} + \frac{\sqrt{n}(a\gamma_n - \sqrt{n}|\eta|)_+}{(a-1)} I(\sqrt{n}|\eta| > \gamma_n)$$

for some $a > 2$, where $(\cdot)_+$ denotes the positive part of a quantity. Figure 3.1 shows a plot of the SCAD penalty with $n = 100$ and $\gamma_n = n^{1/4} \log n$.

Note that

$$p_n(\eta) = \begin{cases} \gamma_n \sqrt{n}|\eta| & \text{if } \sqrt{n}|\eta| \le \gamma_n \\ \dfrac{2\gamma_n \sqrt{n}|\eta|(a-1) - (\sqrt{n}|\eta| - \gamma_n)^2}{2(a-1)} & \text{if } \gamma_n < \sqrt{n}|\eta| \le a\gamma_n \\ \dfrac{\gamma_n^2(a+1)}{2} & \text{if } \sqrt{n}|\eta| > a\gamma_n \end{cases}$$

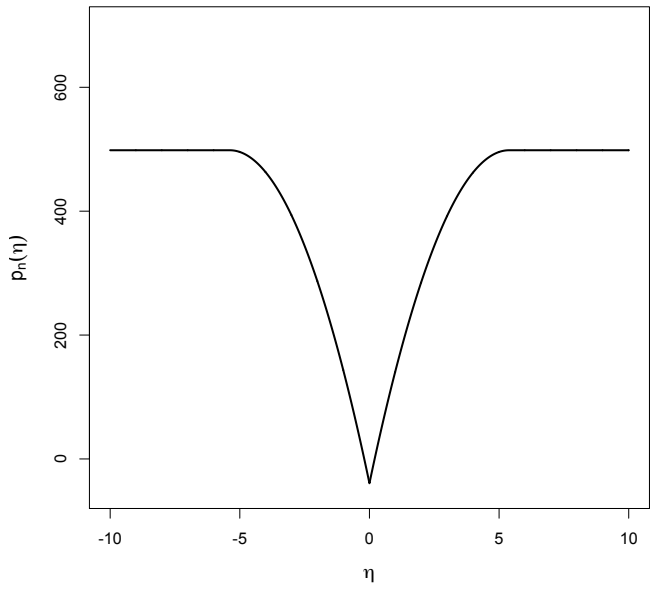

Figure 3.1: Plot of the SCAD penalty function.

so that the SCAD penalty function $p_n(\eta)$ is constant when $|\eta| > an^{-1/2}\gamma_n$. The SCAD penalty is chosen since it shrinks values of η that are close to 0 to exactly 0 with positive probability, preventing the second type of overfitting. The L_1-norm penalty $p_n(\eta) = \gamma_n\sqrt{n}|\eta|$ had also been considered. For small values of $|\eta|$, the L_1-norm penalty has the same behaviour as the SCAD penalty. However, in the development of the asymptotic theory of this method, we require that the penalty function be constant for sufficiently large values of $|\eta|$. Since the L_1-norm penalty increases linearly with $|\eta|$, the SCAD penalty is used in this method instead. A discussion of the selection of the tuning parameters a and γ_n can be found in Section 3.5.3.

The roles of the penalty functions can therefore be summarized as follows. By keeping the estimated values of π_k away from 0, the first penalty clusters the atoms θ_k around the atoms of the true mixing distribution of the marginal distribution. The SCAD penalty then merges each cluster into a single atom, obtaining the estimated order in a single maximization procedure.

3.5.1 Some Asymptotic Properties of the Maximum Penalized Quasi-Likelihood Estimator

We now study some asymptotic properties of the maximum penalized quasi-likelihood estimator. Recall that the mixing distribution G of the marginal distribution of Y_t is

given by

$$G(\theta) = \sum_{k=1}^{K} \pi_k I(\theta_k \leq \theta),$$

where $I(\cdot)$ is an indicator function, $\theta_1, \theta_2, ..., \theta_K \in \Theta$, and $\pi_1, \pi_2, ..., \pi_K$ satisfy $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

Let the class of all finite mixing distributions with at most K support points be given by

$$\mathcal{M}_K = \left\{ G(\theta) = \sum_{k=1}^{K} \pi_k I(\theta_k \leq \theta) : \theta_1 \leq \theta_2 \leq ... \leq \theta_K, \sum_{k=1}^{K} \pi_k = 1, \pi_k > 0 \right\}.$$

Note that $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq ... \subseteq \mathcal{M}_{K-1} \subseteq \mathcal{M}_K$ since the atoms θ_k are allowed to be equal with positive mixing proportions.

Let K_0 be the true number of support points of the finite mixing distribution G. The true mixing distribution G_0 is given by

$$G_0(\theta) = \sum_{k=1}^{K_0} \pi_{0k} I(\theta_{0k} \leq \theta),$$

where $\theta_{01} < \theta_{02} < ... < \theta_{0K_0}$ are K_0 distinct interior points of Θ and $(\pi_{01}, \pi_{02}, ..., \pi_{0K_0})$ are the true stationary mixing proportions.

Denote the maximizer of the penalized quasi-log-likelihood $\tilde{\ell}_n^Q(G)$ by $\hat{G}_n$, which we refer to as the maximum penalized quasi-likelihood estimator (MPQLE) of G. To study some of the asymptotic properties of the MPQLE $\hat{G}_n$, we must first define some notation.

We let $\hat{G}_n = \sum_{j=1}^{K} \hat{\pi}_j I(\hat{\theta}_j \leq \theta)$. Following Chen and Khalili (2008), we then define $I_k = \{j : \theta_{0,k-1} + \theta_{0,k} < 2\hat{\theta}_j \leq \theta_{0,k} + \theta_{0,k+1}\}$ for $k = 1, 2, ..., K_0$ with $\theta_{0,0} = -\infty$ and $\theta_{0,K_0+1} = \infty$, and

$$\hat{H}_k = \frac{\sum_{j \in I_k} \hat{\pi}_j I(\hat{\theta}_j \leq \theta)}{\sum_{j \in I_k} \hat{\pi}_j}$$

so that with $\hat{\alpha}_k = \sum_{j \in I_k} \hat{\pi}_j$,

$$\hat{G}_n(\theta) = \sum_{k=1}^{K_0} \hat{\alpha}_k \hat{H}_k(\theta).$$

Note that $\hat{\alpha}_1$ is the probability assigned to the support points $\hat{\theta}_j \leq (\theta_{01} + \theta_{02})/2$, $\hat{\alpha}_2$ is the probability assigned to the support points $(\theta_{01}+\theta_{02})/2 < \hat{\theta}_j \leq (\theta_{02}+\theta_{03})/2$, and so on.

The aim of our proposed method is to cluster the atoms of $\hat{H}_k$ into a small neighbourhood of θ_{0k} using the penalty on the mixing proportions π_k and then to merge the cluster of atoms into a single atom using the SCAD penalty.

In Lemma 3.1, we show that the MPQLE $\hat{G}_n$ satisfies $0 < \hat{\pi}_k < 1$ for $k = 1, 2, ..., K$ in probability as $n \to \infty$. We use similar proof techniques to those in Chen and Khalili (2008).

Lemma 3.1. *Let $Y_1, Y_2, ..., Y_n$ be a sample from a homogeneous and stationary HMM satisfying Conditions 1 to 3 in Section 2.2 and Assumptions 1 to 6 in Appendix B. Let $f(y; G_0)$ denote the true density function of the marginal mixture distribution of Y_t. Then the MPQLE $\hat{G}_n$ has the property*

$$\sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1)$$

as $n \to \infty$.

Proof. In what follows, the expectations are taken with respect to the true marginal mixture distribution G_0. Note that by Jensen's inequality,

$$\mathbb{E}\left[-\log\left\{\frac{f(Y;G)}{f(Y;G_0)}\right\}\right] \geq -\log\left\{\mathbb{E}\left[\frac{f(Y;G)}{f(Y;G_0)}\right]\right\} = -\log 1 = 0$$

under Assumption 1 for the existence of the expectation and Condition 2 for identifiability. Therefore, for any $G \neq G_0$, we have that

$$\mathbb{E}\left[\log f(Y;G)\right] - \mathbb{E}\left[\log f(Y;G_0)\right] < 0. \tag{3.3}$$

Now, as shown by Poskitt and Zhang (2005), since Y_t is stationary and ergodic from Condition 1,

$$\frac{1}{n}\ell_n^Q(G) \xrightarrow{a.s.} \mathbb{E}\left[\log f(Y;G)\right] \text{ as } n \to \infty.$$

Therefore,

$$\lim_{n\to\infty}\left\{\frac{1}{n}\ell_n^Q(G) - \frac{1}{n}\ell_n^Q(G_0)\right\} = \mathbb{E}\left[\log f(Y;G)\right] - \mathbb{E}\left[\log f(Y;G_0)\right] \tag{3.4}$$

almost surely. Poskitt and Zhang (2005) and references therein also show that this convergence is uniform.

Now combining results (3.3) and (3.4), we obtain

$$\ell_n^Q(G) - \ell_n^Q(G_0) < -Cn$$

almost surely for any $G \neq G_0$ with some $C > 0$. Due to the compactness of the space of G from Assumption 1, we can use a finite open coverage result in topology to strengthen the inequality to

$$\sup_{G \in N} \left\{ \ell_n^Q(G) - \ell_n^Q(G_0) \right\} < -Cn$$

for any compact neighbourhood N not containing G_0. In other words, the difference $\ell_n^Q(G) - \ell_n^Q(G_0)$ is negative in the order of n, uniformly for any G outside a neighbourhood of G_0.

Now when $\gamma_n = n^{1/4} \log n$, the SCAD penalty $p_n(\cdot)$ for any $a > 2$ becomes

$$p_n(\eta) = \begin{cases} n^{3/4} \log n |\eta| & \text{if } |\eta| \leq n^{-1/4} \log n \\ \dfrac{2n^{3/4} \log n |\eta| (a-1) - \sqrt{n}(n^{1/4}|\eta| - \log n)^2}{2(a-1)} & \text{if } n^{-1/4} \log n < |\eta| \leq a n^{-1/4} \log n \\ \dfrac{\sqrt{n} (\log n)^2 (a+1)}{2} & \text{if } |\eta| > a n^{-1/4} \log n \end{cases}$$

and it is straightforward to check that $\lim_{n \to \infty} \frac{p_n(\eta)}{n} = 0$. Due to this property when $\gamma_n = n^{1/4} \log n$, the SCAD penalty $p_n(\cdot)$ for any $a > 2$ satisfies

$$\sum_{k=1}^{K-1} p_n(\eta_k) - \sum_{k=1}^{K_0 - 1} p_n(\eta_{0k}) = o(n).$$

Therefore, since the addition of the SCAD penalty to $\ell_n^Q(G)$ does not change the order assessment and the term $C_K \sum_{k=1}^{K} \log \pi_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k}$ is constant with respect to n, we have that

$$\sup_{G \in N} \left\{ \tilde{\ell}_n^Q(G) - \tilde{\ell}_n^Q(G_0) \right\} \leq -Cn.$$

Thus, since $\hat{G}_n$ is the maximizer of $\tilde{\ell}_n^Q(G)$, it must be in a small neighbourhood of G_0. That is, $\hat{G}_n \xrightarrow{P} G_0$ and so it has at least K_0 distinct support points.

Since each η_{0k} is positive and approximated by one of the estimated differences in atoms $\hat{\eta}_k$, we must have that $p_n(\hat{\eta}_k) = p_n(\eta_{0k})$ in probability due to the fact that the SCAD

penalty is constant outside a neighbourhood of 0. Therefore,

$$\sum_{k=1}^{K-1} p_n(\hat{\eta}_k) - \sum_{k=1}^{K_0-1} p_n(\eta_{0k}) \geq 0$$

in probability.

Now let $\bar{G}_n$ be the maximum quasi-likelihood estimator (MQLE) of G_0 so that $\bar{G}_n$ has at most K support points. Then, by definition,

$$\begin{aligned}
0 &\leq \tilde{\ell}_n^Q(\hat{G}_n) - \tilde{\ell}_n^Q(G_0) \\
&= \left\{\ell_n^Q(\hat{G}_n) - \ell_n^Q(G_0)\right\} - \left\{\sum_{k=1}^{K-1} p_n(\hat{\eta}_k) - \sum_{k=1}^{K_0-1} p_n(\eta_{0k})\right\} \\
&+ \left\{C_K \sum_{k=1}^{K} \log \hat{\pi}_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k}\right\} \\
&\leq \left\{\ell_n^Q(\hat{G}_n) - \ell_n^Q(G_0)\right\} + \left\{C_K \sum_{k=1}^{K} \log \hat{\pi}_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k}\right\} \\
&\leq \left\{\ell_n^Q(\bar{G}_n) - \ell_n^Q(G_0)\right\} + \left\{C_K \sum_{k=1}^{K} \log \hat{\pi}_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k}\right\}.
\end{aligned}$$

Now from Holzmann and Schwaiger (2012), who establish the asymptotic distribution of the quasi-likelihood ratio test statistic under the assumptions of stationarity and ergodicity of Y_t,

$$\ell_n^Q(\bar{G}_n) - \ell_n^Q(G_0) = O_p(1).$$

In addition, since the quantity $C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k}$ is constant with respect to n, we have that

$$C_K \sum_{k=1}^{K} \log \hat{\pi}_k \geq -\left\{\ell_n^Q(\bar{G}_n) - \ell_n^Q(G_0)\right\} + C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k} = O_p(1).$$

□

In Theorem 3.1, we show that the MPQLE $\hat{G}_n$ is a consistent estimator of the true mixing distribution G_0 of the marginal distribution. We use similar proof techniques to those in Khalili (2005).

Theorem 3.1. *Let $Y_1, Y_2, ..., Y_n$ be a sample from a homogeneous and stationary HMM satisfying Conditions 1 to 3 in Section 2.2 and Assumptions 1 to 6 in Appendix B. Let*

$f(y; G_0)$ denote the true density function of the marginal mixture distribution of Y_t. Suppose that we apply the SCAD penalty with $\gamma_n = n^{1/4} \log n$. Then

(a) $\hat{G}_n$ is a consistent estimator of G_0

(b) all support points of $\hat{H}_k$ converge in probability to θ_{0k} for each $k = 1, 2, ..., K_0$.

Proof of (a). Let $\|H_1 - H_2\| = \sup_\theta |H_1(\theta) - H_2(\theta)|$. We will show the following two results:

(i) for $k = 1, 2, ..., K_0$, $\hat{\alpha}_k = \pi_{0k} + o_p(1)$,

(ii) for $k = 1, 2, ..., K_0$, $\|\hat{H}_k - H_{0k}\| = o_p(1)$, where $H_{0k} = I(\theta_{0k} \leq \theta)$.

In what follows, the expectations are taken with respect to the true mixing distribution G_0. From Lemma 3.1, we saw that

$$\frac{1}{n} \left\{ \ell_n^Q(G) - \ell_n^Q(G_0) \right\} \to \mathbb{E} \left[\log f(Y; G) \right] - \mathbb{E} \left[\log f(Y; G_0) \right]$$

as $n \to \infty$ almost surely and uniformly over the compact space of G. Now since $\lim_{n\to\infty} \frac{1}{n} \left\{ C_K \sum_{k=1}^{K} \log \pi_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k} \right\} = 0$ and the SCAD penalty $p_n(\eta)$ with $\gamma_n = n^{1/4} \log n$ and $a > 2$ satisfies $\lim_{n\to\infty} \frac{p_n(\eta)}{n} = 0$, we have that

$$\frac{1}{n} \left\{ \tilde{\ell}_n^Q(G) - \tilde{\ell}_n^Q(G_0) \right\} \to \mathbb{E} \left[\log f(Y; G) \right] - \mathbb{E} \left[\log f(Y; G_0) \right] \tag{3.5}$$

as $n \to \infty$ almost surely and uniformly over the compact space of G.

Now suppose for the sake of contradiction that parts (i) and (ii) do not hold and consider the set

$$A = \{ G \in \mathcal{M}_K : \|H_k - H_{0k}\| > \epsilon_1, |\alpha_k - \pi_{0k}| > \epsilon_2, 1 \leq k \leq K_0, \pi_l \in [\delta_{1l}, \delta_{2l}], 1 \leq l \leq K \}$$

for some $\epsilon_1, \epsilon_2 > 0$ and $0 < \delta_{1l}, \delta_{2l} < 1$. Then due to the compactness of the parameter space Θ and the results of Lemma 3.1, there must exist a subsequence $\hat{G}_{n_s}$ of $\hat{G}_n$ satisfying $P(\hat{G}_{n_s} \in A) > \epsilon$ for some $\epsilon > 0$ and sufficiently large n_s. This implies that

$$P \left\{ \frac{1}{n_s} \left[\tilde{\ell}_{n_s}^Q(\hat{G}_{n_s}) - \tilde{\ell}_{n_s}^Q(G_0) \right] = \sup_{G \in A} \frac{1}{n_s} \left[\tilde{\ell}_{n_s}^Q(G) - \tilde{\ell}_{n_s}^Q(G_0) \right] \right\} > \epsilon$$

for all n_j. On the other hand, $\mathbb{E} \left[\log f(Y; G) \right] - \mathbb{E} \left[\log f(Y; G_0) \right] < 0$ for any $G \in A$ under

Condition 2 for identifiability. Then from Equation (3.5),

$$P\left\{\frac{1}{n_s}\left[\tilde{\ell}^Q_{n_s}(G) - \tilde{\ell}^Q_{n_s}(G_0)\right] < 0\right\} > \epsilon \tag{3.6}$$

for sufficiently large n_s. This, however, contradicts the fact that $\hat{G}_{n_s}$ is the maximizer of the penalized quasi-likelihood $\tilde{\ell}^Q_n(G)$. Therefore, the result of part (a) must hold. □

Proof of (b). In Lemma 3.1, we had shown that the mixing proportion on each atom of $\hat{G}_n$ is positive in probability. On the other hand, from part (a)-(i), we obtained the result

$$\|\hat{H}_k - H_{0k}\| = \sup_\theta \left|\hat{H}_k(\theta) - H_{0k}(\theta)\right| = \sup_\theta \left|\hat{H}_k(\theta) - I(\theta_{0k} \leq \theta)\right| = o_p(1).$$

This means that for all $\theta \in \Theta$,

$$\hat{H}_k(\theta) = \begin{cases} 1 & \text{if } \theta \geq \theta_{0k} \\ 0 & \text{if } \theta < \theta_{0k} \end{cases}$$

for sufficiently large n in probability. Thus, we must have that the atoms of $\hat{H}_k$ converge to the true atom θ_{0k} in probability for each $k = 1, 2, ..., K_0$. □

Note that we have not yet shown that the MPQLE is consistent in estimating the true order K_0. From Theorem 3.1, the order may still be overestimated if any $\hat{H}_k$ has more than one atom. Due to the more complex structure of a HMM, we have yet to be able to extend the consistency of MSCAD in estimating the true order in the finite mixture setting to the HMM context, and must therefore defer this task to future work.

3.5.2 Numerical Computation

Let $Y_1, Y_2, ..., Y_n$ be a sample from a homogeneous, stationary K-state HMM, where the marginal distribution of (Y_t) is given by the finite mixture

$$f(y_t; \mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k f(y_t; \theta_k),$$

with vector of parameters $\mathbf{\Psi} = (\pi_1, \pi_2, ..., \pi_{K-1}, \theta_1, \theta_2, ..., \theta_K)$. The quasi-log-likelihood function of $\mathbf{\Psi}$ based on the above sample is given by

$$\ell^Q_n(\mathbf{\Psi}) = \sum_{t=1}^{n} \log f(y_t; \mathbf{\Psi}).$$

Our goal is to maximize the penalized quasi-log-likelihood function

$$\tilde{\ell}^Q_n(\mathbf{\Psi}) = \ell^Q_n(\mathbf{\Psi}) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_n(\eta_k)$$

over the space $\mathcal{M}_K$ for a pre-specified K, where $\eta_k = \theta_{k+1} - \theta_k$ for $k = 1, 2, ..., K-1$ and $C_K > 0$.

Since the SCAD penalty function $p_n(\eta)$ is singular at $\eta = 0$ and does not have continuous second order derivatives, we follow the suggestion of Fan and Li (2001) of replacing $p_n(\eta)$ by a local quadratic approximation (LQA) in a neighbourhood of η_0, given by

$$p_n(\eta) \simeq p_n(\eta_0) + \frac{p_n'(\eta_0)}{2\eta_0}(\eta^2 - \eta_0^2).$$

In the context of variable selection in regression, a local linear approximation (LLA; Zou and Li, 2008) of $p_n(\eta)$ for $\eta \approx \eta_0$ was also proposed:

$$p_n(\eta) \simeq p_n(\eta_0) + p_n'(\eta_0)(\eta - \eta_0).$$

We consider this approximation and discuss its performance in simulation studies later on in this section.

To perform the maximization of the penalized quasi-log-likelihood, we use a revised EM algorithm (Dempster et al., 1977) as follows.

First we define the unobserved indicator variables

$$u_{tk} = \begin{cases} 1 & \text{if } Z_t = k \\ 0 & \text{otherwise} \end{cases},$$

which denote the state membership of the observation y_t in the HMM. Using the complete data $(y_1, \boldsymbol{u}_1), ..., (y_n, \boldsymbol{u}_n)$, where $\boldsymbol{u}_t = (u_{t1}, u_{t2}, ..., u_{tK})^T$, the complete quasi-likelihood function is defined as

$$\mathcal{L}_n^{Q(C)}(\boldsymbol{\Psi}) = \prod_{t=1}^{n} \prod_{k=1}^{K} \{\pi_k f(y_t; \theta_k)\}^{u_{tk}}$$

so that the complete quasi-log-likelihood function is

$$\ell_n^{Q(C)}(\boldsymbol{\Psi}) = \sum_{t=1}^{n} \sum_{k=1}^{K} u_{tk} \{\log \pi_k + \log f(y_t; \theta_k)\}.$$

The penalized complete quasi-log-likelihood function is then

$$\tilde{\ell}_n^{Q(C)}(\boldsymbol{\Psi}) = \ell_n^{Q(C)}(\boldsymbol{\Psi}) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_n(\eta_k).$$

The EM algorithm maximizes $\tilde{\ell}_n^{Q(C)}(\boldsymbol{\Psi})$ by repeatedly alternating between two steps: the E-step and the M-step.

E-step: The E-step consists of computing the conditional expectation of $\tilde{\ell}_n^{Q(C)}(\boldsymbol{\Psi})$ with respect to u_{tk}, given the observed data $\boldsymbol{y}$ and the current parameter estimate $\boldsymbol{\Psi}^{(m)}$:

$$\begin{aligned} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) &= \mathbb{E}\left[\tilde{\ell}_n^{Q(C)}(\boldsymbol{\Psi}) \mid \boldsymbol{y}, \boldsymbol{\Psi}^{(m)}\right] \\ &= \sum_{t=1}^{n}\sum_{k=1}^{K}\left\{w_{tk}^{(m)} + \frac{C_K}{n}\right\}\log \pi_k + \sum_{t=1}^{n}\sum_{k=1}^{K} w_{tk}^{(m)} \log f(y_t; \theta_k) - \sum_{k=1}^{K-1} p_n(\eta_k), \end{aligned}$$

where, for $k = 1, 2, ..., K$,

$$w_{tk}^{(m)} = \mathbb{E}\left[u_{tk} \mid y_t, \boldsymbol{\Psi}^{(m)}\right] = P(Z_t = k \mid Y_t = y_t, \boldsymbol{\Psi}^{(m)}) = \frac{\pi_k^{(m)} f(y_t; \theta_k^{(m)})}{\sum_{l=1}^{K} \pi_l^{(m)} f(y_t; \theta_l^{(m)})}$$

is the conditional expectation of u_{tk} given the observation y_t and the current parameter estimate $\boldsymbol{\Psi}^{(m)}$.

M-step: The M-step consists of maximizing $Q(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(m)})$ with respect to $\boldsymbol{\Psi}$ on the $(m+1)^{\text{th}}$ iteration. That is, we need to find

$$\boldsymbol{\Psi}^{(m+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}).$$

By maximizing $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$ with respect to the mixing proportion π_k, we obtain

$$\pi_k^{(m+1)} = \frac{\sum_{t=1}^{n} w_{tk}^{(m)} + C_K}{n + KC_K} \text{ for } k = 1, 2, ..., K$$

as the updated estimate of π_k.

To maximize $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$ with respect to θ_k, we replace the penalty $p_n(\eta_k)$ by

$$\tilde{p}_n(\eta_k; \eta_k^{(m)}) = p_n(\eta_k^{(m)}) + \frac{p_n'(\eta_k^{(m)})}{2\eta_k^{(m)}}(\eta_k^2 - \eta_k^{(m)2})$$

and solve the following system of equations

$$\begin{aligned} \sum_{t=1}^{n} w_{t1}^{(m)} \frac{\partial}{\partial \theta_1}\left\{\log f(y_t; \theta_1)\right\} + \frac{\partial \tilde{p}_n(\eta_1; \eta_1^{(m)})}{\partial \theta_1} &= 0, \\ \sum_{t=1}^{n} w_{tk}^{(m)} \frac{\partial}{\partial \theta_k}\left\{\log f(y_t; \theta_k)\right\} - \frac{\partial \tilde{p}_n(\eta_{k-1}; \eta_{k-1}^{(m)})}{\partial \theta_k} + \frac{\partial \tilde{p}_n(\eta_k; \eta_k^{(m)})}{\partial \theta_k} &= 0, \quad k = 2, 3, ..., K-1, \\ \sum_{t=1}^{n} w_{tK}^{(m)} \frac{\partial}{\partial \theta_K}\left\{\log f(y_t; \theta_K)\right\} - \frac{\partial \tilde{p}_n(\eta_{K-1}; \eta_{K-1}^{(m)})}{\partial \theta_K} &= 0 \end{aligned}$$

to obtain the update equations of the state-dependent parameters θ_k.

Starting from an initial value $\boldsymbol{\Psi}^{(0)}$, the E and M steps are iterated until some convergence criterion is met. We use the convergence criterion

$$\|\boldsymbol{\Psi}^{(m+1)} - \boldsymbol{\Psi}^{(m)}\| < \epsilon$$

for some pre-specified value $\epsilon > 0$. We take the initial value of θ_k to be the $100(k - 1/2)/K\%$ sample quantile and the initial value of π_k to be $1/K$ for each $k = 1, 2, ..., K$.

For Poisson state-dependent distributions with mean parameters λ_k, the system of equations presented above do not yield closed-form update equations for λ_k. We considered two possible ways of addressing this issue. The first approach is to use the following quadratic approximation of $\log f(y; \lambda_k)$ in the function $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$ defined in the E-step:

$$\begin{aligned}
\log f(y; \lambda_k) &= \log \left\{ \frac{e^{-\lambda_k} \lambda_k^y}{y!} \right\} \\
&= -\lambda_k + y \log \lambda_k - \log y! \\
&\approx -\lambda_k + y \left\{ (\lambda_k - 1) - \frac{(\lambda_k - 1)^2}{2} \right\} - \log y! \\
&= \lambda_k (y - 1) - \frac{y(\lambda_k - 1)^2}{2} - \log y! - y.
\end{aligned}$$

For the second approach, we considered the reparametrization

$$\begin{aligned}
\lambda_1 &= \theta \\
\lambda_2 &= \theta + \eta_1 \\
&\vdots \\
\lambda_K &= \theta + \eta_1 + ... + \eta_{K-1}
\end{aligned}$$

as well as replacing the penalty function by a LLA of $p_n(\eta)$ in a neighbourhood of η_0.

In our simulation studies with Poisson HMMs, presented in Section 3.6, we use the first approach to obtain closed-form update equations for λ_k. We also assessed the performance of the second approach via simulation and found that using the local linear approximation of the penalty along with the reparametrization does not have a considerable effect on the success rate of the method compared to when using the local quadratic approximation.

3.5.3 Tuning Parameter Selection

In many penalized likelihood or regularization methods, the performance of the method depends strongly on the tuning parameter controlling the extent of penalization. Our proposed method requires the selection of three tuning parameters, γ_n and a from the SCAD penalty as well as C_K from the penalty on the mixing proportions π_k. According to Fan and Li (2001), one may do a two-dimensional grid search to find the optimal pair (λ_n, a), where $\lambda_n = \gamma_n/\sqrt{n}$, based on cross-validation (Stone, 1974) or generalized

cross-validation (Craven and Wahba, 1979). However, this can be computationally expensive. As another approach for selecting the tuning parameter a, Fan and Li (2001) use Bayesian risk analysis. They find that certain Bayes risk criteria are minimized at $a \approx 3.7$ and therefore take $a = 3.7$. This value of a was found to perform suitably for many variable selection problems. Thus, in our simulation studies, we also take $a = 3.7$. As for the constant C_K, Chen, Chen and Kalbfleisch (2001) suggest using $C_K = \log M$ if the parameters θ_k are restricted to $[-M, M]$ for large M. Therefore, we take $C_K = \log 20$.

For the selection of the tuning parameter γ_n, we use cross-validation. Cross-validation (CV) consists of the repeated partition of the data into two parts, where the first part is used to fit the model and the second part is used to evaluate the fitted model. The tuning parameter γ_n is chosen so that the prediction error is minimized.

To use cross-validation, we first partition the data into N equal parts with each part referred to as the *test* data set since it is used to evaluate the fitted model. For each $i = 1, 2, ..., N$, the i^{th} part is then removed from the data and the set of remaining observations, which is referred to as the *training* data set, is used to fit the model. Let us consider this procedure in the context of our problem. Let $\hat{\boldsymbol{\Psi}}_{n,-i}$ be the MPQLE of $\boldsymbol{\Psi}$ based on the training set. Further, let $\ell^Q_{n,i}(\hat{\boldsymbol{\Psi}}_{n,-i})$ be the quasi-log-likelihood function evaluated on the test set using the MPQLE $\hat{\boldsymbol{\Psi}}_{n,-i}$ for $i = 1, 2, ..., N$. The cross-validation criterion is given by

$$CV(\gamma_n) = \frac{1}{N} \sum_{i=1}^{N} \ell^Q_{n,i}(\hat{\boldsymbol{\Psi}}_{n,-i}).$$

We select the value of γ_n that minimizes $CV(\gamma_n)$. This procedure is called N-fold cross-validation. When $N = n$, we obtain what is called leave-one-out cross-validation. When the sample size n is large, the computational burden associated with the leave-one-out CV criterion can be considerable. In this case, one may want to consider 5-fold or 10-fold CV.

3.6 Simulation Studies

In this section, the performance of the proposed order selection method for HMMs as well as the two information criteria AIC and BIC, based on both the full-model likelihood and quasi-likelihood, are studied via simulation.

We focus on the problem of order selection in normal and Poisson HMMs. We report the percentage of times out of 500 replications that the estimated order equals a given value of K with sample sizes $n = 100, 400$. To select the tuning parameter γ_n, we use cross-validation for samples of size $n = 100$ and 5-fold cross-validation for samples of size $n = 400$. As in Chen and Khalili (2008), we take [0.2, 1.5] as the range of $\gamma_n/\sqrt{n}$ for the normal HMMs and [0.4, 1.6] for the Poisson HMMs. These ranges meet the conditions stipulated in the theory on the order of γ_n for the sample sizes under consideration.

The performance of the methods are assessed for HMMs of order 2, 3, 4 and 6 with the dependence structures S1-S13 displayed in Tables 3.1, 3.2, and 3.3. Notice that from S1 to S3 in Table 3.1, the stationary distribution $\boldsymbol{\pi}$ is the same, but the transition matrices $\mathbb{P}$ are different. For the 2-state HMM with dependence structure S1, there is no expected correlation between the observations. For the 2-state HMM with dependence structure S2, negative and positive correlation are expected between the observations, while positive correlation is expected for the model with dependence structure S3. Therefore, we will be able to observe the influence of the different dependence structures on the performance of the methods under study. In Figure 3.2, we plot the theoretical ACF for the normal HMMs.

Example 3.6.1. For the normal HMMs under consideration, the marginal distribution of the observations is given by the finite mixture

$$f(y; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \frac{\pi_k}{\sigma} \Phi\left(\frac{y - \theta_k}{\sigma}\right),$$

where $\boldsymbol{\Psi} = (\theta_1, \theta_2, ..., \theta_K, \sigma, \pi_1, \pi_2, ...\pi_{K-1})$ and $\Phi(\cdot)$ is the standard normal density function. We set the upper bound $K = 15$ for all models in this study.

We simulated data from four 2-state normal HMMs with $\boldsymbol{\theta} = (0, 3)$, $\boldsymbol{\sigma} = (1, 1)$, and dependence structures S1-S4, displayed in Table 3.1. The parameters of the marginal mixture distributions are the same as those in the simulation studies of Ishwaran et al. (2001) for 2-component normal mixtures.

The results can be found in Tables 3.4 and 3.5. For sample size $n = 100$, MSCAD based on the quasi-likelihood performs reasonably well. It outperformed AIC based on

the full-model likelihood and BIC based on the quasi-likelihood in all cases. It also appears that the performance of MSCAD_Q is not significantly affected by the dependence structure. Once dependence within the observations was introduced, the number of times MSCAD_Q selected the correct order decreased, but only slightly. For sample size $n = 400$, both BIC and BIC_Q performed very well. When AIC and AIC_Q did not select the correct order, they overfit the data. MSCAD_Q was not the best, but it did outperform AIC based on the full likelihood in all cases.

Now let us consider the 3-state normal HMMs with $\boldsymbol{\theta} = (0, 3, 6)$, $\boldsymbol{\sigma} = (1, 1, 1)$, and dependence structures S5-S7 in Table 3.1. The results can be found in Tables 3.6 and 3.7. For sample size $n = 100$, MSCAD_Q outperformed the other methods in all three cases. BIC and BIC_Q had the tendency to underfit. For sample size $n = 400$, BIC based on the full-model likelihood was the best in two out of three cases. MSCAD_Q was on par with the other methods.

The 4-state normal HMMs under consideration have dependence structures S8-S10 (see Table 3.2), $\boldsymbol{\theta} = (0, 3, 6, 9)$ and standard deviation $\sigma = 1$ across all states. In Table 3.8, MSCAD_Q was indisputably the best for sample size $n = 100$. BIC and BIC_Q both had the tendency to underfit, often selecting models of order 2. For sample size $n = 400$ (see Table 3.9), MSCAD_Q outperformed the other methods in two out of three cases. AIC and AIC_Q also performed reasonably well.

The final normal HMMs under consideration are of order 6 with $\boldsymbol{\theta} = (0, 3, 6, 9, 12, 15)$, common standard deviation $\sigma = 1$ and dependence structures S11-S13 in Table 3.3. The results are displayed in Table 3.10. AIC, AIC_Q and MSCAD_Q all performed reasonably well. Both BIC and BIC_Q had the tendency to underfit.

Structure	$\mathbb{P}$	$\boldsymbol{\pi}$
S1	$\begin{pmatrix} 0.50 & 0.50 \\ 0.50 & 0.50 \end{pmatrix}$	(0.50, 0.50)
S2	$\begin{pmatrix} 0.25 & 0.75 \\ 0.75 & 0.25 \end{pmatrix}$	(0.50, 0.50)
S3	$\begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$	(0.50, 0.50)
S4	$\begin{pmatrix} 0.20 & 0.80 \\ 0.40 & 0.60 \end{pmatrix}$	$(\frac{1}{3}, \frac{2}{3})$
S5	$\begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
S6	$\begin{pmatrix} 0.70 & 0.20 & 0.10 \\ 0.20 & 0.60 & 0.20 \\ 0.10 & 0.10 & 0.80 \end{pmatrix}$	$\approx (0.316, 0.263, 0.421)$
S7	$\begin{pmatrix} 0.10 & 0.20 & 0.70 \\ 0.20 & 0.60 & 0.20 \\ 0.80 & 0.10 & 0.10 \end{pmatrix}$	$\approx (0.374, 0.275, 0.352)$

Table 3.1: Transition matrices and corresponding stationary distributions in simulation studies for 2-state and 3-state HMMs (S1-S7).

Structure	$\mathbb{P}$	$\boldsymbol{\pi}$
S8	$\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
S9	$\begin{pmatrix} 0.70 & 0 & 0.10 & 0.20 \\ 0 & 0.80 & 0.20 & 0 \\ 0.10 & 0 & 0.70 & 0.20 \\ 0.10 & 0.10 & 0 & 0.80 \end{pmatrix}$	$(0.20, 0.20, 0.20, 0.40)$
S10	$\begin{pmatrix} 0.10 & 0.10 & 0.40 & 0.40 \\ 0.30 & 0.10 & 0.50 & 0.10 \\ 0.10 & 0.60 & 0.10 & 0.20 \\ 0.40 & 0.10 & 0.20 & 0.30 \end{pmatrix}$	$\approx (0.222, 0.244, 0.289, 0.244)$

Table 3.2: Transition matrices and corresponding stationary distributions in simulation studies for 4-state HMMs (S8-S10).

Structure	$\mathbb{P}$	$\boldsymbol{\pi}$
S11	$\begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$	$(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
S12	$\begin{pmatrix} 0.2 & 0.4 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.4 & 0.1 & 0.1 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.1 & 0.2 & 0.2 \\ 0.1 & 0.2 & 0.1 & 0.2 & 0.1 & 0.3 \\ 0.1 & 0.2 & 0.1 & 0.2 & 0.3 & 0.1 \\ 0.1 & 0.2 & 0.1 & 0.2 & 0.1 & 0.3 \end{pmatrix}$	$\approx (0.122, 0.281, 0.100, 0.150, 0.173, 0.175)$
S13	$\begin{pmatrix} 0.1 & 0.2 & 0.1 & 0.1 & 0.1 & 0.4 \\ 0.1 & 0.2 & 0.1 & 0.1 & 0.4 & 0.1 \\ 0.2 & 0.1 & 0.1 & 0.4 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.4 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.4 & 0.1 & 0.1 & 0.2 & 0.1 \\ 0.4 & 0.2 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}$	$\approx (0.163, 0.208, 0.143, 0.143, 0.180, 0.163)$

Table 3.3: Transition matrices and corresponding stationary distributions in simulation studies for 6-state HMMs (S11-S13).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
1	2	1	0.048	0.410	0.018	0.198	0.062
		2	**0.792**	**0.586**	**0.928**	**0.802**	**0.934**
		3	0.122	0.004	0.054	0.000	0.004
		4	0.038	0.000	0.000	0.000	0.000
2	2	1	0.002	0.032	0.026	0.220	0.052
		2	**0.878**	**0.966**	**0.904**	**0.768**	**0.898**
		3	0.090	0.002	0.070	0.012	0.048
		4	0.003	0.000	0.000	0.000	0.002
3	2	1	0.000	0.038	0.028	0.204	0.042
		2	**0.892**	**0.962**	**0.908**	**0.792**	**0.906**
		3	0.082	0.000	0.060	0.004	0.052
		4	0.026	0.000	0.004	0.000	0.000
4	2	1	0.014	0.272	0.012	0.186	0.024
		2	**0.860**	**0.728**	**0.946**	**0.806**	**0.918**
		3	0.110	0.000	0.042	0.008	0.054
		4	0.016	0.000	0.000	0.000	0.004

Table 3.4: Simulation results for 2-state normal HMMs: the percentage of times in which order K_0 was estimated by each method (n=100).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
1	2	1	0.000	0.000	0.000	0.000	0.028
		2	**0.830**	**1.000**	**0.952**	**1.000**	**0.940**
		3	0.120	0.000	0.048	0.000	0.032
		4	0.050	0.000	0.000	0.000	0.000
2	2	1	0.000	0.000	0.000	0.000	0.016
		2	**0.892**	**1.000**	**0.954**	**1.000**	**0.952**
		3	0.078	0.000	0.046	0.000	0.032
		4	0.030	0.000	0.000	0.000	0.000
3	2	1	0.000	0.000	0.000	0.000	0.006
		2	**0.876**	**1.000**	**0.940**	**1.000**	**0.974**
		3	0.094	0.000	0.060	0.000	0.020
		4	0.030	0.000	0.000	0.000	0.000
4	2	1	0.000	0.000	0.004	0.006	0.006
		2	**0.874**	**1.000**	**0.936**	**0.994**	**0.954**
		3	0.104	0.000	0.060	0.000	0.040
		4	0.022	0.000	0.000	0.000	0.000

Table 3.5: Simulation results for 2-state normal HMMs: the percentage of times in which order K_0 was estimated by each method (n=400).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
5	3	1	0.010	0.432	0.004	0.164	0.000
		2	0.310	**0.530**	0.136	0.400	0.004
		3	**0.580**	0.038	**0.800**	**0.436**	**0.836**
		4	0.100	0.000	0.060	0.000	0.160
6	3	1	0.000	0.000	0.002	0.052	0.000
		2	0.178	**0.976**	0.210	**0.574**	0.012
		3	0.000	0.000	**0.666**	0.366	**0.820**
		4	**0.822**	0.024	0.114	0.008	0.166
		5	0.000	0.000	0.008	0.000	0.002
7	3	1	0.000	0.002	0.008	0.082	0.000
		2	0.024	**0.418**	0.136	0.404	0.018
		3	**0.500**	0.286	**0.780**	**0.510**	**0.822**
		4	0.476	0.294	0.076	0.004	0.156
		5	0.000	0.000	0.000	0.000	0.004

Table 3.6: Simulation results for 3-state normal HMMs: the percentage of times in which order K_0 was estimated by each method (n=100).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
5	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.434	0.000	0.026	0.002
		3	**0.866**	**0.566**	**0.950**	**0.974**	**0.982**
		4	0.096	0.000	0.048	0.000	0.016
		5	0.038	0.000	0.002	0.000	0.000
6	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.060	0.028
		3	**0.924**	**1.000**	**0.850**	**0.928**	**0.970**
		4	0.068	0.000	0.144	0.012	0.002
		5	0.008	0.000	0.006	0.000	0.000
7	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.022	0.002
		3	**0.914**	**1.000**	**0.964**	**0.978**	**0.972**
		4	0.072	0.000	0.036	0.000	0.026
		5	0.014	0.000	0.000	0.000	0.000

Table 3.7: Simulation results for 3-state normal HMMs: the percentage of times in which order K_0 was estimated by each method (n=400).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
8	4	1	0.008	0.336	0.000	0.086	0.000
		2	**0.404**	**0.648**	0.128	**0.622**	0.000
		3	0.234	0.016	0.180	0.132	0.024
		4	0.302	0.000	**0.624**	0.158	**0.704**
		5	0.028	0.000	0.060	0.002	0.264
		6	0.024	0.000	0.008	0.000	0.008
9	4	1	0.000	0.000	0.002	0.034	0.000
		2	0.000	0.050	0.134	**0.522**	0.000
		3	0.178	**0.484**	0.344	0.326	0.024
		4	**0.670**	0.466	**0.378**	0.110	**0.672**
		5	0.132	0.000	0.120	0.008	0.292
		6	0.020	0.000	0.022	0.000	0.012
10	4	1	0.002	0.218	0.020	0.282	0.000
		2	0.048	**0.626**	0.176	**0.492**	0.000
		3	0.030	0.040	0.160	0.090	0.100
		4	**0.820**	0.116	**0.586**	0.132	**0.832**
		5	0.090	0.090	0.052	0.004	0.068
		6	0.020	0.000	0.006	0.000	0.000

Table 3.8: Simulation results for 4-state normal HMMs: the percentage of times in which order K_0 was estimated by each method (n=100).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
8	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.008	**0.844**	0.000	0.046	0.000
		3	0.028	0.126	0.002	0.072	0.020
		4	**0.806**	0.030	**0.942**	**0.880**	**0.952**
		5	0.134	0.000	0.042	0.002	0.028
		6	0.024	0.000	0.014	0.000	0.000
9	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.048	0.000
		3	0.000	0.014	0.024	0.230	0.004
		4	**0.760**	**0.778**	0.322	**0.464**	**0.876**
		5	0.198	0.182	**0.654**	0.258	0.120
		6	0.042	0.026	0.000	0.000	0.000
10	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.012	0.000	0.084	0.000
		3	0.000	0.000	0.002	0.048	0.028
		4	**0.898**	**0.988**	**0.942**	**0.868**	**0.950**
		5	0.078	0.000	0.056	0.000	0.022
		6	0.024	0.000	0.000	0.000	0.000

Table 3.9: Simulation results for 4-state normal HMMs: the percentage of times in which order K_0 was estimated by each method (n=400).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
11	6	1	0.000	0.000	0.000	0.000	0.000
		2	0.002	**0.648**	0.000	0.026	0.000
		3	0.146	0.350	0.000	0.368	0.000
		4	0.240	0.002	0.016	0.186	0.000
		5	0.128	0.000	0.006	0.030	0.066
		6	**0.400**	0.000	**0.934**	**0.390**	**0.892**
		7	0.084	0.000	0.044	0.000	0.042
12	6	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	**0.636**	0.000	0.122	0.000
		3	0.044	0.364	0.010	**0.516**	0.000
		4	0.022	0.000	0.002	0.016	0.016
		5	0.162	0.000	0.066	0.080	0.132
		6	**0.668**	0.000	**0.810**	0.264	**0.832**
		7	0.104	0.000	0.112	0.002	0.020
13	6	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.214	0.000	0.238	0.000
		3	0.000	**0.676**	0.004	**0.352**	0.000
		4	0.000	0.098	0.002	0.078	0.000
		5	0.010	0.012	0.016	0.034	0.070
		6	**0.944**	0.000	**0.922**	0.298	**0.878**
		7	0.046	0.000	0.056	0.000	0.052

Table 3.10: Simulation results for 6-state normal HMMs: the percentage of times in which order K_0 was estimated by each method (n=400).

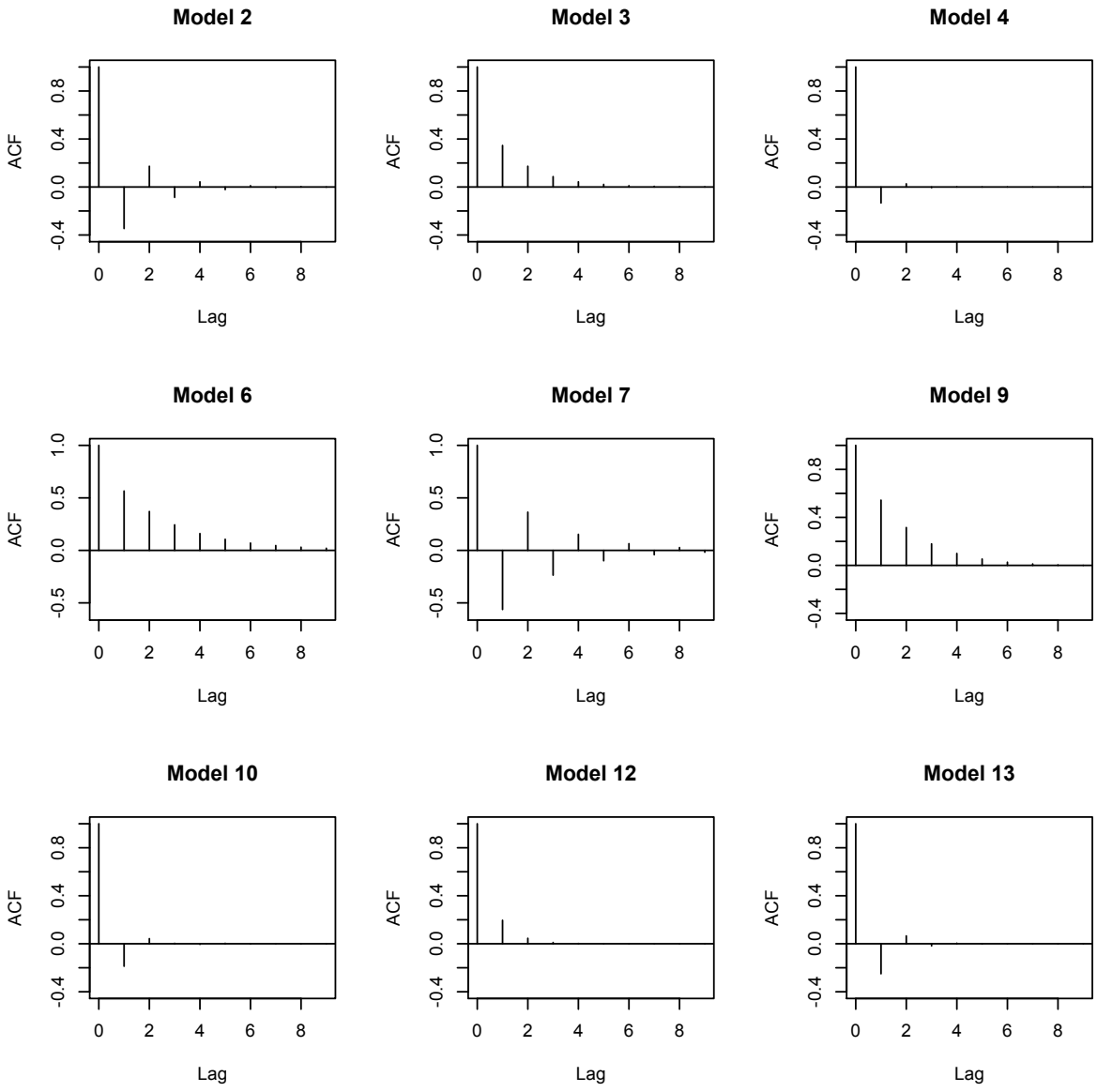

Figure 3.2: Theoretical ACF for normal HMMs.

Example 3.6.2. We then simulated data from Poisson HMMs. The marginal density of the HMM observations is given by

$$f(y;\mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k \left(\frac{e^{-\theta_k}\theta_k^y}{y!} \right),$$

where $\mathbf{\Psi} = (\theta_1, \theta_2, ..., \theta_K, \pi_1, \pi_2, ...\pi_{K-1})$. Again we set the upper bound $K = 15$.

For the 2-state Poisson HMMs under consideration, $\boldsymbol{\theta} = (1, 9)$ with dependence structures S1-S4 shown in Table 3.1. The results are displayed in Tables 3.11 and 3.12. For sample size $n = 100$, BIC performed the best and BIC_Q performed the second best in all cases. MSCAD_Q performed better than AIC in all cases. For sample size $n = 400$, BIC, BIC_Q and MSCAD_Q all performed well in detecting the true order. From these simulations, it appears that the proposed method is not affected by the dependence structure. In some cases, MSCAD_Q had a perfect success rate.

Now let us consider 3-state Poisson HMMs with $\boldsymbol{\theta} = (1, 5, 10)$ and dependence structures S5-S7 from Table 3.1. For sample size $n = 100$ (see Table 3.13), MSCAD_Q was the best in all three cases. Both BIC and BIC_Q had the tendency to underfit. For sample size $n = 400$ (see Table 3.14), either AIC or AIC_Q was the best out of the three cases. MSCAD_Q was on par with the other methods.

The 4-state Poisson HMMs have $\boldsymbol{\theta} = (1, 5, 10, 15)$, and $\mathbb{P}$ and $\boldsymbol{\pi}$ shown in Table 3.2. The results can be found in Tables 3.15 and 3.16. For sample size $n = 100$, MSCAD_Q was indisputably the best. The other methods most often selected a model of order 3. For sample size $n = 400$, the performance of AIC based on the full-model likelihood improved, but MSCAD_Q was either the best or second best out of the three cases.

Finally, the 6-state Poisson HMMs under consideration have $\boldsymbol{\theta} = (1, 5, 10, 15, 20, 25)$ and dependence structures S11-S13, displayed in Table 3.3. For sample size $n = 400$ (see Table 3.17), none of the methods performed well. In all three cases, both BIC and BIC_Q never selected the correct order out of the 500 simulated data sets. MSCAD_Q, however, provided an estimate closer to the true order most often. The poor performance by all methods is likely due to the fact that the state means of the true models are relatively close to each other for count data.

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
1	2	1	0.000	0.000	0.000	0.000	0.032
		2	**0.938**	**1.000**	**0.956**	**0.994**	**0.968**
		3	0.060	0.000	0.044	0.006	0.000
		4	0.002	0.000	0.000	0.000	0.000
2	2	1	0.000	0.000	0.000	0.000	0.016
		2	**0.956**	**1.000**	**0.952**	**0.994**	**0.984**
		3	0.038	0.000	0.048	0.006	0.000
		4	0.006	0.000	0.000	0.000	0.000
3	2	1	0.000	0.000	0.000	0.000	0.000
		2	**0.944**	**1.000**	**0.970**	**0.998**	**0.948**
		3	0.052	0.000	0.030	0.002	0.052
		4	0.004	0.000	0.000	0.000	0.000
4	2	1	0.000	0.000	0.000	0.000	0.000
		2	**0.928**	**1.000**	**0.962**	**0.996**	**0.972**
		3	0.070	0.000	0.038	0.004	0.028
		4	0.002	0.000	0.000	0.000	0.000

Table 3.11: Simulation results for 2-state Poisson-HMMs: the percentage of times in which order K_0 was estimated by each method (n=100).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
1	2	1	0.000	0.000	0.000	0.000	0.000
		2	**0.932**	**1.000**	**0.936**	**0.998**	**0.998**
		3	0.054	0.000	0.064	0.002	0.002
		4	0.014	0.000	0.000	0.000	0.000
2	2	1	0.000	0.000	0.000	0.000	0.000
		2	**0.936**	**1.000**	**0.922**	**1.000**	**1.000**
		3	0.062	0.000	0.076	0.000	0.000
		4	0.002	0.000	0.002	0.000	0.000
3	2	1	0.000	0.000	0.000	0.000	0.002
		2	0.082	**0.802**	**0.718**	**0.908**	**0.972**
		3	0.300	0.162	0.272	0.090	0.026
		4	**0.618**	0.036	0.010	0.002	0.000
4	2	1	0.000	0.000	0.000	0.000	0.000
		2	**0.942**	**1.000**	**0.946**	**1.000**	**1.000**
		3	0.052	0.000	0.054	0.000	0.000
		4	0.006	0.000	0.000	0.000	0.000

Table 3.12: Simulation results for 2-state Poisson-HMMs: the percentage of times in which order K_0 was estimated by each method (n=400).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
5	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.440	**0.962**	0.314	**0.690**	0.028
		3	**0.530**	0.038	**0.684**	0.308	**0.752**
		4	0.030	0.000	0.002	0.002	0.212
		5	0.000	0.000	0.000	0.000	0.008
6	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.290	**0.878**	0.424	**0.806**	0.060
		3	**0.700**	0.120	**0.576**	0.194	**0.766**
		4	0.010	0.002	0.000	0.000	0.172
		5	0.000	0.000	0.000	0.000	0.002
7	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.328	**0.886**	0.394	**0.792**	0.244
		3	**0.660**	0.114	**0.604**	0.208	**0.738**
		4	0.012	0.000	0.002	0.000	0.018
		5	0.000	0.000	0.000	0.000	0.000

Table 3.13: Simulation results for 3-state Poisson-HMMs: the percentage of times in which order K_0 was estimated by each method (n=100).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
5	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.008	**0.574**	0.006	0.076	0.014
		3	**0.890**	0.426	**0.970**	**0.924**	**0.942**
		4	0.096	0.000	0.024	0.000	0.044
		5	0.006	0.000	0.000	0.000	0.000
6	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.002	0.080	0.018	0.190	0.060
		3	**0.966**	**0.920**	**0.956**	**0.810**	**0.902**
		4	0.032	0.000	0.026	0.000	0.038
		5	0.000	0.000	0.000	0.000	0.000
7	3	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.048	0.008	0.156	0.050
		3	**0.932**	**0.952**	**0.960**	**0.844**	**0.908**
		4	0.068	0.000	0.032	0.000	0.042
		5	0.000	0.000	0.000	0.000	0.000

Table 3.14: Simulation results for 3-state Poisson-HMMs: the percentage of times in which order K_0 was estimated by each method (n=400).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
8	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.050	**0.520**	0.010	0.110	0.000
		3	**0.846**	0.478	**0.932**	**0.888**	0.302
		4	0.104	0.002	0.058	0.002	**0.526**
		5	0.000	0.000	0.000	0.000	0.166
		6	0.000	0.000	0.000	0.000	0.006
9	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.040	0.240	0.054	0.248	0.006
		3	**0.888**	**0.758**	**0.898**	**0.750**	0.362
		4	0.072	0.002	0.048	0.002	**0.508**
		5	0.000	0.000	0.000	0.000	0.118
		6	0.000	0.000	0.000	0.000	0.006
10	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.038	0.488	0.020	0.136	0.000
		3	**0.750**	**0.512**	**0.890**	**0.864**	0.266
		4	0.210	0.000	0.090	0.000	**0.572**
		5	0.002	0.000	0.000	0.000	0.156
		6	0.000	0.000	0.000	0.000	0.006

Table 3.15: Simulation results for 4-state Poisson-HMMs: the percentage of times in which order K_0 was estimated by each method (n=100).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
8	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000
		3	**0.736**	**1.000**	**0.652**	**0.978**	0.082
		4	0.258	0.000	0.348	0.022	**0.610**
		5	0.006	0.000	0.000	0.000	0.262
		6	0.000	0.000	0.000	0.000	0.046
9	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000
		3	0.300	**0.894**	**0.764**	**0.990**	0.158
		4	**0.660**	0.106	0.236	0.010	**0.606**
		5	0.040	0.000	0.000	0.000	0.210
		6	0.000	0.000	0.000	0.000	0.026
10	4	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000
		3	0.134	**0.976**	**0.640**	**0.984**	0.070
		4	**0.846**	0.024	0.360	0.016	**0.654**
		5	0.020	0.000	0.000	0.000	0.240
		6	0.000	0.000	0.000	0.000	0.034
		7	0.000	0.000	0.000	0.000	0.002

Table 3.16: Simulation results for 4-state Poisson-HMMs: the percentage of times in which order K_0 was estimated by each method (n=400).

Model	K_0	$\hat{K}_0$	AIC	BIC	AIC_Q	BIC_Q	MSCAD_Q
11	6	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000
		3	0.012	**0.816**	0.002	0.058	0.000
		4	**0.850**	0.184	**0.832**	**0.940**	0.098
		5	0.130	0.000	0.166	0.002	**0.410**
		6	0.008	0.000	0.000	0.000	0.326
		7	0.000	0.000	0.000	0.000	0.138
12	6	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000
		3	0.012	**0.632**	0.000	0.038	0.000
		4	**0.876**	0.368	**0.870**	**0.958**	0.132
		5	0.110	0.000	0.130	0.004	**0.436**
		6	0.002	0.000	0.000	0.000	0.312
		7	0.000	0.000	0.000	0.000	0.100
13	6	1	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000
		3	0.006	0.482	0.002	0.046	0.000
		4	**0.708**	**0.518**	**0.834**	**0.952**	0.098
		5	0.280	0.000	0.162	0.002	**0.452**
		6	0.006	0.000	0.002	0.000	0.272
		7	0.000	0.000	0.000	0.000	0.152

Table 3.17: Simulation results for 6-state Poisson HMMs: the percentage of times in which order K_0 was estimated by each method (n=400).

To summarize, our simulation studies demonstrate that MSCAD based on the quasi-likelihood is an appealing alternative to the information criteria AIC and BIC, based on both the full-model likelihood and the quasi-likelihood. While BIC and BIC_Q performed very well in detecting the true order when it was low, that is, $K_0 = 2$, they had strong tendencies to underfit models of higher order, that is, $K_0 = 3, 4$ and 6. MSCAD_Q, on the other hand, had higher success rates than the other methods when the true order was high, especially when $K_0 = 4$ and $K_0 = 6$. Furthermore, MSCAD_Q is computationally more efficient than the information criteria under consideration as it does not require K separate model fittings. Starting with a large number of states, the method is able to obtain a model of lower order in a single optimization procedure through the clustering and merging of states. Another advantage of MSCAD_Q is that it does not require the estimation of the transition matrix $\mathbb{P}$, as is the case for the information criteria based on the full-model likelihood.

Now let us compare the time taken by MSCAD_Q and the information criteria to complete the analysis of 500 simulated data sets for the most difficult normal HMM with 6 states and sample size $n = 400$. On a typical Unix machine, MSCAD_Q with 5-fold cross-validation to select the tuning parameter took about 13 minutes. The computation of AIC and BIC values for models of order 1 to 7, on the other hand, took over 2 hours. We had used the R code provided in Zucchini and MacDonald (2009) for fitting a stationary HMM by direct numerical maximization of the likelihood function. Their program is for the case of state-dependent Poisson distributions, but can be easily altered for the case of state-dependent normal distributions. While our code for the computation of AIC_Q and BIC_Q values for models of order 1 to 9 took only about 2 minutes to run, one should also consider that time is taken to modify the code after every model fit. The maximum quasi-likelihood estimators were obtained using the EM algorithm. Therefore, it is clear that using a quasi-likelihood instead of the full-model likelihood greatly reduces the computational effort in estimating the order.

3.7 Applications

3.7.1 Poisson HMMs for Movement Counts by Fetal Lambs

We consider a time series of overdispersed count data, originally analyzed in Leroux and Puterman (1992). The data set consists of the numbers of movements by a fetal lamb observed through ultrasound in 240 consecutive 5-second intervals. We plot the data in Figure 3.3.

Let $y_1, y_2, ..., y_{240}$ denote the observations. Leroux and Puterman (1992) suggest fitting

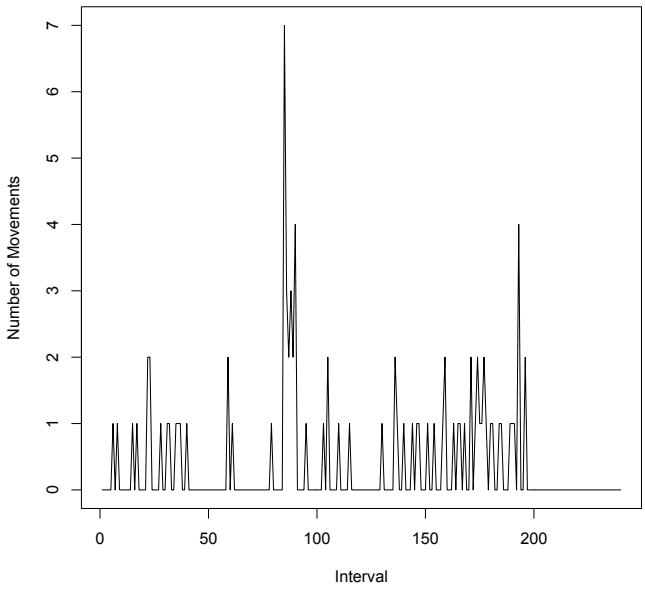

Figure 3.3: Number of movements by a fetal lamb in one of 240 consecutive 5-second intervals.

Poisson HMMs to this data for two reasons. The first reason is to accommodate the overdispersion that is present (the sample variance $s^2 = 0.658$ is larger than the sample mean $\bar{y} = 0.358$). The second reason is to capture the serial dependence in the observations, which can be seen from the sample autocorrelation function found in Figure 3.4.

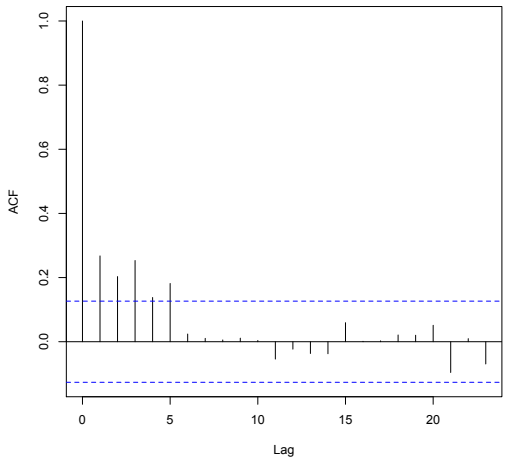

Figure 3.4: Sample ACF for the fetal lamb movement count data.

To select the number of states of the Poisson HMM, Leroux and Puterman (1992) had used the information criteria AIC and BIC based on the full likelihood. They found that

K	Log-likelihood	AIC	BIC	Quasi-log-likelihood	AIC_Q	BIC_Q
1	-201.044	404.087	407.568	-201.044	404.087	407.568
2	-177.519	363.038	**376.960**	-186.990	**379.979**	**390.421**
3	-166.488	**350.976**	382.302	-185.796	381.592	398.995
4	-164.297	360.594	416.285	-185.793	385.587	409.951
5	-164.255	378.511	465.527	-185.790	389.581	420.907

Table 3.18: AIC and BIC values, based on both the full likelihood and quasi-likelihood, for the fetal lamb movement count data.

AIC selects a model of 3 states, whereas BIC selects a model of 2 states. In Table 3.18, we present the AIC and BIC values based on the quasi-likelihood, and for completeness, we also present the AIC and BIC values based on the full likelihood. Note that our log-likelihood values differ from those in Leroux and Puterman (1992) since all terms that do not depend on the parameters of interest were dropped in their computation.

Frühwirth-Schnatter (2006) had used a Bayesian approach to select the number of states. Through a comparison of marginal likelihoods, she had found that the 3-state model was favoured over the 2-state model.

We apply our proposed method to this data set. Starting from $K = 8$ states, we found that MSCAD based on the quasi-likelihood, using CV, AIC and BIC to select the tuning parameter, all favour a 2-state model.

In Table 3.19, we present the maximum likelihood estimates for models of order 2 and 3 obtained using direct numerical maximization of the likelihood. While the 2-state and 3-state models both provide adequate fits, the 2-state model has the additional advantage of being easily interpreted. As pointed out by Leroux and Puterman (1992), the states may correspond to a relaxed state with regular levels of fetal activity and an excited state with higher levels of fetal activity, which are possibly triggered by physical factors such as the development of the central nervous system or empty space within the uterus. The relaxed state has an estimated movement rate $\hat{\lambda}_1 = 0.2564$ and the excited state has an estimated movement rate $\hat{\lambda}_2 = 3.1148$. The fetus occupies an excited state only about 3.5% of the time. Furthermore, it appears that the number of movement counts in any time interval depends strongly on the number of movement counts in the previous time interval. If the fetus is in a relaxed state, it remains in this state with a high probability ($\hat{p}_{11} = 0.989$) and if the fetus is in an excited state, it is more likely to remain in that state in the next time interval ($\hat{p}_{22} = 0.690$).

To assess the marginal properties of both the 2-state and 3-state models, we compare

K	$\hat{\boldsymbol{\lambda}}$	$\hat{\mathbb{P}}$	$\hat{\boldsymbol{\pi}}$
2	(0.256, 3.115)	$\begin{pmatrix} 0.989 & 0.011 \\ 0.310 & 0.690 \end{pmatrix}$	$(0.965, 0.035)$
3	(0.041, 0.495, 3.413)	$\begin{pmatrix} 0.950 & 0.040 & 0.010 \\ 0.041 & 0.959 & 0 \\ 0.188 & 0 & 0.812 \end{pmatrix}$	$(0.490, 0.483, 0.027)$

Table 3.19: Parameter estimates for Poisson HMMs of order 2 and 3 fitted to the movement count data in fetal lambs.

the observed numbers of movement counts to those expected under each of the models, which are displayed in Table 3.20. As a measure of the goodness-of-fit, we use Pearson's chi-squared test statistic

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed frequency, E_i is the expected frequency and $N = 8$. The test statistic is asymptotically chi-squared distributed with $N - p - 1$ degrees of freedom, where p is the number of parameters estimated from the data. We find $\chi^2 = 7.796$ and $\chi^2 = 4.965$ for models of order 2 and 3, respectively. Using the 0.05 level of significance, the critical values are $\chi^2_{0.95,4} = 9.488$ for the 2-state model and $\chi^2_{0.95,2} = 5.991$ for the 3-state model. Since in both cases the test statistic is smaller than the critical value, it appears that both models fit the data well.

# of Movements	Observed Frequency	Expected Frequency	
		$K = 2$	$K = 3$
0	182	179.587	183.911
1	41	47.108	40.292
2	12	7.702	9.969
3	2	2.385	2.814
4	2	1.497	1.358
5	0	0.914	0.823
6	0	0.474	0.460
7	1	0.211	0.224

Table 3.20: Observed numbers of movement counts, compared with those expected under models of order 2 and 3.

3.7.2 Normal HMMs for Waiting Times of the Old Faithful Geyser

We consider a time series relating to eruptions of the Old Faithful geyser in Yellowstone National Park in the U.S. state of Wyoming. The data set, which was originally presented in Azzalini and Bowman (1990), consists of 299 observations of continuous measurement from August 1st to August 15th, 1985. The observations are times between the starts of successive eruptions. From the sample autocorrelation function displayed in Figure 3.5, we see that there is strong serial dependence in the behaviour of the geyser.

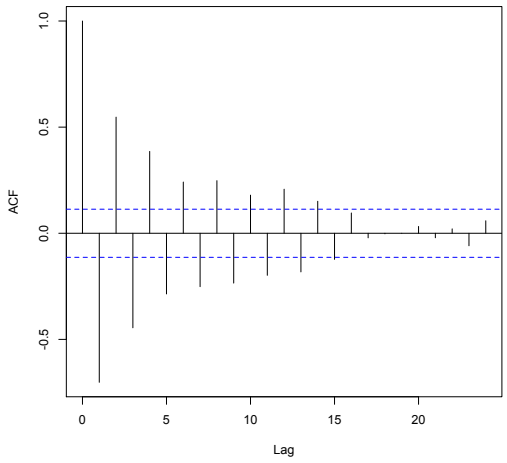

Figure 3.5: Sample ACF for the waiting times of the Old Faithful geyser.

This data set has been previously analyzed by Zucchini and MacDonald (2009), who fit a series of normal HMMs to this data set with unequal variances. Since the variances for their final 3-state model do not differ substantially across states, we decided to fit a series of normal HMMs with equal variances and consider the problem of estimating the number of states. In Table 3.21, we compare models of order 1 to 5 on the basis of AIC and BIC, based on both the full-model likelihood and the quasi-likelihood. We see that AIC selects a 4-state model, while its quasi-likelihood counterpart selects a 3-state model. We also see that BIC selects a 3-state model, while the quasi-likelihood counterpart selects a 2-state model.

We then apply our proposed method, starting with an upper bound $K = 15$. For the tuning parameter $\gamma_n/\sqrt{n}$, we considered the range [0.40, 1.75]. MSCAD$_Q$, using CV and BIC to select the tuning parameter, decided on a 3-state model. With AIC as the method for tuning parameter selection, however, a 6-state model was chosen by MSCAD$_Q$. This example highlights the importance of choosing an appropriate tuning parameter.

K	Log-likelihood	AIC	BIC	Quasi-log-likelihood	AIC_Q	BIC_Q
1	-1210.488	2424.977	2432.378	-1210.488	2424.977	2432.378
2	-1099.632	2209.264	2227.766	-1161.709	2331.419	**2346.220**
3	-1053.391	2126.783	**2163.787**	-1158.522	**2329.044**	2351.246
4	-1046.130	**2126.261**	2189.169	-1157.288	2330.575	2360.179
5	-1034.787	2121.574	2217.786	-1157.288	2330.575	2360.179

Table 3.21: AIC and BIC values, based on both the full likelihood and quasi-likelihood, for the Old Faithful waiting times.

The parameter estimates for models of order 2, 3 and 4 are displayed in Table 3.22. They were obtained using direct numerical maximization of the likelihood. In the left panel of Figure 3.6, we plot the fitted densities of models with 2, 3 and 4 states on the histogram of waiting times. We also plot the density of the fitted normal HMMs with unequal variances in the right panel of Figure 3.6. From the left panel of Figure 3.6, it appears that the 4-state model does not result in a substantial improvement in fit over the 3-state model. Our preference is thus with the normal HMM of order 3. We also see that the models with unequal variances provide slightly better fits than the models with equal variances. This is expected since, in general, the addition of more parameters should improve the overall fit of the model.

K	2	3	4
$\hat{\boldsymbol{\mu}}$	(57.206, 81.921)	(54.764, 75.414, 85.091)	(53.168, 62.544, 75.978, 85.091)
$\hat{\sigma}$	6.867	5.287	4.933
$\hat{\mathbb{P}}$	$\begin{pmatrix} 0.000 & 1.000 \\ 0.638 & 0.362 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 0.000 & 1.000 \\ 0.251 & 0.635 & 0.114 \\ 0.667 & 0.296 & 0.037 \end{pmatrix}$	$\begin{pmatrix} 0.000 & 0.000 & 0.000 & 1.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \\ 0.105 & 0.235 & 0.621 & 0.039 \\ 0.605 & 0.072 & 0.280 & 0.043 \end{pmatrix}$
$\hat{\boldsymbol{\pi}}$	(0.390, 0.610)	(0.325, 0.302, 0.373)	(0.256, 0.092, 0.277, 0.375)

Table 3.22: Parameter estimates for normal HMMs of order 3 and 4 fitted to the waiting times of the Old Faithful geyser.

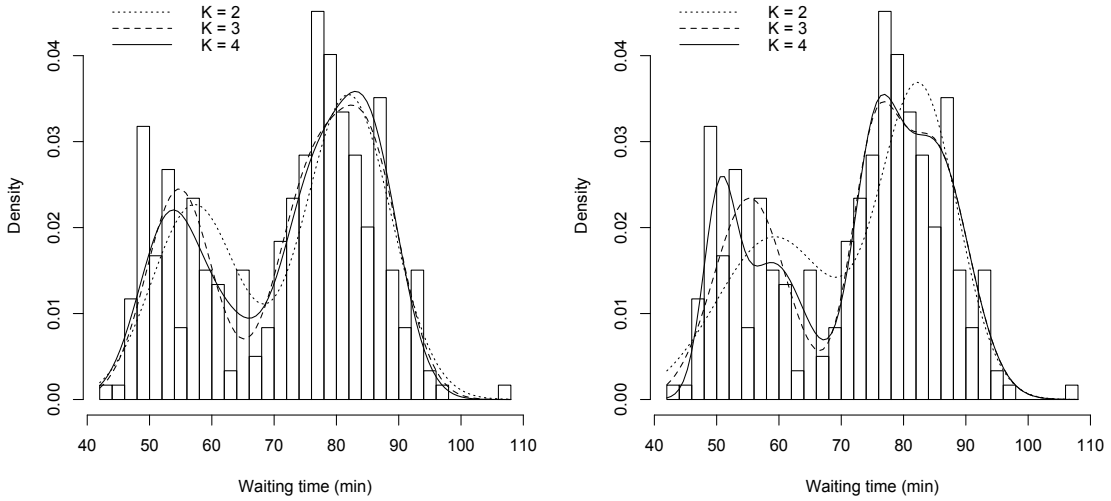

Figure 3.6: Normal HMMs of order 2, 3 and 4 with equal variances (left) and unequal variances (right) fitted to the waiting times of the Old Faithful geyser.

3.8 Discussion

In this chapter, we considered the problem of estimating the number of states in a stationary HMM. We reviewed existing order estimation procedures, such as AIC and BIC, based on both the full-model likelihood and the quasi-likelihood. We then proposed a new method to deal with this order estimation problem.

Theoretically we showed that the maximum penalized quasi-likelihood estimator $\hat{G}_n$ is a consistent estimator of the true mixing distribution G_0. However, this result does not give us consistency in estimating the true order K_0 since $\hat{G}_n$ may have more than K_0 support points. It remains to show that $\hat{G}_n$ is consistent in estimating K_0. We will continue to work on this problem following the completion of this thesis.

The implementation of this method was then discussed. We presented a revised EM algorithm for performing the maximization of the penalized quasi-likelihood. We also introduced likelihood-based cross-validation for selecting the tuning parameter used in the SCAD penalty.

We next evaluated the performance of the proposed method through simulation, comparing it to the performance of the information criteria, based on both the full-model likelihood and the quasi-likelihood. Our simulation results indicate that when K_0 is small, MSCAD_Q is on par with these methods, but when K_0 is large, its success rates in detecting the true order are generally higher than those of the information criteria.

We also demonstrated the method by analyzing two well-known data sets in the HMM literature.

One of the main advantages of this method is that it only requires the fitting of one model. Starting with a HMM with a large number of states, the two penalty functions work simultaneously to cluster and merge states so that a model with the proper order is obtained in a single optimization procedure. This is not the case for AIC and BIC, which must fit all the candidate models in order to select the best one.

Chapter 4

Conclusion

The focus of this thesis was to consider the problem of order estimation in stationary hidden Markov models. We proposed a new method for order estimation that is based on the penalization of a so-called quasi-likelihood, which is constructed from the marginal mixture distributions of the HMM observations. We now summarize this thesis and look towards future work, both on the proposed method itself as well as on possible extensions of this work.

In Chapter 2, we formally introduced HMMs and discussed the estimation of HMM parameters in the case where the order is known. We focused on the the most common approach in the HMM literature for estimating model parameters, which is to perform maximum likelihood estimation. In particular, we compared two methods for finding the maximum likelihood estimators of HMM parameters, namely the EM algorithm and direct numerical maximization.

In Chapter 3, we presented our penalized quasi-likelihood method, MSCAD$_Q$, for estimating the number of hidden states. We investigated some of the asymptotic properties of the proposed method and assessed its performance via simulation. While MSCAD$_Q$ was found to perform well in simulation, there are a number of issues relating to this method that require further investigation. Firstly, the asymptotic theory of this method is incomplete. While the maximum penalized quasi-likelihood estimator was shown to be a consistent estimator of the true mixing distribution, it remains to show that the MPQLE is consistent in estimating the true order. Secondly, the method assumes that there is an optimal range of the tuning parameter for detecting the true order. How to select a suitable tuning parameter used in the SCAD penalty in a computationally efficient manner is still unclear. Thirdly, the optimization of the penalized quasi-likelihood is not trivial due to the singularity of the SCAD penalty at the origin. While the revised EM algorithm, presented in Chapter 3, is relatively straightforward to implement, it does not guarantee convergence to the global maximum of the objective function. Thus, the

optimization of the penalized quasi-likelihood is worthy of further study.

One possible extension of this method would be to the case where the state-dependent parameters $\boldsymbol{\theta}_k$ are multi-dimensional, which is the subject of a future research project. Our proposed method is limited to the one-dimensional setting.

Appendix A: Proofs

In Proposition 2.3.1, we will show that the forward probabilities $\alpha_t(j)$ are indeed probabilities, but first we need the following lemma.

Lemma 2.3.1. For $t = 1, 2, ..., n-1$ and $i, j = 1, 2, ..., K$,

$$P(Y_1 = y_1, ..., Y_{t+1} = y_{t+1}, Z_t = i, Z_{t+1} = j) =$$

$$P(Z_{t+1} = j \mid Z_t = i)P(Y_{t+1} = y_{t+1} \mid Z_{t+1} = j)P(Y_1 = y_1, ..., Y_t = y_t, Z_t = i).$$

Proof. Note that

$$\begin{aligned} P(Y_1, ..., Y_{t+1}, Z_1, ..., Z_{t+1}) &= P(Z_1)\prod_{s=2}^{t+1} P(Z_s \mid Z_{s-1})\prod_{s=1}^{t+1} P(Y_s \mid Z_s) \\ &= P(Z_{t+1} \mid Z_t)P(Y_{t+1} \mid Z_{t+1})P(Y_1, ..., Y_t, Z_1, ..., Z_t). \end{aligned}$$

Now summing over $Z_1, ..., Z_{t-1}$, the result follows. □

Proposition 2.3.1. For $t = 1, 2, ..., n$ and $j = 1, 2, ..., K$,

$$\alpha_t(j) = P(Y_1 = y_1, ..., Y_t = y_t, Z_t = j).$$

Proof. By induction. For t = 1:

$$\begin{aligned} \alpha_1(j) &= \pi_j f(y_1; \theta_j) \\ &= P(Z_1 = j)P(Y_1 = y_1 \mid Z_1 = j) \\ &= P(Y_1 = y_1, Z_1 = j). \end{aligned}$$

Now suppose the claim holds for some $t \geq 1$. Then

$$\begin{aligned}
\alpha_{t+1}(j) &= \left\{ \sum_{i=1}^{K} \alpha_t(i)\, p_{ij} \right\} f(y_{t+1}; \theta_j) \\
&= \sum_{i=1}^{K} P(Y_1 = y_1, ..., Y_t = y_t, Z_t = i) P(Z_{t+1} = j \mid Z_t = i) P(Y_{t+1} = y_{t+1} \mid Z_{t+1} = j) \\
&= \sum_{i=1}^{K} P(Y_1 = y_1, ..., Y_{t+1} = y_{t+1}, Z_t = i, Z_{t+1} = j) \\
&= P(Y_1 = y_1, ..., Y_{t+1} = y_{t+1}, Z_{t+1} = j),
\end{aligned}$$

where the second to last equality follows from Lemma 2.3.1. □

In Proposition 2.3.2, we also verify that the backward probabilities are indeed probabilities. First, we must present the following lemma.

Lemma 2.3.2. For $t = 1, 2, ..., n-1$,

$$\begin{aligned}
&(1) P(Y_{t+1}, ..., Y_n \mid Z_{t+1}) = P(Y_{t+1} \mid Z_{t+1}) P(Y_{t+2}, ..., Y_n \mid Z_{t+1}) \\
&(2) P(Y_{t+1}, ..., Y_n \mid Z_{t+1}) = P(Y_{t+1}, ..., Y_n \mid Z_t, Z_{t+1}).
\end{aligned}$$

Proof of (1). Note that

$$\begin{aligned}
&P(Y_{t+1}, ..., Y_n, Z_{t+2}, ..., Z_n \mid Z_{t+1}) P(Z_{t+1}) \\
&\qquad = P(Y_{t+1}, ..., Y_n \mid Z_{t+1}, ..., Z_n) P(Z_{t+1}, ..., Z_n) \\
&\qquad = P(Y_{t+1} \mid Z_{t+1}) \left\{ P(Z_{t+1}) \prod_{s=t+2}^{n} P(Z_s \mid Z_{s-1}) \prod_{s=t+2}^{n} P(Y_s \mid Z_s) \right\} \\
&\qquad = P(Y_{t+1} \mid Z_{t+1}) P(Y_{t+2}, ..., Y_n, Z_{t+1}, ..., Z_n)
\end{aligned}$$

and summing over $Z_{t+2}, ..., Z_n$ and dividing by $P(Z_{t+1})$, we obtain the result. □

Proof of (2). We have that

$$\begin{aligned}
P(Y_{t+1},...,Y_n \mid Z_t, Z_{t+1}) &= \frac{P(Y_{t+1},...,Y_n, Z_t, Z_{t+1})}{P(Z_t, Z_{t+1})} \\
&= \frac{1}{P(Z_t, Z_{t+1})} \sum_{Z_{t+2},...,Z_n} P(Y_{t+1},...,Y_n, Z_t,...,Z_n) \\
&= \frac{1}{P(Z_t, Z_{t+1})} \sum_{Z_{t+2},...,Z_n} P(Z_t,...,Z_n) P(Y_{t+1},...,Y_n \mid Z_t,...,Z_n) \\
&= \frac{P(Z_t)}{P(Z_t, Z_{t+1})} \sum_{Z_{t+2},...,Z_n} \prod_{s=t+1}^{n} P(Z_s \mid Z_{s-1}) \prod_{s=t+1}^{n} P(Y_s \mid Z_s) \\
&= \sum_{Z_{t+2},...,Z_n} \prod_{s=t+2}^{n} P(Z_s \mid Z_{s-1}) \prod_{s=t+1}^{n} P(Y_s \mid Z_s) \\
&= \frac{1}{P(Z_{t+1})} \sum_{Z_{t+2},...,Z_n} P(Y_{t+1},...,Y_n, Z_{t+1},...,Z_n) \\
&= P(Y_{t+1},...,Y_n \mid Z_{t+1}).
\end{aligned}$$

□

Proposition 2.3.2. For $t = 1, 2, ..., n-1$,

$$\beta_t(i) = P(Y_{t+1} = y_{t+1}, ..., Y_n = y_n \mid Z_t = i).$$

Proof. By induction. For $t = n - 1$:

$$\begin{aligned}
\beta_{n-1}(i) &= \sum_{j=1}^{K} p_{ij} f(y_n; \theta_j) \beta_n(j) \\
&= \sum_{j=1}^{K} P(Z_n = j \mid Z_{n-1} = i) P(Y_n = y_n \mid Z_n = j) \\
&= \sum_{j=1}^{K} P(Z_n = j \mid Z_{n-1} = i) P(Y_n = y_n \mid Z_n = j, Z_{n-1} = i) \\
&= \sum_{j=1}^{K} P(Y_n = y_n, Z_n = j \mid Z_{n-1} = i) \\
&= P(Y_n = y_n \mid Z_{n-1} = i),
\end{aligned}$$

where the third equality follows from Lemma 2.3.2 (2).

Now assume the result holds for some $t+1$. Then we have that

$$\begin{aligned}
\beta_t(i) &= \sum_{j=1}^{K} p_{ij} f(y_{t+1}; \theta_j) \beta_{t+1}(j) \\
&= \sum_{j=1}^{K} P(Z_{t+1} = j \mid Z_t = i) P(Y_{t+1} \mid Z_{t+1} = j) P(Y_{t+2}, ..., Y_n \mid Z_{t+1} = j) \\
&= \sum_{j=1}^{K} P(Z_{t+1} = j \mid Z_t = i) P(Y_{t+1}, ..., Y_n \mid Z_{t+1} = j) \\
&= \sum_{j=1}^{K} P(Z_{t+1} = j \mid Z_t = i) P(Y_{t+1}, ..., Y_n \mid Z_{t+1} = j, Z_t = i) \\
&= \sum_{j=1}^{K} P(Y_{t+1}, ..., Y_n, Z_{t+1} = j \mid Z_t = i) = P(Y_{t+1}, ..., Y_n \mid Z_t = i),
\end{aligned}$$

where the third and fourth equalities follow from Lemma 2.3.2 (1) and Lemma 2.3.2 (2), respectively. □

Lemma 2.3.3. For $t = 1, 2, ..., n-1$,

$$P(Y_1, ..., Y_n \mid Z_t) = P(Y_1, ..., Y_t \mid Z_t) P(Y_{t+1}, ..., Y_n \mid Z_t).$$

Proof. First note that

$$\begin{aligned}
P(Y_1, ..., Y_n, Z_1, ..., Z_n) &= P(Z_1) \prod_{s=2}^{n} P(Z_s \mid Z_{s-1}) \prod_{s=1}^{n} P(Y_s \mid Z_s) \\
&= P(Y_1, ..., Y_t, Z_1, ..., Z_t) \prod_{s=t+1}^{n} P(Z_s \mid Z_{s-1}) \prod_{s=t+1}^{n} P(Y_s \mid Z_s) \\
&= P(Y_1, ..., Y_t, Z_1, ..., Z_t) \frac{P(Y_{t+1}, ..., Y_n, Z_t, ..., Z_n)}{P(Z_t)}.
\end{aligned}$$

Now summing over $Z_1, ..., Z_{t-1}$ and $Z_{t+1}, ..., Z_n$, we obtain

$$P(Y_1, ..., Y_n, Z_t) = P(Y_1, ..., Y_t, Z_t) \frac{P(Y_{t+1}, ..., Y_n, Z_t)}{P(Z_t)}$$

and dividing by $P(Z_t)$, the result follows. □

Proposition 2.3.3. For $t = 2, ..., n$ and $i, j = 1, 2, ..., K$,

(1) $P(Z_t = i \mid Y_1, ..., Y_n) = \dfrac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{K} \alpha_t(i)\beta_t(i)}$ and

(2) $P(Z_{t-1} = i, Z_t = j \mid Y_1, ..., Y_n) = \dfrac{\alpha_{t-1}(i) p_{ij} f(y_t; \theta_j) \beta_t(j)}{\sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_{t-1}(i) p_{ij} f(y_t; \theta_j) \beta_t(j)}$.

Proof of (1). From Lemma 2.3.3, we obtained the result

$$P(Y_1, ..., Y_n \mid Z_t = i) = P(Y_1, ..., Y_t \mid Z_t = i)P(Y_{t+1}, ..., Y_n \mid Z_t = i)$$

and thus multiplying both sides by $P(Z_t = i)$, we have that

$$P(Y_1, ..., Y_n, Z_t = i) = P(Y_1, ..., Y_t, Z_t = i)P(Y_{t+1}, ..., Y_n \mid Z_t = i) = \alpha_t(i)\beta_t(i).$$

Dividing $P(Y_1, ..., Y_n, Z_t = i)$ by $P(Y_1, ..., Y_n) = \sum_{i=1}^{K} \alpha_t(i)\beta_t(i)$, the result follows.

□

Proof of (2). First note that

$$\begin{aligned}
&P(Y_1, ..., Y_n, Z_1, ..., Z_n) \\
&= P(Y_1, ..., Y_{t-1}, Z_1, ..., Z_{t-1})P(Z_t \mid Z_{t-1}) \prod_{s=t+1}^{n} P(Z_s \mid Z_{s-1}) \prod_{s=t}^{n} P(Y_s \mid Z_s) \\
&= P(Y_1, ..., Y_{t-1}, Z_1, ..., Z_{t-1})P(Z_t \mid Z_{t-1}) \frac{P(Y_t, ..., Y_n, Z_t, ..., Z_n)}{P(Z_t)} \\
&= P(Y_1, ..., Y_{t-1}, Z_1, ..., Z_{t-1})P(Z_t \mid Z_{t-1})P(Y_t, ..., Y_n, Z_{t+1}, ..., Z_n \mid Z_t)
\end{aligned}$$

Now summing over $Z_1, ..., Z_{t-2}$ and $Z_{t+1}, ..., Z_n$, we obtain

$$\begin{aligned}
P(Y_1, ..., Y_n, Z_{t-1}, Z_t) &= P(Y_1, ..., Y_{t-1}, Z_{t-1})P(Z_t \mid Z_{t-1})P(Y_t, ..., Y_n \mid Z_t) \\
&= \alpha_{t-1}(i)p_{ij}f(y_t; \theta_j)\beta_t(j)
\end{aligned}$$

since $P(Y_t, ..., Y_n \mid Z_t) = P(Y_t \mid Z_t)P(Y_{t+1}, ..., Y_n \mid Z_t)$ from Lemma 2.3.2 (1). Dividing $P(Y_1, ..., Y_n, Z_{t-1}, Z_t)$ by $P(Y_1, ..., Y_n) = \sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_{t-1}(i)p_{ij}f(y_t; \theta_j)\beta_t(j)$, the result follows.

□

Proposition 2.4.1: The initial distribution $\boldsymbol{\pi}$ is a stationary distribution if and only if $\boldsymbol{\pi}(I_K - \mathbb{P} + O) = \mathbf{1}$, where I_k is the $K \times K$ identity matrix, $\mathbb{P}$ is the transition matrix, O is the $K \times K$ matrix of ones and $\mathbf{1}$ is the K-dimensional row vector of ones.

Proof. Suppose that $\boldsymbol{\pi}(I_K - \mathbb{P} + O) = \mathbf{1}$. Then we have that

$$\boldsymbol{\pi}(I_K + O) - \mathbf{1} = \boldsymbol{\pi}\mathbb{P} \tag{4.1}$$

and evaluating the left-hand side of Equation (4.1) gives us

$$(\pi_1\pi_2 \ldots \pi_K) \begin{pmatrix} 2 & 1 & ... & 1 \\ 1 & 2 & ... & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & ... & 2 \end{pmatrix} - (11 \ldots 1)$$

$= (2\pi_1 + \pi_2 + ... + \pi_K, \pi_1 + 2\pi_2 + ... + \pi_K, ..., \pi_1 + \pi_2 + ... + 2\pi_K) - (1, 1, \ldots, 1)$

$= (\pi_1 + 1, \pi_2 + 1, ..., \pi_K + 1) - (1, 1, \ldots, 1)$ since $\sum_{j=1}^{K} \pi_j = 1$

$= (\pi_1, \pi_2, ..., \pi_K)$.

Thus, equating the left-hand side and right-hand side of Equation (4.1) gives us $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$.

Now suppose that $\boldsymbol{\pi}$ is a stationary distribution, that is, $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$, then reversing the argument as follows

$$\boldsymbol{\pi}\mathbb{P} = \boldsymbol{\pi} = \boldsymbol{\pi}(I_K + O) - \mathbf{1}$$

will give us that $\boldsymbol{\pi}(I_K - \mathbb{P} + O) = \mathbf{1}$. □

Appendix B: Regularity Conditions

In what follows, the expectations are under the true mixing distribution G_0. Assumptions 3 to 6 correspond to those in Poskitt and Zhang (2005), who show that the sequence $\frac{1}{n}\ell_n^Q(G)$ converges to $\mathbb{E}[\log f(Y;G)]$ almost surely and uniformly over the compact space of G.

Assumption 1. The parameter space Θ is compact. Furthermore, the following two conditions hold.

(i) $\mathbb{E}[|\log f(Y_1;\theta)|] < \infty \ \forall \theta \in \Theta$.

(ii) There exists $\epsilon > 0$ such that for each $\theta \in \Theta$, $f(y;\theta)$ is measurable and $\mathbb{E}[|\log f(Y_1;\theta,\epsilon)|] < \infty$, where $f(y;\theta,\epsilon) = 1 + \sup_{|\theta-\theta'|\leq\epsilon} f(y;\theta')$.

Assumption 2. The family $\{f(y;\theta);\theta \in \Theta\}$ is strongly identifiable in the sense that for K distinct $\theta_1, \theta_2, ..., \theta_K$,

$$\sum_{j=1}^{K} \{a_j f(y;\theta_j) + b_j f'(y;\theta_j) + c_j f''(y;\theta_j)\} = 0$$

for all y implies that $a_j = b_j = c_j = 0$ for $j = 1, 2, ..., K$.

Assumption 3. The density $f(y;\theta)$ is differentiable with respect to $\theta \in \Theta$ and y, and three times continuously differentiable with respect to θ.

Assumption 4. There exists a continuous function $h(y)$ such that $f(y;\theta) \leq h(y)$ and $\mathbb{E}[|\log h(y)|] < \infty$.

Assumption 5. First- and second-order partial derivatives of $\log f(y;\theta)$ satisfy

$$\left|\frac{\partial \log f(y;\theta)}{\partial \theta_u}\right| < h_u(y), \quad \left|\frac{\partial \log f(y;\theta)}{\partial \theta_u \partial \theta_v}\right| < h_{u,v}(y),$$

where $\mathbb{E}[h(y)] < \infty$ and $\mathbb{E}[h_{u,v}(y)] < \infty$.

Assumption 6. There exists $\delta > 0$ such that $\int \|y\|^{2+\delta} f(y;\theta)\, dy < \infty$ for all $\theta \in \Theta$, where $\|\cdot\|$ is Euclidean distance.

Further, in addition to the identifiability of the marginal finite mixture model, Lindgren (1978) had shown that under the following conditions the maximum quasi-likelihood estimator is consistent and asymptotically normal.

Assumption 7: There exists a neighbourhood S of θ_0 such that

(i) $\mathbb{E}\left[\sup_{\theta\in S} |\log f(Y_1;\theta)|\right] < \infty$,

(ii) $\int \sup_{\theta\in S} \left|\frac{\partial f(y;\theta)}{\partial \theta_i}\right| \mathrm{d}y < \infty$,

(iii) $\int \sup_{\theta\in S} \left|\frac{\partial^2 f(y;\theta)}{\partial \theta_i \partial \theta_j}\right| \mathrm{d}y < \infty$,

(iv) $\mathbb{E}\left[\sup_{\theta\in S} \left|\frac{\partial^3 f(y;\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}\right|\right] < \infty$,

(v) For some $\delta > 0$, $\mathbb{E}\left[\left|\frac{\partial \log f(Y_1;\theta_0)}{\partial \theta_i}\right|^{2+\delta}\right] < \infty$.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest: Academiai Kiado.

[2] Albert, P.S. (1991). A Two-State Markov Mixture Model for a Time Series of Epileptic Seizure Counts. *Biometrics*, **47**, 1371-1381.

[3] Altman, R.M. and Petkau, A.J. (2005). Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine*, **24**, 2335–2344.

[4] Azzalini, A. and Bowman, A.W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, **39**, 357–365.

[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, *The 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. *The Annals of Statistics*, **26**, 1614-1635.

[7] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models.* Springer, New York.

[8] Chen, H., Chen, J. and Kalbfleisch, J.D. (2004). Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society: Series B*, **66**, 95-115.

[9] Chen, H., Chen, J. and Kalbfleisch, J.D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B*, **63**, 19-29.

[10] Chen, J. and Kalbfleisch, J.D. (1996). Penalized minimum-distance estimates in finite mixture models. *The Canadian Journal of Statistics*, **24**, 167-175.

[11] Chen, J. and Khalili, A. (2008). Order Selection in Finite Mixture Models With a Nonsmooth Penalty. *Journal of the American Statistical Association*, **103**, 1674-1683.

[12] Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313– 1321.

[13] Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, **51**, 79-94.

[14] Craven, P. and Wahba, G. (1979). Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik*, **31**, 377–403.

[15] Dannemann, J. and Holzmann, H. (2008). Testing for two states in a hidden Markov model. *The Canadian Journal of Statistics*, **4**, 505-520.

[16] Dempster, A.P., Laird, N.M. and Rubin D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

[17] Fan, J. and Li, R. (2001). Variable Selection via Noncave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

[18] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, New York.

[19] Gassiat, E. and Keribin, C. (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM: Probability and Statistics*, **4**, 25–52.

[20] Holzmann, H. and Schwaiger, F. (2012). Testing for the number of states in regime-switching models. *Preprint.*

[21] Ishwaran, H., James, L.F., and Sun, J. (2001). Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions. *Journal of the American Statistical Association*, **96**, 1316-1332.

[22] Keribin, C. (2000). Consistent Estimation of the Order of Mixture Models. *Sankhyā: The Indian Journal of Statistics, Series A*, **62**, 49-66.

[23] Khalili, A. (2005). *Order Selection in Classical Finite Mixture Models, and Variable Selection and Inference in Finite Mixture of Regression Models*. Ph.D. thesis, Department of Statistics and Actuarial Sciences, University of Waterloo, Ontario.

[24] Leroux, B.G. (1992a). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and Their Applications*, **40**, 127-143.

[25] Leroux, B.G. (1992b). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20**, 1350-1360.

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

[27] Levinson, S., Rabiner, R., and M. Sondhi. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Systems Technical Journal*, **62**, 1035-1074.

[28] Lindgren, G. (1978). Markov Regime Models for Mixed Distributions and Switching Regressions. *Scandinavian Journal of Statistics*, **5**, 81-91.

[29] MacKay, R.J. (2002). Estimating the order of a hidden markov model. *The Canadian Journal of Statistics*, **30**, 573-589.

[30] MacKay, R.J. (2003). *Hidden Markov Models: Multiple Processes and Model Selection.* Ph.D. thesis, Department of Statistics, University of British Columbia, Vancouver.

[31] Poskitt, D.S. and Zhang, J. (2005). Estimating Components in Finite Mixtures and Hidden Markov Models *Australian and New Zealand Journal of Statistics*, **47**, 269-286.

[32] Rabiner, L.R. and Juang, B.H. (1986). An introduction to hidden Markov models. IEEE ASSP Magazine, **3**, 4–16.

[33] Rydén, T. (1995). Estimating the Order of Hidden Markov Models. *Statistics*, **26**, 345-354.

[34] Rydén, T., Teräsvirta, T. and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden markov model. *Journal of Applied Econometrics*, **13**, 217–244.

[35] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461–464.

[36] Stone, M. (1974). Cross-validation choice and assessment of statistical predictions (with Discussion). *Journal of the Royal Statistical Society: Series B*, **36**, 111–147.

[37] Turner, R. (2008). Direct maximization of the likelihood of a hidden Markov model. *Computational Statistics and Data Analysis*, **52**, 4147-4160.

[38] Wilks, S.S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, **9**, 60–62.

[39] Zou H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics*, **36**, 1509–1533.

[40] Zucchini, W. and MacDonald, I.L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R.* Boca Raton, FL: Chapman & Hall/CRC.