MARKOV REGIME MODELS FOR MIXED DISTRIBUTIONS
AND SWITCHING REGRESSIONS

by
Georg Lindgren

American Mathematical Society 1970 subject classification: Primary 62M05;

Key words and phrases: Incomplete data, Markov chains, maximum likelihood,
mixtures, partial observation, robustness, switching regression.
ABSTRACT

The paper deals with estimation of unknown parameters in a finite mixture of distributions or populations. The basic assumption is the existence of unobservable regime variables $X_t$, $t = 1, \ldots, T$ which select the distribution to be observed for each $t$. If the $X$-variables are independent this is the classical mixed distribution problem.

Here we suppose that $X_t$, $t = 1, \ldots, T$ is a stationary Markov process, which means that successive observations either tend to come from different or from similar populations. Such a dependence is reasonable if the observations are obtained sequentially in time.

The ML-estimator of marginal parameters and transition probabilities are derived using a maximization technique due to Baum. In a simulation study the ML-estimators are compared to the estimators derived under the assumption that the observations are independent, and it is shown that when there actually is a dependence, then the estimators based on the full Markov model are superior. A FORTRAN-routine for estimation is presented.
1. Introduction

This paper deals with the statistical estimation of unknown parameters in a finite mixture

\[ f(y) = \sum_{j=1}^{r} \pi_j f_j(y) \]

of distributions with densities \( f_j, j = 1, \ldots, r \) from a sequence of observations \( y_t, t = 1, \ldots, T \) from the density (1). Such mixtures often appear as the result of a physical mixture of populations with different characteristics, e.g. different sexes or races in biology, different polymer molecules in chemistry, or different demand and supply markets in econometrics.

The basic assumption made here is the existence of regime variables \( X_t, t = 1, \ldots, T \) which for each \( t \) select one of the distributions \( f_j \) which is then observed, i.e. the conditional density of \( y_t \) given that \( X_t = j \) is equal to \( f_j \). If the \( X \)-variables are independent with \( P(X_t = j) = \pi_j \) we have the classical mixed distribution problem with independent observations, which has been tackled by several authors, starting with Karl Pearson (1894) and Charlier (1906). As is well known this is a very difficult estimation problem, which requires large samples and good initial guesses in order to be successfully solved.

The estimation procedures for mixtures of normal distributions are of mainly three types: moment estimators as those of Pearson and Charlier and more recently Cohen (1967), graphical methods, e.g. Hald (1952) and Bhattacharya (1967), and, most promising, maximum likelihood estimators as developed by Rao (1948), Hasselblad (1966), Day (1969), and Wolfe (1970). For a rather complete review of normal decomposition, see also Holgersson & Jorner (1976).
General principles for separation have been developed by Hasselblad (1969), Orchard & Woodbury (1972), and Sundberg (1974), who all employ a missing information technique in connexion with exponential families. Behboodian (1975) contains some more references.

All the abovementioned authors have been concerned with independent observations, i.e. the regime variables $X_t$ are independent and un-observable.

Some effort has also been made to utilize additional information in order to improve the estimates. Hosmer (1973) and Dick & Bowden (1973) showed that if one has an additional small sample which is known to come from one specific of the components then the estimates are considerably improved. Formulated in terms of the regime variables $X_t$ this means that we know, for some $t$, the exact value of $X_t$, while for most $t$-values the regimes are unknown and still independent.

In this paper we shall use the regime variables more systematically in that we suppose that $X_t, t = 1, \ldots, T$ is a realization of a stochastic process with known or partially known distribution. In other words, we postulate some type of dependence between $X_s$ and $X_t$ for $s \neq t$, and then use this dependence in order to improve the estimates of parameters in the marginal distribution (1).

Such a dependence is often reasonable, especially if the observations $y_t, t = 1, \ldots, T$ are obtained sequentially in time. Then it is often natural to assume that adjacent observations will either tend to come from the same or tend to come from different populations.

Closely related to mixed distributions is what is sometimes called switching regressions, i.e. one has a sequence of observations $y_t, t = 1, \ldots, T$ where each $y_t$ is generated by one of $r$ possible regression expressions,

$$y_t = \beta^{(j)} u_t^{(j)} + \epsilon_t^{(j)}.$$
where the residuals $\varepsilon_t^{(j)}$ are independent with $E(\varepsilon_t^{(j)}) = 0$, $V(\varepsilon_t^{(j)}) = \sigma_j^2$, $j = 1, \ldots, r$, $t = 1, \ldots, T$.

Formulated by the aid of the regimes $X_t$, $t = 1, \ldots, T$ one has that

$$y_t | X_t = j \overset{\text{L}}{=} \beta_{(j)u_t}^{(j)} + \varepsilon_t^{(j)}.$$  

Such models are of frequent use in econometric theory for unstable markets, and there have been several attempts to determine reasonable regime processes. Fair & Kelejian (1974) let the outcome of the different regressions determine which equation will actually be observed, which however introduces an intractable dependence between the regime $X_t$ and the observation $y_t$. Goldfeld & Quandt (1973) pretend to use a Markov regime process for similar problems. We will return to such applications in Section 6.

The use of a regime process in order to describe parametric instability can be a useful alternative to different adaptive estimation schemes, such as discounted least squares, direct and discounted Kalman estimation and other techniques. The regime procedure can be used when there is a finite, and small, number of regimes which alternate in governing the parameters, which will vary from time to time but always take values chosen from a finite set of possible values.

## 2. Markov regime models

Let $\{X_t\}_{t=1}^{\infty}$ be a finite state Markov chain with state space $\{1, \ldots, r\}$, stationary transition probabilities

$$\pi_{jk} = P(X_{t+1} = k | X_t = j)$$

and stationary distribution

$$\pi_j = P(X_t = j).$$
Let $f_j, j = 1, \ldots, r$ be probability density functions with joint continuous or discrete state space and let $\{Y_t\}_{t=1}^{\infty}$ be a sequence of random variables such that, given that $X_t = j$, the conditional distribution of $Y_t$ has the density $f_j$. We furthermore assume that the $Y$-variables are conditionally independent, given the $X$-variables, i.e. the conditional density of

$$Y_1, \ldots, Y_T | X_1 = j_1, \ldots, X_T = j_T$$

is

$$\prod_{t=1}^{T} f_j (y_t).$$

The marginal density of $Y_t$ is then

$$f(y) = \sum_{j=1}^{r} \pi_j f_j(y),$$

so that we have a mixed distribution like (1), the difference being that the variables may now be dependent.

For regression models we will sometimes let the densities depend also on independent variables $u_{t(j)}$ in which case we replace (2) by

$$\prod_{t=1}^{T} f_j (y_t, u_{t(j)}).$$

In some literature a process $\{Y_t\}_{t=1}^{\infty}$ of the specified type is called a sequence of probabilistic functions of the Markov chain $\{X_t\}_{t=1}^{\infty}$. Such probabilistic functions appear in different technical and economical applications. They have been given a systematic treatment by Petrie (1969) and others. Most work has been done for functions with finite state space, since this easily reduces to the problem of partially observed finite Markov chains. Especially useful is the treatment by Baum et al (1970) from which many of the ideas in this paper have been taken.
EXAMPLE 2.1 A mixture of Poisson distributions,

\[ \sum_{j=1}^{r} \pi_j \exp(-\theta_j) \theta_j^{y/y!} \], \quad y = 0, 1, ...

appears to be one of the most simple mixed distributions. It will here be considered in connexion with a special mixed Poisson process.

Let \( Y_t = N_t - N_{t-1} \) be the number of events in the interval \((t-1,t]\) in an inhomogeneous Poisson process with an intensity function \( \lambda_t \) which is constant over every interval \((t-1,t], t = 1, 2, ...\) but otherwise changes randomly according to a Markov chain. This is a special type of a doubly stochastic Poisson process. Each \( Y_t \) has a mixed Poisson distribution with probability function (3), where \( \theta_j, j = 1, ..., r \) are the possible values of the intensity function, and \( \pi_j \) the corresponding stationary probabilities in the Markov chain.

EXAMPLE 2.2 The normal mixture

\[ f(y) = \sum_{j=1}^{r} \pi_j \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right) \]

is a typical example of a location and scale parameter mixture, and as such it has several interesting features; see Theorem 4.2 and Section 4.3.

EXAMPLE 2.3 "Switching autoregression" When autoregressive and moving average processes are used as models for long time series the fitted coefficients are often found to be instable over time. One could then sometimes consider using a Markov regime model for abrupt parametric changes. For a pure autoregression this would mean that if the regime \( X_t \) is equal to \( j \) then

\[ y_t = \beta^{(j)} u_t + \epsilon^{(j)} \]

where \( \beta^{(j)} = (\beta_1^{(j)}, ..., \beta_p^{(j)}) \) and \( u_t = (y_{t-1}, ..., y_{t-p}) \); with independent residuals \( \epsilon_t^{(j)} \).
3. Reconstruction of states

It the regimes \( X_1, \ldots, X_T \) which govern the distributions of the observations \( y_1, \ldots, y_T \) are completely known, the estimation of the parameters in the marginal distribution (1) is a simple statistical problem involving certain, but random, numbers of observations from each of the separate distributions \( f_j, j = 1, \ldots, r \). In order to carry out the estimation when the regimes are unobservable we need some reconstruction technique based on what is actually observed. We will here employ a missing information principle which utilizes the posterior distribution of the regime \( X_t \) given the observations \( y_1, \ldots, y_t \); cf. Orchard & Woodbury (1972) and Sundberg (1974). Reconstruction of \( X_t \) from observations \( y_1, \ldots, y_t \) is usually called a filtering problem, while reconstruction from \( y_1, \ldots, y_T \) is called an interpolation problem. All information about \( X_t \) is contained in the conditional probabilities

\[
\hat{\pi}_j(t|s) = P(X_t = j | y_1, \ldots, y_s)
\]

of which we shall make frequent use of

\[
\hat{\pi}_j(t|t) = P(X_t = j | y_1, \ldots, y_t)
\]

and

\[
\hat{\pi}_j(t|T) = P(X_t = j | y_1, \ldots, y_T).
\]

The filtering probabilities \( \hat{\pi}_j(t|t) \) can be calculated in real time, while the interpolation probabilities \( \hat{\pi}_j(t|T) \) can be used retrospectively after the whole experiment has been completed.

We also need the conditional probability that a transition has taken place between any two states \( j \) and \( k \),

\[
\hat{\pi}_{jk}(t|T) = P(X_t = j, X_{t+1} = k | y_1, \ldots, y_T),
\]
and the conditional density for \( y_{t+1}, \ldots, y_T \) given \( X_t = j \),

\[
\phi_j(t) = f_{y_{t+1}, \ldots, y_T | X_t = j}(y_{t+1}, \ldots, y_T).
\]

As starting values we define

\[
\hat{\pi}_j(1|0) = \pi_j = P(X_1 = j),
\]

\[
\phi_j(T) = 1, j = 1, \ldots, r.
\]

For notational convenience we introduce the normalizing operator \( N \) such that if \( \alpha_j, j = 1, \ldots, r \) are non-negative real numbers then

\[
\alpha_j N = \frac{\alpha_j}{\sum_{k=1}^{r} \alpha_k}, j = 1, \ldots, r.
\]

The following forward and backward recursion formulas follow from the Markov structure; see also Baum et al (1970).

**LEMMA 3.1**

\[
\hat{\pi}_j(t|t-1) = \sum_{i=1}^{r} \hat{\pi}_i(t-1|t-1) \pi_{ij},
\]

\[
\hat{\pi}_j(t|t) = \sum_{i=1}^{r} \hat{\pi}_i(t-1|t-1) \pi_{ij} f_j(y_t) N = \hat{\pi}_j(t|t-1) f_j(y_t) N,
\]

(4) \[
\phi_i(t) = \sum_{j=1}^{r} \pi_{ij} f_j(y_{t+1}) \phi_j(t+1),
\]

\[
\hat{\pi}_j(t|T) = \hat{\pi}_j(t|t) \phi_j(t) N,
\]

(5) \[
\hat{\pi}_{jk}(t|T) = \hat{\pi}_j(t|T) \pi_{jk} f_k(y_{t+1}) \phi_k(t+1) / \phi_j(t).
\]

**PROOF** We show only (4) and (5). The other formulas are even simpler.

The Markov structure gives

\[
\phi_i(t) = \sum_{j=1}^{r} f_{y_{t+1}, \ldots, y_T, X_{t+1} | X_t = i}(y_{t+1}, \ldots, y_T) =
\]

\[
= \sum_{j=1}^{r} P(X_{t+1} = j | X_t = i) f_{y_{t+1}}(y_{t+1}) \cdot f_{y_{t+2}, \ldots, y_T | X_{t+1} = j}(y_{t+2}, \ldots, y_T) =
\]

\[
= \sum_{j=1}^{r} \pi_{ij} f_j(y_{t+1}) \phi_j(t+1),
\]
and
\[ \hat{\pi}_{jk}(t|T) = \frac{\hat{f}_{y_{1}, \ldots, y_{t}}(y_{1}, \ldots, y_{t})}{\hat{f}_{y_{1}, \ldots, y_{t}}(y_{1}, \ldots, y_{t})} \cdot P(X_t = j|y_{1}, \ldots, y_{t}) \]

\[ P(X_{t+1} = k|X_{t} = j)f_{y_{t+1}}|X_{t+1} = k^{y_{t+1}}|y_{t+2}, \ldots, y_{T}|X_{t} = k(y_{t+2}, \ldots, y_{T}) = \]

\[ = C_{t} \hat{\pi}_j(t|t)\pi_{jk}f_k(y_{t+1})\phi_k(t+1), \text{ say.} \]

Since \( \Sigma_{k=1}^{r} \hat{\pi}_{jk}(t|T) = \hat{\pi}_j(t|T) \) and \( \Sigma_{k=1}^{r} \pi_{jk}f_k(y_{t+1})\phi_k(t+1) = \phi_j(t) \),
we find that \( C_{t} \hat{\pi}_j(t|t) = \hat{\pi}_j(t|T)/\phi_j(t) \), which implies (5). \( \square \)

**EXAMPLE 3.1** "Switching autoregression" A mixed AR(2)-process switches between the models
\[ Y_t = 1.6Y_{t-1} - 0.7Y_{t-2} + \varepsilon_t^{(1)} \]
and
\[ Y_t = 1.4Y_{t-1} - 0.7Y_{t-2} + \varepsilon_t^{(2)} \]
where the residuals \( \varepsilon_t^{(i)} \) are normal with mean zero and with variances such that under stable conditions, i.e. no switching, \( V(Y_t) = 1 \) in both models. It should be noted that the actual value of \( V(Y_t) \) depends on the degree of switching.

Figures 1-4(a) show simulated series with \( T = 500 \) where the regimes are mixed in equal proportions with different transition matrices \( P = (\pi_{jk}) \). Figures 1-4(b) show the filtering probability \( \hat{\pi}_j(t|t) \) for \( j = 1; n_T \) denotes the number of correctly classified states, i.e. the number of times that \( \hat{\pi}_j(t|t) \) attains its maximum for \( j = X_t \). As is seen, the tendency for \( \hat{\pi}_j(t|t) \) to point at the correct state is highly dependent of the degree of prevalence for the regime to remain in its current state.
Figure 1(a)
Switching auto-regression; \( T=500 \)
\[ P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \]

Figure 1(b)
Filtering probability, \( \hat{n}_1(t|t) = \)
\[ P(X_t=1|y_1, \ldots, y_t) \]
\( n_T = 362 \)

Figure 2(a)&(b)
\[ P = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix} \]
\( n_T = 398 \)

Figure 3(a)&(b)
\[ P = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix} \]
\( n_T = 444 \)
4. ML-estimation of marginal and transition parameters

Of all decomposition techniques for mixtures, maximum likelihood estimation has proved to be the most effective, not only for large samples, but also for small ones. It is furthermore the only technique which can readily be generalized to dependent regimes.

From now on we let the densities $f_j$ in the marginal distribution (1) depend on an unknown, possibly multidimensional, parameter $\theta_j$ so that each component in the mixture has its own parameter, $f_j(y) = f_j(y; \theta_j)$. The marginal density for each $Y_t$ is then

$$f(y) = \sum_{j=1}^{r} \pi_j f_j(y; \theta_j).$$

To begin with we suppose that the proportions $\pi_j$ and the transition probabilities $\pi_{jk}$ are known, and concentrate on the estimation of $\theta = (\theta_1, \ldots, \theta_r)$. We will return to the estimation of the Markov parameters in Section 4.3.

Maximum likelihood estimation of parameters in mixed distributions for independent variables is usually performed either by direkt maximization of the likelihood function by a modified Newton-Raphson technique,
or via an iterative scheme where the observations \( y_t \) are weighted with the posterior probabilities \( P(X_t = j | y_t) \); see Hasselblad (1966) and Behboodian (1970). Both principles can be used when the regimes are Markov dependent.

4.1 Direct maximization. The likelihood function and its derivatives can be calculated recursively using Lemma 3.1 and the posterior probabilities \( \hat{\pi}_j(t|t-1) \).

**THEOREM 4.1** The loglikelihood function for the parameter \( \Theta = (\Theta_1, \ldots, \Theta_r) \) given the observations \( y_1, \ldots, y_T \) is

\[
L(\Theta) = \sum_{t=1}^{T} \log f(y_t | y_1, \ldots, y_{t-1})
\]

where

\[
f(y_t | y_1, \ldots, y_{t-1}) = f_t = \sum_{j=1}^{r} \hat{\pi}_j(t|t-1)f_j(y_t; \Theta_j)
\]

is the posterior density for \( y_t \) given \( y_1, \ldots, y_{t-1} \). Writing \( \Theta_k \) for the vector of partial derivatives with respect to the components in \( \Theta_k \) one also has

\[
\frac{\partial L}{\partial \Theta_k} = \sum_{t=1}^{T} f_t^{-1} \left( \pi_k(t|t-1) \frac{\partial}{\partial \Theta_k} f_k(y_t; \Theta_k) \right) + \sum_{j=1}^{r} \frac{\partial}{\partial \Theta_k} \hat{\pi}_j(t|t-1)f_j(y_t; \Theta_j)
\]

and

\[
\frac{\partial}{\partial \Theta_k} \hat{\pi}_j(t|t-1) = \sum_{i=1}^{r} \frac{\partial}{\partial \Theta_k} \pi_i(t-1|t-1)\pi_{ij},
\]

\[
\frac{\partial}{\partial \Theta_k} \hat{\pi}_j(t|t) = \hat{\pi}_k(t-1|t-1)\pi_{kj} \frac{\partial}{\partial \Theta_k} f_k(y_t; \Theta_k) + \sum_{i=1}^{r} \frac{\partial}{\partial \Theta_k} \pi_i(t-1|t-1)\pi_{ij}f_j(y_t; \Theta_j).
\]

\[\square\]
The likelihood function and its derivatives can be calculated via the recursive formulas, and thus numerical maximization is possible. Even the second partial derivatives can be computed recursively, but the formulas are complicated and can give rise to doubts about the numerical accuracy.

4.2 An iterative maximization scheme. The iterative scheme for the ML-estimates derived by Hasselblad is intuitively appealing in that it replaces the missing information on the regime $X_t$ with the likelihood for the different outcomes. Orchard & Woodbury (1972) and Sundberg (1974) have given a systematic account of this missing information technique for independent variables.

Without coupling it to mixed distributions Baum et al (1970) have derived an iterative technique analogous to that of Hasselblad for estimation of parameters in Markov regime models; see also Baum (1972). We will here summarize Baum’s results in Theorem 4.2.

The iteration starts with initial estimates $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)$ of the marginal parameters, from which we can compute the posterior probabilities for interpolation

$$\hat{p}_j(t|T) = P_\theta(X_t=j|y_1, \ldots, y_T)$$

by means of Lemma 3.1. The dependence on $\theta$ is indicated by the index $P_\theta$.

For each $j$ separately one then maximizes the weighted loglikelihood function

$$\frac{1}{T} \sum_{t=1}^{T} \hat{p}_j(t|T) \log f_j(y_t; \theta'_j)$$

as a function of $\theta'_j$. Let us for the moment use the notation $\theta'_j$ also for the solution of this maximization problem. Then $\theta' = (\theta'_1, \ldots, \theta'_r)$ is a new and, as will be seen, better set of estimates of the marginal
parameters. The whole procedure can be repeated, and under mild restrictions the estimates will converge towards a local maximum of the likelihood function.

The transformation \( \theta \rightarrow \tau(\theta) = \theta' \) is often easy to express explicitly, and the new estimates will have a simple form. The following theorem contains the results by Baum et al (1970) concerning the transformation \( \tau \).

**THEOREM 4.2** Let the transformation \( \tau \) be defined by \( \tau(\theta) = \theta' \), where \( \theta' \) maximizes

\[
\sum_{t=1}^{T} \hat{\pi}_j(t|T) \log f_j(y_t; \theta'_j), \quad j = 1, \ldots, r.
\]

Then \( \tau \) increases the likelihood function, i.e.

\[
L(\tau(\theta)) \geq L(\theta),
\]

with strict inequality unless \( \theta \) is a stationary point of \( L \), in which case \( \tau(\theta) = \theta \). The transformation is unique and continuous if the densities \( f_j \) fulfill some mild regularity and extremal conditions, e.g. if \( \log f_j(y; \theta_j) \) is strictly concave for each \( y \), or if \( \theta_j = (\mu_j, \sigma_j) \) is a location and scale parameter and

\[
f_j(y; \theta_j) = \sigma_j^{-1} f((y-\mu_j)/\sigma_j), \quad j = 1, \ldots, r,
\]

where \( \log f(u) \) is strictly concave. Of course one also need that for all \( j \) the posterior probability \( \hat{\pi}_j(t|T) \) is strictly positive for at least one \( t \).

**PROOF** We will here only give the proof of (7). The reader is referred to the original article by Baum et al for a detailed discussion of uniqueness and regularity conditions.

To show (7) let \( x = (x_1, \ldots, x_T) \) denote a sequence of outcomes for the regime process \( \{X_t, t=1, \ldots, T\} \), and write
\[ p(x, \theta) = \prod_{t=1}^{T} f_{x_t}(y_t; \theta_{x_t}) \]

for the conditional likelihood of \( y_t, t = 1, \ldots, T \) given that \( x_t = x_{t_t}, t = 1, \ldots, T \). If \( \mu(x) \) is the probability for the sequence \( x \), i.e.

\[ \mu(x) = p_{\theta}(x_t = x_{t_t}, t = 1, \ldots, T), \]

then the total likelihood is

\[ L(\theta) = \sum_{x} \mu(x)p(x, \theta). \]

Now separate the product \( p(x, \theta) \) into \( r \) factors according to the value of \( x_t \), i.e.

\[ p(x, \theta) = \prod_{j=1}^{r} \left( \prod_{t; x_t = j} f_j(y_t; \theta_j) \right) = \prod_{j=1}^{r} p_j(x, \theta_j), \]

say, and define

\[ Q_j(\theta; \theta'_j) = \sum_{x} \mu(x)p(x, \theta) \log p_j(x, \theta'_j). \]

Then (7) will follow if we show that there is a constant \( C_\theta \), depending only on \( y_1, \ldots, y_T \) and \( \theta \), such that for \( j = 1, \ldots, r \)

\[ Q_j(\theta; \theta'_j) = C_\theta \sum_{t=1}^{T} \hat{\eta}_j(t|T) \log f_j(y_t; \theta'_j) \]

and that

\[ Q_j(\theta; \theta'_j) \geq Q_j(\theta; \theta_j), j = 1, \ldots, r \Rightarrow L(\theta') \geq L(\theta). \]

We first show (8):

\[ Q_j(\theta; \theta'_j) = \sum_{x} \mu(x)p(x, \theta) \log \left( \prod_{t; x_t = j} f_j(y_t; \theta'_j) \right) = \]

\[ = \sum_{x} \mu(x)p(x, \theta) \sum_{t; x_t = j} \log f_j(y_t; \theta'_j) = \]

\[ = \sum_{t=1}^{T} \log f_j(y_t; \theta'_j) \sum_{x; x_t = j} \mu(x)p(x, \theta). \]


Here

\[(10) \quad \sum_{x; x_t=j} \mu(x)p(x, \theta) = p_{\theta}(X_t=j|y_1, \ldots, y_T)L(\theta) = \hat{\pi}_j(t|T)L(\theta)\]

is the combined likelihood for \(X_t = j\) and \(y_1, \ldots, y_T\), so that

\[Q_j(\theta; \theta'_j) = L(\theta) \sum_{t=1}^{T} \hat{\pi}_j(t|T) \log f_j(y_t; \theta'_j).\]

This shows (8) with \(C_\theta = L(\theta)\), which obviously does not depend on \(\theta'_j\), and we have shown that \(\theta' = \tau(\theta) = (\theta'_1, \ldots, \theta'_r)\) maximizes each \(Q_j(\theta; \theta'_j)\) as a function of \(\theta'_j\).

We now turn to the implication (9), which will follow from Jensen's inequality by summing the different \(Q_j\),

\[Q(\theta; \theta') = \sum_{j=1}^{r} Q_j(\theta; \theta'_j) = \sum_{x} \mu(x)p(x, \theta) \log p(x, \theta').\]

Writing \(E_{\theta}\) for expectation with respect to the measure \(L(\theta)^{-1} \mu(\cdot)p(\cdot, \theta)\), we have

\[
\log \frac{L(\theta')}{L(\theta)} = \log \frac{\sum_{x} \mu(x)p(x, \theta')/L(\theta)}{\sum_{x} \mu(x)p(x, \theta)/L(\theta)} = \\
\log \frac{\sum_{x} \mu(x)p(x, \theta)}{L(\theta)} \cdot \frac{p(x, \theta')}{p(x, \theta)} = \\
= \log E_{\theta}(p(x, \theta')/p(x, \theta)) \geq E_{\theta}(\log p(x, \theta')/p(x, \theta)) = \\
= \sum_{x} \frac{\mu(x)p(x, \theta)}{L(\theta)} \log \frac{p(x, \theta')}{p(x, \theta)} = \\
= \frac{1}{L(\theta)} \{Q(\theta; \theta') - Q(\theta; \theta)\} \geq 0
\]

if \(Q(\theta; \theta') \geq Q(\theta; \theta)\). Now since

\[\sup_{\theta'} Q(\theta; \theta') = \sum_{j=1}^{r} \sup_{\theta'_j} Q_j(\theta; \theta'_j)
\]

we get (9).
EXAMPLE 4.1 For a mixed Poisson distribution with

\[ f_j(y; \theta_j) = \exp(-\theta_j) \theta_j^y y! , \]

\[ \sum_{t=1}^{T} \hat{\pi}_j(t|T) \log f_j(y_t; \theta_j') = \sum_{t=1}^{T} \hat{\pi}_j(t|T) \{-\theta_j' + y_t \log \theta_j' - \log y_t! \} \]

is maximized by

\[ \theta_j' = \frac{\sum_{t=1}^{T} \hat{\pi}_j(t|T) y_t}{\sum_{t=1}^{T} \hat{\pi}_j(t|T)} . \]

EXAMPLE 4.2 For mixed normal distributions with unknown location and scale parameters one gets the same iterative scheme as Hasselblad (1966) and Behboodian (1970), but with \( \hat{\pi}_j(t|T) \) computed from the Markov model,

\[ \mu_j' = \frac{\sum_{t=1}^{T} \hat{\pi}_j(t|T) y_t}{\sum_{t=1}^{T} \hat{\pi}_j(t|T)} , \]

\[ (\sigma_j')^2 = \frac{\sum_{t=1}^{T} \hat{\pi}_j(t|T) y_t^2}{\sum_{t=1}^{T} \hat{\pi}_j(t|T)} - \mu_j'^2 . \]

EXAMPLE 4.3 In a switching regression model with normal residuals,

\[ (Y_t | X_t = j) \sim N(\alpha_j + \beta_j u_t(j), \sigma_j^2) , \]

the parameters are \( \theta_j = (\alpha_j, \beta_j, \sigma_j^2) , j = 1, ... , r \). If we drop the primes from \( \alpha_j', \beta_j', (\sigma_j')^2 \), the maximization of \( Q_j(\theta; \theta_j') \) gives the equations

\[ \frac{\partial Q_j}{\partial \alpha_j} = \sum_{t=1}^{T} \hat{\pi}_j(t|T)(y_t - \alpha_j - \beta_j u_t(j)) = 0 , \]

\[ \frac{\partial Q_j}{\partial \beta_j} = \sum_{t=1}^{T} \hat{\pi}_j(t|T) u_t(j)(y_t - \alpha_j - \beta_j u_t(j)) = 0 , \]

\[ \frac{\partial Q_j}{\partial \sigma_j^2} = \sum_{t=1}^{T} \hat{\pi}_j(t|T)((y_t - \alpha_j - \beta_j u_t(j))^2 - \sigma_j^2)/2\sigma_j^4 = 0 , \]

which are only slight modifications of the familiar normal equations. \( \square \)
4.3 Estimation of transition probabilities. In reality, the mixing proportions $\pi_j$ are unknown, and so are of course the transition probabilities $\pi_{jk}$ in the postulated Markov chain. The maximum likelihood estimates of the transition matrix $\mathbf{\pi} = (\pi_{ij})$, and consequently also of the stationary mixing proportions $\mathbf{\pi_s} = (\pi_{1}, \ldots, \pi_{r})$, can be obtained by a similar iteration technique that yields the estimates of the marginal parameters $\theta = (\theta_1, \ldots, \theta_r)$. This is most clearly seen from an analogy with the situation when the Markov chain $X_t$, $t = 1, \ldots, T$ can be observed. Then the maximum likelihood estimates of the transition probabilities $\pi_{jk}$ are obtained by a maximization of

$$
\log \pi_{x_1} + \sum_{t=1}^{T-1} \log \pi_{x_t x_{t+1}} = \log \pi_{x_1} + \sum_{j=1}^{r} \left( \sum_{k=1}^{r} n_{jk} \log \pi_{jk} \right)
$$

under the restriction $\pi_{jk} > 0$, $\sum_k \pi_{jk} = 1$. The number of transitions $j \sim k$ in the sequence $x = (x_1, \ldots, x_T)$ is denoted by $n_{jk}$,

$$n_{jk} = \sum_{t=1}^{T-1} 1 \{x_t = j, x_{t+1} = k\}.$$

Maximization of (11) can be performed for each $j$ separately, and we see that $\sum_k n_{jk} \log \pi_{jk}$ attains its maximum for

$$\pi_{jk} = \frac{n_{jk}}{\sum_{k=1}^{r} n_{jk}}.$$

In the present situation we have to replace the indicator $1 \{x_t = j, x_{t+1} = k\}$ by the posterior probability of a transition $j \sim k$,

$$\hat{\pi}_{jk}(t|T) = P_{\theta, \pi}(X_t=j, X_{t+1}=k|y_1, \ldots, y_T),$$

calculated as in Lemma 3.1. The iteration procedure then goes as follows.

Take some initial estimates $\theta = (\theta_1, \ldots, \theta_r)$ and $\pi = (\pi_{jk})$ of the marginal and transition parameters, and let similarly $\pi_0 = (\pi_{01}, \ldots, \pi_{0r})$. 
be initial estimates of the starting distribution for $X_1$. Of course, estimation of $\pi_{0j} = P(X_1 = j)$ from only one sample makes no sense if we do not assume it to be the stationary distribution corresponding to the transition matrix $\pi$. In this context $\pi_0$ is introduced separate from $\pi_s$ only as a computational aid.

Using $\theta$, $\pi_0$, and $\pi$, compute the posterior probabilities for states and transitions, $\hat{\pi}_j(t|T)$ and $\hat{\pi}_{jk}(t|T)$ by Lemma 3.1, and finally take as new estimates

$$\pi_{0j} = \hat{\pi}_j(1|T)$$

and

$$\pi_{jk} = \frac{\sum_{t=1}^{T-1} \hat{\pi}_{jk}(t|T)}{\sum_{r=1}^{r} \sum_{t=1}^{T-1} \sum_{k=1}^{K} \hat{\pi}_{jk}(t|T)}$$

in analogy with (12). The analogy with the completely observable case goes even further, and as will be shown in the following theorem, $\pi_{jk}'$ maximizes the analog to (11),

$$\sum_{k=1}^{r} \left\{ \sum_{t=1}^{T-1} \hat{\pi}_{jk}(t|T) \right\} \log \pi_{jk}'$$

The whole procedure can then be repeated as far as necessary.

The stationary mixing proportions $\pi_j$ are most naturally estimated by

$$\pi_{j}' = T^{-1} \sum_{t=1}^{T} \hat{\pi}_j(t|T)$$

even if this does not give the maximum likelihood estimator.

**THEOREM 4.3** The transformations $(\pi_0, \pi) \sim (\pi_0', \pi')$ defined by (13) and (14) and $\theta \sim \theta'$ defined in Theorem 4.2 increase the likelihood function, i.e.
\[ L(\theta', \pi_0', \pi') \geq L(\theta, \pi_0, \pi), \]

with strict inequality unless \((\theta, \pi_0, \pi)\) is a stationary point.

**Proof** The following proof is a rearrangement of a proof given by Petrie (1969). We can write the total likelihood as

\[ L(\theta, \pi_0, \pi) = \sum_x \mu(x, \pi_0, \pi)p(x, \theta), \]

where the likelihood for the unobservable \(x\)-sequence is

\[ \mu(x, \pi_0, \pi) = P_{\pi_0, \pi}(X_t = x_t, t=1, \ldots, T) = \prod_{t=1}^{T} \pi_{0x_t} \pi_{x_t x_{t+1}}. \]

We find, as in (9) in the proof of Theorem 4.2, that in order to increase the likelihood, we can maximize

\[ Q(\theta, \pi_0, \pi; \theta', \pi_0', \pi') = \sum_x \mu(x, \pi_0, \pi)p(x, \theta) \log \{\mu(x, \pi_0', \pi')p(x, \theta')\} \]

as a function of \((\theta', \pi_0', \pi')\). This \(Q\)-function can be split into three terms, which can be maximized separately, as functions of \(\pi_0', \pi',\) and \(\theta',\) respectively:

\[ Q = \sum_x \mu(x, \pi_0, \pi)p(x, \theta) \left\{ \log \pi_{0x_1} + \sum_{t=1}^{T-1} \log \pi_{x_t x_{t+1}} + \log p(x, \theta') \right\}. \]

Here we already know, from Theorem 4.2, that \(\theta'\) maximizes

\[ \sum_x \mu(x, \pi_0, \pi)p(x, \theta) \log p(x, \theta'), \]

so what we actually have to prove is that \(\pi_0'\) and \(\pi'\) defined by (13) and (14) maximize

\[ \sum_x \mu(x, \pi_0, \pi)p(x, \theta) \log \pi_{0x_1} \]

and

\[ \sum_x \mu(x, \pi_0, \pi)p(x, \theta) \sum_{t=1}^{T-1} \log \pi_{x_t x_{t+1}}, \]

respectively.
Using (10) we can write (15) as
\[ \sum_{j=1}^{r} \sum_{x; x_1 = j}^{\infty} \mu(x; \pi_0; \pi) p(x; \theta) \log \pi'_{0x_1} = L(\theta; \pi_0; \pi) \sum_{j=1}^{r} \hat{n}_j(1|T) \log \pi'_{0j}, \]
which is maximized by \( \hat{\pi}'_{0j} = \hat{n}_j(1|T) \) as required.

Similarly, (16) can be written as the sum of \( r \) terms,
\[ \sum_{j=1}^{r} \sum_{t=1}^{T-1} \sum_{x; x_1 = j}^{\infty} \mu(x; \pi_0; \pi) p(x; \theta) \log \pi'_{jx_{t+1}}, \]
which can be maximized for each \( j \) separately as functions of the vectors \( (\pi'_{j1}, \ldots, \pi'_{jr}) \) under the restrictions \( \pi'_{jk} \geq 0, \Sigma_k \pi'_{jk} = 1 \). The terms can be written
\[
\begin{align*}
\sum_{t=1}^{T-1} \sum_{x; x_1 = j}^{\infty} \mu(x; \pi_0; \pi) p(x; \theta) \log \pi'_{jx_{t+1}} & = \sum_{k=1}^{r} \log \pi'_{jk} \sum_{t=1}^{T-1} \sum_{x; x_1 = j, x_{t+1} = k}^{\infty} \mu(x; \pi_0; \pi) p(x; \theta) \\
& = L(\theta; \pi_0; \pi) \sum_{k=1}^{r} \log \pi'_{jk} \hat{n}_j(1|T),
\end{align*}
\]
(17)
since, in analogy with (10),
\[ \sum_{x; x_1 = j, x_{t+1} = k}^{\infty} \mu(x; \pi_0; \pi) p(x; \theta) = \]
\[ L(\theta; \pi_0; \pi) p_{\theta; \pi_0; \pi}(X_t = j, X_{t+1} = k|y_1, \ldots, y_T) \]
\[ = L(\theta; \pi_0; \pi) \hat{n}_{jk}(t|T). \]

Under the imposed restrictions, (17) is maximized by (14). This completes the proof of Theorem 4.3. \( \square \)

The iterative scheme for ML-estimation of \( \theta, \pi_0 \) and \( \pi \) has been coded by the author in a FORTRAN-subroutine MARMIX which has been implemented and tested at a CDC Cyber 172 of Umeå University Computing Center.

4.4 Robustness and consistency of ML-estimates. As many authors have
noted the ML-estimates of marginal parameters in mixed distributions are not necessarily consistent. In a location and scale parameter family such as in Example 2.2 the likelihood function is unbounded if $\mu_i$ is equal to some $y_i$ and $\sigma_i$ is near zero. This means that the likelihood is unbounded near the boundary of the parameter space $\Theta = \{(\mu_i, \sigma_i) ; \mu_i \in \mathbb{R}, \sigma_i > 0\}$. This can be remedied in different ways. Hasselblad (1966) makes a grouping of the observations which transforms the problem to an estimation problem in a multinomial distribution. Behboodian (1970) proposes that one shall take the greatest local maximum in the interior of $\Theta$ as the estimator. Hasselblad's method is natural from a practical point of view but leads to some numerical complications. More motivated from principles of statistical model building is to use a compact parameter space, and for the location and scale mixture take

$$\Theta = \{(\mu_i, \sigma_i) ; |\mu_i| \leq c, 0 < c_1 < \sigma_i < c_2, i=1,\ldots,r\}.$$ 

Besides from being mathematically attractive, this has the important statistical relevance that it will prevent the distributions from being too different and exclude aberrant forms from the class of permitted distributions. To be sure of this we of course also have to impose some continuity or differentiability restrictions. For example, if we fit a mixture of two normal distributions with means and standard deviations $\mu_1, \sigma_1$ and $\mu_2, \sigma_2$ respectively, we presumably have some prior reason for doing so, and then we also have some opinion of what reasonable values of $\sigma_1$ and $\sigma_2$ are. Permitting $\sigma_1$ to tend to zero introduces a one point distribution into the class of distributions, and if this is the intention, we should have done it from the beginning.

We will not embark upon a general treatment of asymptotic properties of the ML-estimates in Markov regime models. The consistency and asymptotic normality of the estimators seem to be established only for
families of discrete distributions \( f_i \) with finite state space; and that was done already by Baum & Petrie (1966) in an early paper on probabilistic functions of Markov chains.

In the rest of this section we will investigate the robustness against a dependence structure of the ML-estimates of marginal parameters derived under the hypothesis of independent regimes. In cases where the Markov dependence between regimes is unclear, or perhaps not even realized, one could be tempted to derive the ML-estimates of the parameters in the density

\[
f(y) = \sum_{j=1}^{r} \pi_j f_j(y; \theta_j)
\]

under the assumption that the regimes are independent. This means that one maximizes the loglikelihood function

\[
L^I(\theta, \pi_s) = \sum_{t=1}^{T} \log f(y_t)
\]

as a function of the marginal parameters \( \theta = (\theta_1, \ldots, \theta_r) \) and the stationary mixing proportions \( \pi_s = (\pi_1, \ldots, \pi_r) \). This is the standard mixing problem with independent observations. We call the solution an MLI-estimator (Maximum Likelihood under Independence assumption) and denote it \( (\hat{\theta}, \hat{\pi}_s) \). Robustness of MLI-estimators against different dependence structures has been studied by Ranneby (1975) for distributions with finite or countably infinite state space. The following is a version of his Theorem 3 for continuous state space.

**THEOREM 4.4** Suppose that the regime process \( \{X_t\}_{t=1}^{\infty} \) is ergodic, and that the parameter space \( \Theta \) for \( \theta \) is compact with the true parameter \( \theta^0 \) as an interior point. Also suppose that the mixture is identifiable in the sense that for each \( \delta > 0 \) there is a set of outcomes for \( Y_t \) with \( P(Y_t \in I_\delta) > 0 \) such that
\[ \inf_{\delta > 0} |f(y; \theta) - f(y; \theta^0)| > 0 \]

for all \( y \in I \). Furthermore suppose there is a neighbourhood \( S \) of \( \theta^0 \) such that

\[
E(\sup_{\theta} |\log f(Y; \theta)|) < \infty,
\]

\[
\int \sup_{\theta \in S} \left| \frac{\partial f(y; \theta)}{\partial \theta_i} \right| dy < \infty,
\]

\[
\int \sup_{\theta \in S} \left| \frac{\partial^2 f(y; \theta)}{\partial \theta_i \partial \theta_j} \right| dy < \infty,
\]

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

\[
E \left| \frac{\partial \log f(Y; \theta)}{\partial \theta_i} \right|^{2+\delta} < \infty,
\]

for some \( \delta > 0 \). Then the ML-estimator \( \hat{\theta} \) is consistent and asymptotically normal,

\[ \sqrt{T}(\hat{\theta} - \theta^0) \xrightarrow{L} N(0, \Sigma^{-1} C \Sigma^{-1}) \]

where \( \Sigma = (\sigma_{ij}), \ C = (c_{ij}), \)

\[
\sigma_{ij} = \text{Cov} \left\{ \frac{\partial \log f(Y)}{\partial \theta_i}, \frac{\partial \log f(Y)}{\partial \theta_j} \right\},
\]

\[
c_{ij} = \lim_{T \to \infty} \text{Cov} \left\{ \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \frac{\partial \log f(Y)}{\partial \theta_i}, \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial \log f(Y)}{\partial \theta_j} \right\}.
\]

Similar results hold for \( \hat{\pi}_s \) with \( \partial \theta_i \) replaced by \( \partial \pi_i \).

**Proof** The proofs of consistency and asymptotic normality given by Ranneby (1975) and (1976) work also for a probabilistic function of a Markov chain. We only have to check that the mixing condition for asymptotic independence of the \( Y_k \)-variables is satisfied, i.e. that if

\[ \sup_{A, B} |P(A \cap B) - P(A)P(B)| = \alpha_y(k) \]
when \( A \in \sigma(Y_1, \ldots, Y_r), B \in \sigma(Y_{t+k+1}, \ldots) \) then

\[
\sum_{k=1}^{\infty} \alpha_y(k)^{\delta/(2+\delta)} < \infty, \quad (\delta > 0).
\]

To see that (20) holds, let \( X(t) = (X_1, \ldots, X_t), X(t+k+1) = (X_{t+k+1}, \ldots) \) and define \( \alpha_x(k) \) by (19) with \( Y_s \) replaced by \( X_s \). Then

\[
\alpha_x(k) \leq c\rho^k, \quad (0<\rho<1),
\]

so that (20) is fulfilled with \( \alpha_y \) replaced by \( \alpha_x \). But it is easy to see that the dependence between \( Y_t \)-events can not be greater than that between \( X_t \)-events. Due to the conditional independence between \( A \) and \( B \) given \( X(t) \) and \( X(t+k+1) \) we have

\[
P(AB) = E(P(AB|X(t), X(t+k+1))) = E(P(A|X(t)) \cdot P(B|X(t+k+1))).
\]

Here \( P(A|X(t)) \) and \( P(B|X(t+k+1)) \) are bounded random variables, measurable with respect to \( \sigma(X(t)) \) and \( \sigma(X(t+k+1)) \), respectively, and it follows from Ibragimov & Linnik (1971), Theorem 17.2.1 that

\[
\alpha_y(k) \leq 4 \alpha_x(k).
\]

Hence (20) follows. The rest of the proof goes as in Ranneby (1975) and (1976).

Due to the Markov dependence structure and the conditional independence of \( Y_t, t = 1, \ldots, T \) given \( X_t, t = 1, \ldots, T \) the quantities \( \sigma_{ij} \) and \( c_{ij} \) can be calculated quite simply. Write

\[
f_j'(y) = \frac{\partial f_j(y; \theta_j)}{\partial \theta_j},
\]

\[
\rho \log f(Y_t) = \pi_j f_j'(Y_t) / f(Y_t),
\]

\[
u_j(t) = \frac{\partial \log f(Y_t)}{\partial \theta_j} = \pi_j f_j'(Y_t) / f(Y_t),
\]
and define
\[
m_j(k) = \mathbb{E}(u_j(t) | X_t = k) = \pi_j \int f_j'(y) f_k(y) / f(y) dy,
\]
\[
\sigma_{ij}(k) = \mathbb{E}(u_i(t) \cdot u_j(t) | X_t = k) = \pi_i \pi_j \int f_i'(y) f_j'(y) f_k(y) / f^2(y) dy.
\]
Conditioning on the sequence \( X = (X_1, X_2, \ldots) \) of regimes we have
\[
\text{Cov} \left\{ \frac{1}{\sqrt{T}} \sum_{s=1}^{T} u_j(s), \frac{1}{\sqrt{T}} \sum_{t=1}^{T} u_j(t) \right\} = \mathbb{E}_X \left\{ \frac{1}{T} \sum_{s,t=1}^{T} \mathbb{E}(u_i(s) u_j(t) | X) \right\}.
\]
To evaluate the conditional expectation we can use that \( Y_s \) and \( Y_t \) are conditionally independent given the sequence \( X \) and that the conditional distribution depends only on \( X_s \) and \( X_t \). Thus
\[
\sum_{s,t=1}^{T} \mathbb{E}(u_i(s) u_j(t) | X) = \sum_{s,t=1}^{T} \mathbb{E}(u_i(t) u_j(t) | X_s) +
\]
\[
+ \sum_{s+t}^{r} \mathbb{E}(u_i(s) u_j(t) | X_s X_t) = \sum_{k=1}^{r} \mathbb{E}(u_i(t) u_j(t) | X_t = k) +
\]
\[
+ \sum_{k=1}^{r} \sum_{l=1}^{r} \sum_{s+t}^{r} \mathbb{E}(u_i(s) | X_s = k) \cdot \mathbb{E}(u_j(t) | X_t = l).
\]
Using
\[
n_k(T) = \# \{ t; 0 < t < T, X_t = k \}
\]
this can be written
\[
\sum_{k=1}^{r} n_k(T) \{ \sigma_{ij}(k) - m_i(k) m_j(k) \} + \sum_{k=1}^{r} n_k(T) n_k(T) m_i(k) m_j(k).
\]
Taking expectations over \( X \) and noting that \( \mathbb{E}(n_k(T)) = T \pi_k \) we get the following expression for \( c_{ij} \),
\[
c_{ij} = \sum_{k=1}^{r} \pi_k \{ \sigma_{ij}(k) - m_i(k) m_j(k) \} +
\]
\[
+ \lim_{T \to \infty} \mathbb{E} \left\{ \frac{1}{T} \sum_{k=1}^{r} n_k(T) m_i(k) \cdot \sum_{l=1}^{r} n_k(T) m_j(l) \right\}.
\]
Since

$$E\left(\frac{1}{T} \sum_{k=1}^{r} \pi_k(T)m_i(k)\right) = \sum_{k=1}^{r} \pi_k m_i(k) = E(u_i(t)) = 0$$

the second term can be written

$$\lim_{T \to \infty} E\left(\sqrt{T} \sum_{k=1}^{r} (T^{-1}n_k(T) - \pi_k)m_i(k) \cdot \sqrt{T} \sum_{s=1}^{r} (T^{-1}n_s(T) - \pi_s)m_j(s)\right).$$

Now it is well known that the \( r \)-variate variable

$$\sqrt{T}(T^{-1}n_k(T) - \pi_k)^{r}_{k=1}$$

is asymptotically normal with covariance matrix \( B = (\beta_{k\ell}) \),

$$\beta_{k\ell} = \delta_{k\ell} \pi_k - \pi_k \pi_{\ell} + \sum_{t=2}^{\infty} \{P(X_1 = k, X_t = \ell) + P(X_1 = \ell, X_t = k) - 2\pi_k \pi_{\ell}\},$$

so we get the simple expression

$$c_{ij} = \sum_{k=1}^{r} \pi_k (\sigma_{ij}(k) - m_i(k)m_j(k)) + \sum_{k, \ell=1}^{r} m_i(k) \beta_{k\ell} m_j(\ell).$$

For \( \sigma_{ij} \) we get the analogue formula

$$\sigma_{ij} = \sum_{k=1}^{r} \pi_k \sigma_{ij}(k).$$

REMARK 4.5 It is slightly surprising that the asymptotic properties of the MLI-estimates depend on the transition probabilities, since what affects the solution of the likelihood equation

$$\frac{\partial L_i(\theta, \pi_s)}{\partial \theta_i} = 0, \ i = 1, \ldots, \ r$$

is only the number of times that \( X_t = k, k = 1, \ldots, r \), not the order in which the regimes appear. However, the transition probabilities determine how close \( T^{-1}n_k(T) \) falls to \( \pi_k \), and as the form of \( c_{ij} \) indicates this affects also the efficiency of the estimators of the marginal parameters \( \theta_i \).
REMARK 4.6 The loglikelihood function is given by (6) under the full Markov model and by (18) under the independence hypothesis. If we choose $\pi_s$ as a starting distribution for $X_1$, then

$$2(\sup_{\theta, \pi_s} L(\theta, \pi_s, \pi) - \sup_{\theta, \pi_s} L^T(\theta, \pi_s))$$

could be used as a $\chi^2$-statistic with $(r-1)^2$ degrees of freedom for a test of independence. □

5. Small sample properties of ML-estimates

There are in essence two methods for obtaining the variance of the ML-estimates $\theta^*, \pi^*_s, \pi^*$ of $\theta, \pi_s, \pi$ for finite $T$ and they are inversion of the matrix of observed second order partial derivatives of the loglikelihood function and simulation of known mixtures.

The asymptotic variance as $T \to \infty$ can be obtained from the matrix of second order partial derivatives of the loglikelihood function at the point $\theta^*, \pi^*_s, \pi^*$; for a recursive method of doing this, see Theorem 4.1. This is the technique used by Hasselblad (1966) for independent mixtures.

However, and this has been shown in several simulation studies with moderately large samples, the actual variance of the estimators are usually greater than those computed from the asymptotic variance; see Dick & Bowden (1973).

In this section we will present the results of a simple simulation study with the aim to show the effect of the dependence between regimes on the ML-estimates and the MLI-estimates. The mixture is composed of two normal distributions with means $\mu_1 = 0$ and $\mu_2 = 10$ and standard deviations $\sigma_1 = \sigma_2 = \sigma$ where $\sigma = 2.5$ and 5.0. The mixing propor-
tions are hold equal, $\pi_1 = \pi_2 = 0.5$ while the transition matrix

\[
\begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix}
\]

is varied from a strong negative dependence with $\pi_{11} = \pi_{22} = 0.1$ to a strong positive dependence with $\pi_{11} = \pi_{22} = 0.9$.

Figure 5.1 Mixture of normal densities with $\mu_1 = 0$, $\mu_2 = 10$, $\sigma_1 = \sigma_2 = \sigma = 2.5$ and $5.0$, $\pi_1 = \pi_2 = 0.5$.

The ML-estimates of $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)$, $\pi_s = (\pi_1, \pi_2)$, and $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ were obtained via the iterative scheme presented in Section 4.2 and 4.3 by means of the FORTRAN-subroutine MARMIX. The true parameter values were always used as the starting points, and the iterations carried on until no two successive approximations differed more than $\varepsilon = 0.002$. In no case more than 20 iterations were permitted.

For comparison the ML-estimators of $\theta$ and $\pi_0$ were computed by a similar iteration scheme, also incorporated in the subroutine MARMIX.

The choice of the true parameters as starting points and the restriction of the number of iterations to 20 will apparently increase the
accuracy of the estimators, but it will probably not affect the compari-
son between the ML- and the MLI-estimators. Further investigations are
necessary to show how the choice of starting points affect the estimates.

Two sample sizes were considered, one very small with \( T = 50 \), and
one moderately small with \( T = 200 \). The number of simulations were 200
and 100, respectively, for every combination. The results are presented
below as the observed RMSE (Root Mean Square Error, \( E((\theta^*_t - \theta_t)^2)^{1/2} \)) for
the different estimates.

**Remark 5.1** The accuracy of the estimates \( \pi^*_j \) of the mixing propor-
tions \( \pi_j \) needs special care. In fact, \( \pi_j \) is only the prior proba-
bility that observation \( t \) comes from regime \( j \) and it equals the long
run proportion of regimes equal to \( j \). In the sample, however, the
real proportion is

\[
\frac{1}{T} n_j(T) = p_j, \text{ say,}
\]

which is a random quantity. The total mean square error can be split
according to

\[
E((\pi^*_j - \pi_j)^2) = E((\pi^*_j - p_j)^2) + 2E((\pi^*_j - p_j)(p_j - \pi_j)) + E((p_j - \pi_j)^2),
\]

where

\[
E((p_j - \pi_j)^2) \approx T^{-1}\beta_{jj}
\]

from (23). It can be argued that \( \pi^*_j \) should be regarded as an estimate
of \( p_j \) and that only \( E((\pi^*_j - p_j)^2) \) has any relevance for the valida-
tion of the estimation procedure. In this the marginal distributions
\( f_j \) are used to split the observations \( y_t \) according to the presumed
value of \( X_t \), which will give an estimate of \( p_j \). On the other hand,
inference about \( \pi_j \) from exact or inexact knowledge of \( p_j \) is a ques-
tion of how much we trust the Markov regime model for the generation of
regimes.
It seems as if previous authors have not made this distinction. Both Hasselblad (1966) and Dick & Bowden (1973) present \( E((\pi^j - \pi_j)^2) \) without comment.

In the Markov regime model the separation of \( E((\pi^j - \pi_j)^2) \) into its components is of special importance, since \( E((p_j - \pi_j)^2) \) is strongly dependent on the transition probabilities \( \pi_{ij} \) and increases drastically as the diagonal elements \( \pi_{jj} \) tend to one; see Figure 5.4 and 5.8.

As is seen from the simulations it is only for a fairly strong positive or negative dependence that the full ML-estimator of \( \mu_1, \mu_2 \) and \( \sigma_1, \sigma_2 \) performs better than the simpler MLI-estimator. The addition of an extra dependence parameter obviously causes an overfit to data, at least for such a small sample size as \( T = 200 \).

It is also seen, as could be anticipated, that for a negative dependence where the regimes alternate, the ML-estimates of \( \mu_1, \mu_2 \) and \( \sigma_1, \sigma_2 \) are almost as good as when the regimes are completely known.

One important fact is that the iterative estimation schemes very seldom converged within 20 iterations for the sample size \( T = 200 \) and \( \sigma_1 = \sigma_2 = 5 \). For \( \pi_{11} = \pi_{22} = 0.1 \) and \( 0.9 \) the ML-estimator converged in about 50% of the simulations. All other combinations gave convergence in less than about 10% of the simulations.

When \( \sigma_1 = \sigma_2 = 2.5 \) all procedures converged in at least 75%, and the ML-estimator in almost 100% of the simulations when there actually was a dependence.

Separate studies indicated that the convergence properties were considerably improved with further iterations, but that the observed variance of the estimator then also increased. Incidentally, this indicates that the ML-estimator is more superior over the MLI-estimator than is hinted from the diagrams, in cases of dependence.
Figure 5.2 Sample RMSE, $E((\mu_i^* - \mu_i)^2)^{1/2}$, $i = 1, 2$; $T = 50$, $NSIM = 200$;

$\sigma$---o MLI-estimator, x---x ML-estimator.

Figure 5.3 Sample RMSE, $E((\sigma_i^* - \sigma_i)^2)^{1/2}$, $i = 1, 2$; $T = 50$, $NSIM = 200$;

$\sigma$---o MLI-estimator, x---x ML-estimator.
Figure 5.4 Sample RMSE, $E((\pi_{1i}^* - \pi_{11})^2)^{1/2}$, $i = 1, 2$; $T = 50$, NSIM = 200; 
- - - - - ML-estimator, x--x ML-estimator; solid curve = $E((p_{1i} - \pi_{1i})^2)^{1/2}$.

Figure 5.5 Sample RMSE, $E((\pi_{1j}^* - \pi_{1j})^2)^{1/2}$, $i, j = 1, 2$; $T = 50$, 
NSIM = 200; x--x ML-estimator.
Figure 5.6 Sample RMSE, $E((\mu_i^* - \mu_i)^2)^{1/2}$, $i = 1, 2; T = 200, NSIM = 100$;
o---o MLI-estimator, x---x ML-estimator.

Figure 5.7 Sample RMSE, $E((\sigma_i^* - \sigma_i)^2)^{1/2}$, $i = 1, 2; T = 200, NSIM = 100$;
o---o MLI-estimator, x---x ML-estimator.
Figure 5.8 Sample RMSE, $E((\pi_{i}^* - \pi_{i})^{2})^{1/2}$, $i = 1, 2$; $T = 200$, NSIM = 100; o- o MLI-estimator, x---x ML-estimator; solid curve = $E((p_{i} - \pi_{i})^{2})^{1/2}$.

Figure 5.9 Sample RMSE, $E((\pi_{ij}^* - \pi_{ij})^{2})^{1/2}$, $i, j = 1, 2$; $T = 200$, NSIM = 100; x---x ML-estimator.
As a whole, the choice of the true parameter values as a starting point in combination with the poor convergence properties will probably not affect the comparison between the two estimators when both behaved poorly. It was noted that it never happened that the MLI-estimator converged considerably more often than the ML-estimator, while in those cases where the opposite happened, the ML-estimator also was better than the MLI-estimator.

6. Final remarks

It has been emphasized by several authors that a mixed model should be applied to empirical data only when theoretical considerations give a clear indication of their relevance. On the other hand, once one has such good reasons for a mixed model, not much more is needed in order to apply a model with dependent regimes, which either alternate or form long runs. The assumption of an exact Markov dependence is then not crucial, but is merely used as the most simple tool to transfer information between successive observations. The transition probabilities \( \pi_{ij} = P(X_{t+1} = j | X_t = i) \) are of course defined for every stationary regime process.

There are several examples from the literature in which observations are collected sequentially in time, and where it is natural to suppose that adjacent regimes are either of the same or of different type. A worked example from the New York Stock Exchange is provided by Baum et al (1969). For unstable economic markets demand and supply equations are often used to describe price or production variation. Fair & Jaffee (1972) have proposed the use of two different regression equations for the number \( y_t \) of housing starts in month \( t \),
\[(19) \quad y_t = \alpha_1 + \frac{3}{4} \sum_{k=1}^{\infty} \beta_k u_{kt}^{(1)} + \varepsilon_t^{(1)},\]

\[(20) \quad y_t = \alpha_2 + \frac{4}{4} \sum_{k=1}^{\infty} \beta_k u_{kt}^{(2)} + \varepsilon_t^{(2)}.
\]

Here \(u_{1t}^{(1)} = u_{1t}^{(2)}\) is a common trend term, \(u_{2t}^{(1)}\) and \(u_{2t}^{(2)}\) reflect the mortgage rate, while \(u_{3t}^{(1)}\) describes the supply of old houses and \(u_{3t}^{(2)}\) and \(u_{4t}^{(2)}\) depend on the amount of money available to house builders. If there is an excess demand then the production is governed by the supply regime (20) while in case of an excess supply the demand regime (19) determines the starting of new houses.

Goldfeld & Quandt (1974) used a Markov chain model for successive demand or supply regimes and looked for the ML-estimator of the regression coefficients in (19) and (20). However, they did not use the correct loglikelihood function as given by (6), but instead

\[\sum_{t=1}^{T} \log \sum_{i=1}^{r} P(X_t = i)f_i(y_t, u_{it}^{(1)})\]

from which they estimated \(\pi_0 = (P(X_1 = i))_{i=1}^{r} \) and \(\pi = (\pi_{ij})\). This resembles the MLI-estimator in a model where the regimes are independent but not necessarily identically distributed. This procedure does not utilize any dependence structure between the regimes.

The iterative ML-procedure presented in Section 4.2 and 4.3 works also on the switching regression; see Example 4.3. However, reliable estimates should require more data than is usually available from monthly economic time series.
REFERENCES


Previous Statistical Research Reports
from the Institute of Mathematics and Statistics, University of Umeå

1971-1 "On necessary and sufficient conditions for B.L.U.E. in repeated sampling from finite populations" by Lennart Gustafsson
1971-2 "A Poisson process with marcovianly varying intensity with applications to process control" by Mats Rudemo
1971-3 "Estimation of the location and scale parameters of a Pareto distribution by linear functions of order statistics" by Gunnar Kulldorff and Kerstin Vännman
1971-4 "The consistency of the maximum likelihood estimates in finite Markov chains" by Bo Ranneby
1971-5 "State estimation for partially observed Markov chains with continuous time" by Mats Rudemo
1971-6 "Robustness of maximum likelihood estimates in a Markov chain model" by Bo Ranneby
1972-1 "Nonparametric estimation of the probability density function of a stationary mixing sequence" by Birger Bjureén
1972-2 "A recurrence formula for the multivariate Poisson distribution" by Olle Carlsson
1972-3 "The consistency and asymptotic normality of the maximum likelihood estimates from partially recorded samples" by Staffan Uvell
1972-4 "On the derivation of some discrete distributions from the hypothesis of elementary errors" by Olle Carlsson
1973-1 "Robustness of maximum likelihood estimates in a continuous time Markov chain model" by Bo Ranneby
1973-2 "Maximum likelihood estimates from partially grouped samples from independent not identically distributed random variables I" by Staffan Uvell
1973-3 "Markov decision processes for Markov dependent observations" by Bo Ranneby and Staffan Uvell
1973-4 "Maximum likelihood estimates from partially grouped samples from independent not identically distributed random variables II" by Staffan Uvell
1973-5 "On maximum likelihood estimates from progressively censored and progressively grouped samples" by Staffan Uvell
1973-6 "Constant-risk estimators with applications to linear functions of order statistics" by Kerstin Vännman
1973-7 "On queues in parallel" by Olle Carlsson
1973-8 "On a two-dimensional birth and death process in connection with migration" by Olle Carlsson
1974-1 "On queues in parallel" by Bengt Klefsjö
1974-2 "A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions" by Jayanti Chotai
1974-3 "On counteracting evidentiary mechanisms" by Hans Stenlund
1974-4 "Estimators based on order statistics from a Pareto distribution" by Kerstin Vännman
1974-5 "Maximum likelihood estimates from progressively censored samples" by Staffan Uvell
1974-6 "Maximum likelihood estimates from progressively grouped samples" by Staffan Uvell
1974-7 "A Monte Carlo study of some sampling designs" by Hans Stenlund and Anders Westlund
1974-8 "Simple estimators of the scale parameter of a Pareto distribution" by Kerstin Vännman

1975-1 "Prediction of catastrophes" by Georg Lindgren

1975-2 "A point estimator and an upper confidence bound for the scale parameter in a fetal monitoring system and other signal processing systems" by Åke Björnham and Dan-Eric Wiklund

1975-3 "Maximum likelihood estimation for discrete time Markov chains" by Bo Ranney

1975-4 "Maximum likelihood estimation for continuous time Markov chains" by Bo Ranney

1975-5 "Maximum likelihood estimates based on a general form of partial grouping" by Staffan Uvell

1975-6 "Maximum likelihood estimates from partially recorded samples" by Staffan Uvell

1975-7 "Maximum likelihood estimates based on partial grouping of type II" by Staffan Uvell

1975-8 "Definitions and some simple tests for heavy-tailed distributions" by Kerstin Vännman

1975-9 "Robustness against dependence of maximum likelihood estimates in i.i.d. models" by Bo Ranney

1975-10 "Frequency estimation by means of crossings of an observed mean level" by Åke Björnham and Georg Lindgren

1975-11 "On the optimum spacing of sample quantiles for estimating the location and scale parameters of a gamma distribution" by Lennart Rhodin

1976-1 "On the optimum spacing of sample quantiles for the ABLUE" by Lennart Rhodin

1976-2 "Functional limits of empirical distributions in crossing theory" by Georg Lindgren

1976-3 "Strong consistency of approximative maximum likelihood estimators" by Bo Ranney

1976-4 "An adaptive sequential decision procedure for analysis of FHR variability" by Åke Björnham and Dan-Eric Wiklund

1976-5 "A simulation model of FHR signals testing the properties of the ASD procedure" by Åke Björnham

1976-6 "On the optimum spacing of sample quantiles for estimating the location and scale parameters of a gamma distribution II" by Lennart Rhodin

1976-7 "Comparison of prediction intervals for failure times when life is exponential" by Kenneth S. Kaminsky

1976-8 "Admissibility of subset selection procedures" by Göran Broström

1976-9 "Optimum spacing for the ABLUE of location or scale parameters in Tukey's lambda distribution" by Lennart Rhodin

1976-10 "Comparison of asymptotically best linear unbiased predictors of order statistics" by Lennart Rhodin