PARAMETER ESTIMATION FOR A MIXTURE OF LINEAR REGRESSIONS

BY

RICHARD D. DE VEAUX

TECHNICAL REPORT NO. 247
APRIL 1986

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
PARAMETER ESTIMATION FOR A MIXTURE OF LINEAR REGRESSIONS

BY

RICHARD D. DE VEAUX

TECHNICAL REPORT NO. 247
APRIL 1986

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
TABLE OF CONTENTS

Chapter 1  Introduction and Historical Background  1

1.1  Introduction  1
1.2  Historical background  5

Chapter 2  Asymptotic Results  12

2.1  Introduction  12
2.2  The EM algorithm  12
2.3  Asymptotic properties  20
2.4  A $\sqrt{n}$-consistent starting point  33

Chapter 3  Starting values  37

3.1  Introduction  37
3.2  Mixture of normals  38
3.2.1  Method of moments  38
3.2.2  Minimum distance estimators  44
3.2.3  Bayes estimators  48
3.2.4  Conclusions  50
3.3  Mixture of regressions  51
3.3.1  A $\sqrt{n}$-consistent starting point  53
3.3.2  Other possible starting points  74
3.4  Conclusions  81

Chapter 4  Monte Carlo Simulations  83

4.1  Introduction  83
4.2  Mixture of normals simulations  83
4.3  Mixture of regressions simulation  100

Chapter 5  Data analysis  115

5.1  Introduction  115
5.2  The experiment  115
5.3  Data analysis  120
5.4  Conclusions  152

Bibliography  155
Parameter Estimation for a Mixture of Linear Regressions

Richard D. DeVeaux, Ph.D.
Stanford University, 1986

Abstract

A problem in music perception gives rise to a model where for each predictor $x$ the response $y$ is from the model $y_i = \alpha_j + \beta_j x_i + \epsilon_{ij}$, $i = 1, 2$ with probability $\lambda$ and $1 - \lambda$, respectively, where the $\epsilon_{ij}$ are iid unnormal $N(0, \sigma^2_j)$. As in the mixture of normals problem, the likelihood has no global maximum. It is shown that asymptotically efficient estimates of all seven parameters in the mixture of relationships problem can be found.
Chapter 1

§1.1 Introduction

Our concern in this dissertation is to estimate the parameters of a mixture of linear regressions. By a mixture of linear regressions we are referring to the following model. Let $x_1, \ldots, x_n$ be a non-random set of values, and let

\begin{equation}
(1.1.1) \quad y_i = \begin{cases} 
\alpha + \beta x_i + \epsilon_i & \text{with probability } \lambda \\
\gamma + \beta x_i + \epsilon_i & \text{with probability } 1-\lambda 
\end{cases}
\end{equation}

where $\epsilon_{j1} \sim N(0, \sigma_j^2)$ (j=1, 2; i = 1, \ldots, n). The problem is to estimate the seven parameters $\theta = (\lambda, \sigma_1, \sigma_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$.

This study was motivated by the data from an experiment on musical perception carried out at the Center for Research on Music and Acoustics (CRMA) at Stanford University. The details of this experiment are given in Chapter 5. A scatterplot of the data from one of the five subjects in the experiment is shown in Figure 1.1.

The two lines which seem to cross in this scatterplot are a common feature of the data of the five subjects, while the pattern is less obvious in some. The scatterplots of all five subjects are displayed in Chapter 5. The scientific question of interest is to estimate the slope and intercepts.
of both the lines as well as the variance of the errors

![Graph of JMS Full Data](image)

**Figure 1.1**

Data from JMS (One of five subjects)

around these lines. The experimenter hypothesized that given $x_i$, $y_1$ would be near 2.0 with some probability $\lambda$, and near $x_1$ with probability $(1-\lambda)$. For the purposes of our study, we
have generalized this model to the model of equation (1.1.1).

If we "collapse" the model (1.1.1) by assuming that the
slopes are zero, and all the data \( y_i \) are at the same \( x_i \),

(1.1.1) reduces to the following mixture of normals model.

Let \( y_1 (i=1,\ldots,n) \) be i.i.d with

\[
(1.1.2) \quad f(y_1) = \lambda f_1(y_1) + 1 - \lambda f_2(y_1),
\]

where \( f_j(y) = 1/(\sqrt{2\pi} \sigma_j) \exp{-1/2\{(y - \mu_j)/\sigma_j^2\}}, \ j=1,2. \)

The problem of estimating the five parameters \( \theta = (\lambda, \mu_1, - \mu_2, \sigma_1^2, \sigma_2^2) \) of (1.1.2) has a long history, dating back to
Karl Pearson. For this reason, as well as the fact that some of our results are best stated for the two models separately,
we will often treat the model (1.1.2) separately, rather than as a special case of (1.1.1). We will refer to model (1.1.1) as the mixture of regressions model, or the seven parameter case, while model (1.1.2) will be the mixture of normals model, or the five parameter case.

For the present discussion, we assume that we are interested in estimating all parameters with equal precision. That is, our implicit loss function gives the distance from each true parameters equal weights. This differs from the approach of several previous authors (e.g. Woodward et al (1984)) who consider \( \lambda \) to be the
parameter of interest, regarding the other four (or six) as nuisance parameters.

The structure of this dissertation is as follows. Chapter 1 contains an introduction as well as the historical background of the mixture of normals problem. Chapter 2 contains asymptotic results concerning both models. The main theoretical contribution of this dissertation is Proposition 2.3.4 which states that the EM algorithm, when started at a \( \sqrt{N} \)-consistent starting point, can be used to produce an asymptotically efficient estimator of \( \theta \) in both the five and (Proposition 2.3.6) seven parameter case. The main result in Chapter 3 is the proof that a proposed estimator of \( \theta \) for the regressions model (1.1.1) is \( \sqrt{N} \)-consistent. This is proved in Proposition 3.3.6. Also discussed in Chapter 3 are the practical problems of implementing various estimating procedures in both the five and seven parameter models. Chapter 4 contains our Monte Carlo simulation where we study the small sample properties of the methods recommended in Chapter 3. In the first part of Chapter 4, we study the five parameter case and compare the performance of the EM algorithm estimate to a minimum distance (MD) estimator. We compute the information matrix and compare the performance of the EM algorithm estimate to that predicted by asymptotic efficiency. We conclude that the EM exhibits desired small
sample properties, including to some degree robustness to choice of starting point. We also conclude, that as predicted by asymptotic efficiency, the EM outperforms the MD estimator. In the second part of Chapter 4, we carry out a similar small sample study of the EM algorithm estimator in the mixture of regression case. As no competitors currently exist for this estimator, we compare its variance to the variances from the information matrix, and to the variances of the least squares estimates that would be used if the population memberships were known. In Chapter 5, we describe in detail the experiment mentioned above, and use the EM algorithm to estimate the parameters for each subject. We discuss data analytic issues for each subject and for the experiment as a whole. We provide confidence intervals for the parameters via bootstrapped standard errors.

§1.2 Historical Background

Efforts to estimate the five parameters $\Theta = (\lambda, \mu_1, - \mu_2, \sigma_1^2, \sigma_2^2)$ of model (1.1.2) can be traced back to the nineteenth century, when biologists were attempting to estimate the mean forehead length of two visually indistinguishable species of crab.

Karl Pearson proposed a method of moments solution to the five parameter problem as early as 1894. We will
summarize his results. For further details see Pearson (1894). Without loss of generality, let \( \nu_1 \leq \nu_2 \). Then, let

\[
\begin{align*}
  m_1 &= (1 - \lambda) (\nu_1 - \nu_2) \quad (m_1 \leq 0) \\
  m_2 &= \lambda (\nu_2 - \nu_1) \quad (m_2 \geq 0)
\end{align*}
\]

(1.2.1)

Let \( \alpha_i \) be the \( i \)-th central moment of the mixture. Then,

\[
\begin{align*}
  \alpha_1 &= \lambda m_1 + (1 - \lambda)m_2 = 0 \\
  \alpha_2 &= \lambda (m_1^2 + \sigma_1^2) + (1 - \lambda)(m_2^2 + \sigma_2^2) \\
  \alpha_3 &= \lambda m_1 (m_1^2 + 3\sigma_1^2) + (1 - \lambda)m_2 (m_2^2 + 3\sigma_2^2) \\
  \alpha_4 &= \lambda (m_1^4 + 6m_1^2\sigma_1^2 + 3\sigma_1^4) + (1 - \lambda)(m_2^4 + 6m_2^2\sigma_2^2 + 3\sigma_2^4) \\
  \alpha_5 &= \lambda m_1 (m_1^4 + 10m_1^2\sigma_1^2 + 15\sigma_1^4) + (1 - \lambda)(m_2^4 + 10m_2^2\sigma_2^2 + 15\sigma_2^4)
\end{align*}
\]

(1.2.2)

The method of moments solution is obtained by substituting the first five sample central moments \( (\alpha_1) \) into the left side of the above equations and solving for \( \lambda, m_1, m_2, \sigma_1^2 \) and \( \sigma_2^2 \). Pearson showed that "after some algebra", the five equations can be reduced to the following nonic (ninth order) equation:
\[ 24z^9 + 84k_4z^7 + 36a_3^2z^6 + (90k_4 + 72a_3k_5)z^5 + (444k_4a_3^2 - 18k_5)z^4 + (288a_3^4 - 108a_3k_4k_5 + 27k_4^3)z^3 - (63a_3^2k_4^2 + 72a_3^3k_5)z^2 - 96a_3^4k_4z - 24a_3^6 = 0 \]

where \( k_4 = a_4 - 3a_2^2 \), and \( k_5 = a_5 - 10a_2a_3 \) are the 4th and 5th sample cumulants.

After finding the negative real root \( z^* \) of (1.2.3), if it exists, define

\[ r^* = \frac{-8a_3z^*^3 + 3k_5z^*^2 + 6a_3k_4z^* + 2a_3^3}{z^*(2z^*^3 + 3k_4z^* + 4a_3^2)}. \]

Then,

\[ \hat{m}_j = \frac{1}{2} \left( r^* + (r^* - 4z^*)^{1/2} \right), \]

and the parameter estimates are:

\[ \hat{\sigma}_j^2 = (\hat{m}_j/3) \left( 2r^* - a_3/z^* \right) + a_2 - \hat{m}_j^2, \]

(1.2.4) \[ \hat{\lambda} = \hat{m}_2 / (\hat{m}_2 - \hat{m}_1), \] and \[ \hat{\mu}_j = \hat{m}_j + \hat{\lambda}, \] \( j = 1, 2 \).

A. C. Cohen (1967) gave a modern interpretation of solution. If the simplifying assumption of equal variances can be made, he showed that the nonic equation (1.2.3) above can be reduced to a cubic, and thus that \( z \) can be found in closed form. For the non-equal variance case, he recommended starting with the cubic equation solution, and iterating to
the solution of the five equations rather than trying to solve the nonic directly.

Day (1969) investigated a maximum likelihood approach to finding the estimators of the five parameters. Let \{y_i\} \(i=1,\ldots,n\) be i.i.d with density (1.1.2). Let \(y\) represent the vector \(y_1,\ldots,y_n\). Then the log likelihood, \(\log L(\theta,y)\) (\(\log\) will be used throughout this dissertation to denote logarithm to the base \(e\)) can be written:

\[
(1.2.5) \quad \log L(\theta,y) = \Sigma \log\{\lambda f_1(y_1) + (1 - \lambda) f_2(y_1)\},
\]

where

\[
f_j(y) = \frac{1}{\sqrt{2\pi} \sigma_j} \exp -\frac{1}{2} \{(y - \mu_j)/\sigma_j^2\}, \quad j=1,2.
\]

He pointed out that if \(\sigma_1 \neq \sigma_2\), one can always make the likelihood unbounded by choosing any point \(y_1\) as the estimate for one of the \(\mu_1\) and 0 as the corresponding estimate of \(\sigma_1\). That is,

\[
(1.2.6) \quad L(\lambda, y_1, \mu_2, 0, \sigma_2^2) = \infty
\]

for any choice of the other three parameters, \(\lambda\), \(\mu_2\), and \(\sigma_2^2\).

Thus the likelihood function does not have a global
maximum (in either the five or seven parameter case). In this case, a maximum likelihood solution, in the strict sense does not exist. Day (1969) concluded that in the unequal variance case, there was little hope of using maximum likelihood techniques. He stated that "maximum likelihood clearly breaks down". We discuss this issue in Chapter 2.

For the equal variance case, \((\sigma_1^2 = \sigma_2^2 = \sigma^2)\), he proposed an interactive technique for finding the now four parameters by updating the conditional probability of population membership at each stage. Much more will be said of this idea in Chapter 2 when we introduce the EM algorithm.

In the 1970's several authors investigated, via Monte Carlo simulation, the possibility of applying maximum likelihood techniques to the estimation problem, despite the "singularity problem" of equation (1.2.6). Summarizing several papers (Tan and Chang (1972), Fryer and Robertson (1972), Hosmer (1973, 1974), Hosmer and Dick (1973)), it was found that in general there were no problems with convergence of iterative solutions to a singularity point \((\sigma_i = 0)\) for either the five or seven parameter case. However, they concluded that the starting point of the iteration was extremely crucial because of numerous local maxima and saddle points of the likelihood function.

On another track, Quandt and Ramsey (1978) proposed an
empirical moment generating function (mgf) technique. The mgf of the five parameter mixture can be written:

\[
E(e^{tY}) = \lambda \exp \left\{ \mu_1 + \left( t^2 \sigma_1^2 \right)/2 \right\} + (1 - \lambda) \exp \left\{ \mu_2 + \left( t^2 \sigma_2^2 \right)/2 \right\}
\]

(1.2.7)

They proposed choosing five points \( t_j \) \((j=1, \ldots, 5)\) in some interval \((a, b)\). Then, they let \( m_j \) be the sample mgf evaluated at \( t_j \),

\[
m_j = \left( \sum \exp \left\{ t_j y_j \right\} \right)/n,
\]

(1.2.8)

and solved for the five parameters by minimizing \( f(\theta, y) \), where,

\[
f(\theta, y) = \sum \left( m_j - E(\exp(t_j Y)) \right)^2
\]

(1.2.9)

The authors claimed that the method could easily be extended to the seven parameter problem, although the question of choosing the \( t_j \), \( a \), \( b \) and problems of minimizing the function \( f \) were left unanswered.

Recently, attempts have been made to use minimum distance estimators in the five parameter problem. Specifi-
cally, Woodward et al (1984) numerically minimize the Cramér-von Mises distance to obtain parameter estimates. We discuss this estimator in Chapter 3 and include it as a competitor to the EM algorithm estimator in Chapters 3 and 4.
Chapter 2

§2.1 Introduction

Although it has been shown that singularities in the likelihood equation exist for the mixture problem (see equation (1.2.6)), we show that these difficulties can in some sense be overcome. We show that if a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$ can be found for $\theta$ (where $\theta \in \mathbb{R}^5$ or $\mathbb{R}^7$), then this can be used as a starting value for the EM algorithm. The main contribution of this chapter is (Proposition 2.3.4), the proof that the EM algorithm will converge to a local maximum of the likelihood function, which will be shown to be an asymptotically efficient estimator for $\theta$.

§2.2 The EM algorithm

Throughout this chapter, we will consider mixture problems in general, and two mixture problems in particular. A general mixture problem in our context is a collection of random variables $\{y_i\}, i=1, \ldots, n$ with density

\begin{equation}
(2.2.1) \quad f(y_i) = \lambda f_1(y_i) + (1 - \lambda) f_2(y_i).
\end{equation}

In the five parameter mixture of normals problem, the $y_i$ are iid, with $\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \in \mathbb{R}^5$, and
(2.2.2) \[ f_j(y_1) = \frac{1}{\sqrt{2\pi}} \sigma_j \exp \left\{ -\frac{1}{2} \left( \frac{(y_1 - \nu_j)}{\sigma_j} \right)^2 \right\}. \]

The seven parameter, mixture of regressions problem is defined by data \((x_1, y_1)\), where conditional on \(x_1\), the \(y_1\) are independent with densities

(2.2.3) \[ f_j(y_1) = \frac{1}{\sqrt{2\pi}} \sigma_j \exp \left\{ -\frac{1}{2} \left( \frac{(y_1 - (\alpha_j + \beta_j x_1))}{\sigma_j} \right)^2 \right\}. \]

This is model (1.1.1) where as before, \( \theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, -\sigma_1^2, \sigma_2^2) \in \mathbb{R}^7 \).

We can view a general mixture problem as one of "incomplete data". That is, if we knew to which population each data point \(y_1\) belonged, we could use standard estimators to estimate the parameters. For example, in the five parameter case, we would use the sample means and variances of each population, while in the seven parameter case, we could use ordinary least squares estimators for \(\alpha_j, \beta_j, \) and \(\sigma_j^2 (j=1,2)\). We thus, following Dempster, Laird and Rubin (1977), consider the problem to be embedded in a larger sample space which includes knowledge of the membership of the \(y_1\). Let our original sample space be denoted by \(Y\). For each \(y_1\), consider the unobservable complete data pair \(\{y_1, z_1\}\). The \(z_1\) indicate


membership in population $i$. That is, $z_i = 1$ if $y_i \sim f_1$ and $z_i = 0$ if $y_i \sim f_2$. Let the pair $(y_i, z_i) = w_i$. Denote the larger sample space by $W$. (We are modifying somewhat the notation of Dempster, Laird and Rubin (1977)).

The density of the unobserved (or complete) data $f(w_i|\theta)$ can be written:

$$f(w_i|\theta) = f(y_i, z_i|\theta)$$

$$(2.2.4) \quad = f(y_i|z_i, \theta) f(z_i|\theta)$$

$$= \{\lambda f_1(y_i)\} z_i \{1-\lambda\} f_2(y_i) (1-z_i).$$

Since $E(z_i) = \lambda$, we have:

$$\int f(w_i|\theta) dw = \lambda f(y_i|\theta) + (1-\lambda) f(y_i|\theta)$$

$$= f(y_i|\theta) \text{ (or simply } f(y_i|\theta)).$$

Thus, the density of the observed data $y_i$, unconditional on knowing the membership $z_i$, can be written as the expectation of the density of the unobserved data $w_i$, where the integral is over the "complete" data space $W$. Let $y = (y_1, \ldots, y_n)$ denote the observed data vector. In vector notation, we can write:

$$(2.2.5) \quad f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

14
\[
\prod_{i=1}^{n} \int_{w} f(w \mid \theta) \, dw = \int_{w} f(w \mid \theta) \, dw.
\]

Let

(2.2.6) \[ k(w \mid y, \theta) = \frac{f(w \mid \theta)}{f(y \mid \theta)} \]

be the conditional density of \( w \) given \( y \) and \( \theta \). We denote the log likelihood of the sample by \( L(\theta) \). That is,

\[ L(\theta) = \log f(y \mid \theta). \]

Define two new functions,

(2.2.7) \[ Q(\theta' \mid \theta) = \mathbb{E} \{ \log f(w \mid \theta') \mid y, \theta \}, \]

and

(2.2.8) \[ H(\theta' \mid \theta) = \mathbb{E} \{ \log k(w \mid y, \theta') \mid y, \theta \}. \]

Then,

(2.2.9) \[ L(\theta') = Q(\theta' \mid \theta) - H(\theta' \mid \theta). \]
The idea of the EM algorithm is to iteratively maximize \( f(y) \) (equivalently \( L(y, \theta) \)), using the best guess of the population membership of the \( y_i \) (and thus the best guess of \( f(w|\theta) \)) at each step. See Dempster, Laird and Rubin (1977) for more details.

Explicitly, let \( \theta(p) \) be the estimate of \( \theta \) at the \( p \)th step of the EM algorithm. Then, from (2.2.7), we have

\[
Q(\theta | \theta(p)) = E \left[ \log f(w|\theta) \mid y, \theta(p) \right].
\]

The EM algorithm \( \theta(p) \rightarrow \theta(p+1) \) is defined by the following two steps. The "E-step" is to compute the expectation \( Q(\theta | \theta(p)) \). The "M-step" is to find \( \theta(p+1) \) which maximizes \( Q(\theta | \theta(p)) \).

For the mixture problem, from (2.2.4) and (2.2.5), we have:

\[
(2.2.10) \quad \log f(w|\theta) = \sum z_i \log f_1(y_i) + (1 - z_i) \log f_2(y_i) \\
\quad \quad \quad + \sum z_i \log \lambda + (1 - z_i) \log (1 - \lambda).
\]

Let

\[
W_{ij}(p) = E(z_i \mid y, \theta(p-1)),
\]

be the conditional probability of \( y_i \) belonging to population
$j$ based on the previous iteration estimates of $\theta$ and the observed data $y$. That is,

$$W_{1i}(p) = \frac{\lambda f_1(p-1)(y_i)}{\lambda f_1(p-1)(y_i) + (1 - \lambda) f_2(p-1)(y_i)}$$

(2.2.11)

$$W_{2i}(p) = \frac{(1 - \lambda) f_2(p-1)(y_i)}{\lambda f_1(p-1)(y_i) + (1 - \lambda) f_2(p-1)(y_i)},$$

where the superscript $(p)$ denotes that the expression is the estimate at the $p$th stage of the EM. For all parameters in the right hand side of the above equations we use the estimate at the $(p-1)^{st}$ iteration.

Thus the $E$-step for the mixture problem follows:

(2.2.12)

$$Q(\theta|\theta(p)) = \sum_{i} W_{1i}(p) \log f_1(p)(y_i) + W_{2i}(p) \log f_2(p)(y_i) + \sum W_{1i}(p) \log \lambda(p) + W_{2i}(p) \log (1 - \lambda(p)).$$

To complete the $M$-step, we maximize $Q$ with respect to $\theta$. This is done explicitly by setting the partial derivative with respect to each $\theta_j$ to 0 and solving for each $\theta_j$.

It follows that, in the five parameter case, the
parameter estimates at the \( p+1 \)st stage are:

\[
\lambda(p+1) = \frac{\sum W_{ij}(p)}{n},
\]

\[
\mu_j(p+1) = \frac{\sum W_{ij}(p) y_i}{\sum W_{ij}(p)}.
\]

\[
\sigma_j^2(p+1) = \frac{\sum W_{ij}(p) (y_i - \mu_j(p))^2}{\sum W_{ij}(p)}.
\]

Notice that these are just the usual Maximum Likelihood estimators weighted by the \( p \)th step estimates of \( W_{ij}(y_i) \), which are the conditional estimates of population membership of \( y_i \). This is essentially the idea to which Day (1969) referred for the equal variance case.

The seven parameter case is similar. The \( p+1 \)st step estimators are the usual ML(least squares) estimators weighted by the \( p \)th step estimates of the \( W_{ij} \). Explicitly,

\[
\lambda(p+1) = \frac{\sum W_{ij}(p)}{n},
\]

\[
\sigma_j(p+1) = \frac{\sum W_{ij}(p) y_i}{\sum W_{ij}(p)} - \beta_j(p+1) \frac{\sum W_{ij}(p) x_i}{\sum W_{ij}(p)},
\]
\[ \beta_{j}(p+1) = \frac{\sum W_{j}(p) x_{i} y_{i}}{\sum W_{j}(p)} - \frac{\sum W_{j}(p) x_{i} \sum W_{j}(p) y_{i}}{\sum W_{j}(p)} \]

\[ \sigma_{j}^{2}(p+1) = \frac{\sum W_{j}(p) (y_{i} - (\alpha_{j}(p) + \beta_{j}(p) x_{i}))^{2}}{\sum W_{j}(p)}. \]

In both problems, the solution is found by first providing an initial guess or starting point for the EM algorithm. We denote this by \( \theta^{(0)} \). The next step is started by computing the weights \( W_{j}(p) \) from (2.2.11). The step is completed by computing \( \theta^{(p+1)} \) explicitly from (2.2.13), (or (2.2.14)). One then cycles back and forth between (2.2.11) and (2.2.13) ((2.2.14)) until no more improvement (increase in \( L \)) is seen. The convergence of this algorithm is proved in the next section. We also discuss the asymptotic properties of the estimator defined as the limit of the EM algorithm sequence. For any sample size \( n \), we denote this limit by

\[ \theta^{*}_{n} = \lim_{p \to \infty} \theta^{(p)}_{n}. \]

We will show that \( \theta^{*}_{n} \) is asymptotically efficient for \( \theta \) as \( n \to \infty \). We have explicitly denoted the sample size \( n \) in
(2.2.15). When no confusion is likely to result, we will leave it out, and write the EM algorithm limit as $\theta^*$. 

§2.3 Asymptotic properties

We have seen that the EM algorithm can be implemented for both the five and seven parameter mixture problems. We have not, however, verified that the algorithm will converge, or if it does converge, whether the resulting estimator (2.2.15) has any desirable statistical properties. In this section we will show that the algorithm does converge (as $p$, the number of steps $\to \infty$) to a local maximum of the likelihood function and that this limit point is asymptotically efficient for estimating $\theta$ as the sample size $n \to \infty$.

We will use a theorem of Wald which states that under certain regularity conditions, there exist asymptotically efficient estimates of $\theta$ which are solutions of the likelihood equations. By a solution $\theta'$ of the likelihood equations we shall mean that

\[
\frac{\partial L(\theta)}{\partial \theta_j} \bigg|_{\theta}, = 0,
\]

for each $j$. This is also referred to as a stationary point of $L(\theta)$ in the EM literature. We shall state the regularity conditions which we shall need and the conditions on the $X$
matrix in the regression problem. We then refer to a Lemma of Kiefer (1978) which insures that the regularity conditions are satisfied for the mixture problems.

For simplicity, we now concentrate on the iid case, the mixture of normals problem. At the end of the section, we show that the arguments carry over to the regression case. We will assume the following conditions (following Lehmann (1983)):

(A0) The distributions $P_\theta$ of the observations are distinct.

(A1) The distributions $P_\theta$ have common support.

(A2) The observations $y_i$, $i=1, \ldots, n$ are iid with probability density $f(y_i, \theta)$.

(A3) The parameter space $\Theta$ contains an open interval $\omega$ of which the true parameter value $\theta_0$ is an interior point.

We will also assume the following regularity conditions:

(B1) For an open subset $\omega \in \Theta$ containing the true parameter $\theta_0$, for almost all $y$, the density $f(y, \theta)$ has all partial derivatives $(\partial^3/\partial \theta_j \partial \theta_k \partial \theta_l) f(y, \theta)$ for all $\theta \in \omega$.

(B2) The first and second logarithmic derivatives of $f$ are absolutely bounded, while the third are uniformly integrable.
(B3) The information matrix is positive definite, i.e.,

$$I_{jk}(\theta) = E_{\theta} \left[ - \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(y, \theta) \right]$$

is a positive definite matrix.

**Lemma 2.3.1**

The mixture density

$$f(y) = \lambda f_1(y) + (1 - \lambda) f_2(y)$$

where the $f_j$ ($j=1,2$) are as in equations (2.2.2) or (2.2.3), satisfies the regularity conditions (B1) - (B3).

**Proof:** Kiefer (1978).

QED.

We now invoke a theorem attributed to Wald which states that under the conditions above, there exist solutions of the likelihood equations which are consistent and asymptotically efficient for estimating $\theta = (\theta_1, \ldots, \theta_k)$. In particular, we use the version as stated in Lehmann (1983):

**Theorem 2.3.2**

Let $y_1, \ldots, y_n$ be iid each with density $f(y, \theta)$ satisfying assumptions (A0) - (A3) and (B1) - (B3). Then with probability tending to 1 as $n \to \infty$, there exist $\theta'_n$ which are solutions of the likelihood equations and such that for each $j$:
1) $\theta_1^n$ is consistent for estimating $\theta$.

2) $\sqrt{n}(\theta_1^n - \theta)$ is asymptotically normal with mean 0 and covariance matrix $[I(\theta)]^{-1}$ and,

3) $\theta_j^n$ is asymptotically efficient in the sense that $\sqrt{n}(\theta_j^n - \theta_j) \rightarrow N(0, [I(\theta)]^{-1}jj)$ in distribution. (That is, the $\theta_j^n$ have asymptotic variance equal to the Cramér-Rao lower bound.)

To find these solutions of the likelihood equations (local maxima) we will use the EM algorithm. We first state some general conditions under which the convergence of the EM algorithm (in $p$, the number of steps) is assured. These are due to Wu (1983). We then proceed to show that for large sample size, $(n \rightarrow \infty)$, starting the EM algorithm at a $\sqrt{n}$-consistent estimate guarantees that the limit point (in $p$) of the EM algorithm sequence is asymptotically efficient. Theorem 2 of Wu (1983) follows.

Theorem 2.3.3

Suppose the conditions below are satisfied. Then all limit points of any instance $\{\theta(p)\}$ of an EM algorithm are stationary points of $L$ and $L(\theta(p))$ converges monotonically to $L^* = L(\theta^*)$ for some stationary point $\theta^*$. The regularity conditions are:
i) $\Omega$ is a subset in $r$-dimensional Euclidean space $\mathbb{R}^r$,
ii) $\Omega_{\theta^0} = \{ \theta \in \Omega : L(\theta) \geq L(\theta^0) \}$ is compact for any $L(\theta^0) > -\infty$,
iii) $L$ in continuous and differentiable in the interior of $\Omega$,
iv) $Q(\theta^' | \theta)$ is continuous in both $\theta'$ and $\theta$ (separately). ($Q$ is defined as in (2.2.7)).

Of the four conditions in Theorem 2.3.3, the second one is the one which causes difficulties for the mixture problem. The four conditions together imply that the sequence \{\text{L}(\theta^{(p)})\}_{p \geq 0} is bounded above for all starting points $\theta^{(0)}$ in $\Omega$ such that $L(\theta^{(0)}) > -\infty$. This, as Day (1969) showed for the five parameter problem (see equation (2.2.6)) is not true for either problem. We have seen that if we include the boundaries of $\Omega$, i.e. \{ $\lambda = 0$ or 1; $\sigma_j = 0$, $j = 1$ or 2 \}, then the likelihood $L(\theta)$ may be unbounded.

Fortunately, we will not need such global convergence results. We will instead restrict our attention to small compact neighborhoods $S_\delta$ around $\theta_0$, the true parameter value. Then, since $\theta_0 \in \Omega_0$, condition ii) is satisfied for small $\delta$. Condition iii) follows from Lemma 2.3.1 and condition iv) follows from noting that $Q(\theta^{(p+1)} | \theta^{(p)})$ is a function of the mixture densities $f_1$ and $f_2$. The continuity
of \( Q \) then follows from the continuity of \( f_1 \) and \( f_2 \). Specifically, from (2.2.12),

\[
Q(\theta(p+1)|\theta(p)) = \sum_{i} W_{i1}(p) \log f_{1}(p)(y_{i}) + W_{i2}(p) \log f_{2}(p)(y_{i}) + \sum_{i} W_{i1}(p) \log \lambda(p) + W_{i2}(p) \log (1-\lambda(p)),
\]

where \( W_{j1}(p) \) and \( \lambda(p) \) as defined in (2.2.11) and (2.2.13), are smooth functions of \( f_j(y) \).

A weaker, but more general result than Theorem 2.3.3, from Dempster, Laird and Rubin (1977) (Theorem 1, p.7) states that the EM algorithm never decreases \( L(\theta) \). That is,

\[
(2.3.1) \quad L(\theta(p)) \geq L(\theta(p-1)).
\]

We will show that for the mixture problem, in certain neighborhoods \( S_\delta(\theta_0) \) around the true parameter value \( \theta_0 \), the likelihood \( L(\theta) \) is a strictly concave function, and thus the EM algorithm will stay on \( S_\delta(\theta_0) \) when started there. Applying Theorem 2.3.3, we will see that \( \{ \theta(p) \} \) converges to a stationary point of \( L(\theta) \). The key to being in such a neighborhood is to start at a \( \sqrt{m} \)-consistent estimator. We will use the following definition of \( \sqrt{m} \)-consistent throughout this dissertation.
Definition: An estimator $\hat{\theta}_n$ is $\sqrt{n}$-consistent for estimating $\theta$ if $\sqrt{n} (\hat{\theta}_n - \theta)$ is bounded in probability.

We are now in a position to state the main result of this chapter. If we start the EM algorithm at a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$ of $\theta$, the limit point of the EM algorithm will be asymptotically efficient for estimating $\theta$. The proof of this proposition is a bit technical and is perhaps better understood by the following figure and discussion.

![Figure 2.1](image_url)
Let $S_\delta(\theta_0)$ be the closed ball of radius $\delta$ centered at the true parameter value $\theta_0$. Let $\hat{\theta}_n$ be a $\sqrt{\frac{\delta}{n}}$-consistent estimator for $\theta$. (We show the existence of such an estimator in §2.4). Since $\hat{\theta}_n$ is $\sqrt{\frac{1}{n}}$-consistent for $\theta$, there will be some $n$ for which $\hat{\theta}_n$ is in the interior of $S_\delta(\theta_0)$ with arbitrarily large probability. The same is true of $\hat{\theta}'_n$, the likelihood solution whose existence is guaranteed by Theorem 2.3.2. We will show that for large $n$, the function $-L(\theta)$ is strictly convex on $S_\delta(\theta_0)$. Since the EM algorithm will be shown to converge on compact sets to a stationary point, by strict convexity, this limit must be $\hat{\theta}'_n$ which is asymptotically efficient. We now proceed to state and prove this claim.

**Proposition 2.3.4**

Let $y_i$, $i=1,\ldots,n$ be iid from a distribution $F_\theta$ satisfying (A0)-(A3) and with densities $f_j(\theta)$ ($j=1,2$) satisfying (B1)-(B3). Let $\{\theta(p)\}_{p \geq 0}$ be the EM algorithm sequence starting at a $\sqrt{\frac{\delta}{n}}$-consistent estimator $\hat{\theta}_n$. That is, $\theta(0) = \hat{\theta}_n$. Then $\theta_n^* = \lim_{p \to \infty} \theta(p)$, $p \geq 0$ is asymptotically efficient for estimating $\theta$.

**Proof:**

For any given $\delta > 0$, consider $S_\delta(\theta_0)$, the closed ball of
radius $\delta$ centered at the true parameter value $\theta_0$. $\theta_0 = 
(\theta_{01}, \ldots, \theta_{0m}) \in \mathbb{R}^m$. We restrict ourselves to $\delta$ such that $S_\delta(\theta_0) \subseteq \Omega^0$. In the interior of $S_\delta(\theta_0)$, $L(\theta)$ is three times differentiable by Lemma 2.3.1. By the same lemma, the information matrix $I(\theta)$ defined by

$$I_{jk}(\theta) = E \left[ -\partial^2 / \partial \theta_j \partial \theta_k \right] L(\theta) \right] \quad j, k = 1, \ldots, m$$

is positive definite.

Since $\theta_n$ is $\sqrt{n}$-consistent for $\theta$, there exists an $n_1(\epsilon, \delta)$ such that for all $n > n_1$, and for each $j$ ($j=1, \ldots, m$),

$$P(\sqrt{n} \mid \theta_{nj} - \theta_{0j} \mid > \delta) < \epsilon.$$ 

That is, with probability at least $1 - \epsilon$, $\theta_n$ is in the interior of $S_\delta(\theta_0)$.

By Theorem 2.3.2 there exists a likelihood solution $\theta'_n$ which is $\sqrt{n}$-consistent for $\theta$ and is a stationary point of $L(\theta)$. Therefore, for $n > n_2(\epsilon, \delta)$, for all $j$, 

$$P(\sqrt{n} \mid \theta'_{nj} - \theta_{0j} \mid > \delta) < \epsilon$$

$\theta'_n$ is thus also in the interior of $S_\delta(\theta_0)$ with probability at least $1 - \epsilon$. By Theorem 2.3.2, $\theta'_n$ is asymptotically
efficient in the sense that

$$\sqrt{n} (\theta'_{nj} - \theta_{0j}) \rightarrow N\{0, [I(\theta)]^{-1}jj\}$$

Recall that $[I(\theta_0)]$ is a positive definite matrix. By the law of large numbers,

$$-\frac{1}{n} \frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta_0) \rightarrow I_{jk}(\theta_0)$$

in probability. Thus, for any $\epsilon$, there exists $n_{jk}(\epsilon, \delta)$ such that

$$P\left( \left| \frac{1}{n} \frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta_0) - I_{jk}(\theta_0) \right| > \delta \right) < \epsilon$$

for all $n > n_{jk}(\epsilon, \delta)$. Let $n_3(\epsilon, \delta) = \max\{n_{jk}\}$. The eigenvalues of $[I(\theta_0)]$ are all positive, and thus there exists an $n_4$ such that, for all $n > n_4$, the eigenvalues of

$$[-\frac{1}{n} \frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta_0)]$$

are all positive (Wilkinson (1965) p. 93). For $n > n_4$, $-L(\theta)$ is strictly convex on $S_{\delta}(\theta_0)$. Now, let $n^* = \max\{n_i\}$, $i = 1,2,3,4$, and consider some particular $n > n^*$.

Let $\{\theta(p)\}_{p \geq 0}$ be the EM algorithm sequence which
starts at the $\sqrt{n}$-consistent estimator $\hat{\theta}_n$. (i.e. $\theta(0) = \hat{\theta}_n$).

The sequence $\{ L(\theta(p)) \}_{p \geq 0}$ is non-decreasing from Theorem 2.3.1. Since $-L(\theta)$ is strictly convex on $S_{\delta}(\theta_0)$ and since $L(\theta_n^*) \geq L(\hat{\theta}_n)$, $\theta(p)$ is in $S_{\delta}(\theta_0)$, for all $p$ and so $\theta_n^*$ is also in $S_{\delta}(\theta_0)$. Thus, by Theorem 2.3.3, $\theta_n^*$ is a stationary point of $L(\theta)$. Since $-L(\theta)$ is strictly convex, this stationary point must be the unique local maximum on $S_{\delta}(\theta_0)$, and the unique solution of $L(\theta)$. Thus, the limit of the EM algorithm, $\theta_{n}^* = \hat{\theta}_n$, the root (solution) of the likelihood $L(\theta)$, which is asymptotically efficient for estimating $\theta$.

QED.

To summarize, if we can find a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$ for $\theta$, we can use the EM algorithm to find a local maximum of the likelihood function near this point. By the above proposition, this local maximum, will be an asymptotically efficient estimator of $\theta$. We are thus strapped with the task of finding a $\sqrt{n}$-consistent estimator. In §2.4, we propose a $\sqrt{n}$-consistent starting point for the five parameter case. In §3.2, we propose one for the regression case. We now proceed to show that the above argument can be extended to the regression case.
Regression problem

The results of the previous subsection are not directly applicable to the seven parameter situation. Specifically, Theorem 2.3.2 and Proposition 2.3.4 require that the $y_j$ be iid, which is not true for the $y_j$ in the regression case. Recall that given $x_i$, the $y_j$ have density

$$\lambda f_1(y_j) + (1 - \lambda) f_2(y_j)$$

where

$$(2.3.2)$$

$$f_j(y_j) = \frac{1}{\sqrt{2\pi} \sigma_j} \exp \left\{ -\frac{1}{2} \left( \frac{(y_j - (\alpha_j + \beta_j x_i))}{\sigma_j} \right)^2 \right\}.$$ 

Let $X$ be the (usual) $2 \times n$ design matrix of $n 1$'s and the $x_i$, $i=1, \ldots, n$. For the rest of this dissertation, we will impose the following condition on the $x_i$.

(R1) $X$ is a $2 \times n$ nonstochastic matrix with the property that

$$\lim_{T \to \infty} \frac{XX'}{T} = Q \text{ is positive definite}.$$ 

Kiefer (1978) provides an extension of Theorem 2.3.2 for the regression case. Adapting the theorem to our notation, we have,
Theorem 2.3.5

Let \( x_i, y_i \) \( i = 1, \ldots, n \) be independent such that conditional on \( x_i, y_i \) have density \( \lambda f_1(y_i) + (1 - \lambda) f_2(y_i) \) where \( f_j(y_i) \) are as in (2.3.3) and the \( x_i \) satisfy (R1) above. Then there exists a unique, consistent estimator \( \theta'_n \) which is a solution of the likelihood equations. Further, \( \sqrt{n}(\theta'_n - \theta_0) \) is asymptotically normal with mean 0 and covariance matrix \( I(\theta_0)^{-1} \), where \( I(\theta_0) \) is the Fisher information matrix.


QED.

Using virtually the same argument as in the five parameter case, with Theorem 2.3.5 instead of Theorem 2.3.2, we have

Proposition 2.3.6

Let \( x_i, y_i \) \( i = 1, \ldots, n \) be independent, satisfying the conditions of Theorem 2.3.5. Let \( \hat{\theta}_n \) be a \( \sqrt{n} \)-consistent estimator of \( \theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2) \). Let \( \{ \theta(p) \}_{p \geq 0} \) be the EM algorithm sequence starting at \( \hat{\theta}_n \). Then \( \theta_n^* = \lim_{p \to \infty} \{ \theta(p) \}_{p \geq 0} \) is asymptotically efficient for estimating \( \theta \).

Proof: See Proposition 2.3.4.

QED.
§2.4 A $\sqrt{n}$-consistent starting point

In general, moment estimators are $\sqrt{n}$-consistent for estimating the moments of the distribution themselves. (For details see Serfling (1980), Theorem 2.2.1B). Recall, however, that the moment estimators we used (1.2.4) are used implicitly to find estimators for the mixture parameters. That is, the mixture parameters $\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, are functions of the moments of the mixture, the $\alpha_i$ of equation (1.2.2).

Suppose, as in (1.2.2) that we have used the five sample moments $a_i$ ($i=1, \ldots, 5$) to estimate the true population moments $\sigma_i$. The $a_i$ are $\sqrt{n}$-consistent for $\sigma_i$, but it is not immediately clear that the estimates $\hat{\theta}_n$ defined in (1.2.4), are $\sqrt{n}$-consistent for $\theta$. Proposition 2.4.1 guarantees that this is indeed the case. Additionally, it provides an explicit method of calculating the asymptotic covariance of $\hat{\theta}_n$ in terms of the asymptotic covariance of $a = (a_1, \ldots, a_5)$.

**Proposition 2.4.1**

Let $\sigma \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ be parameters. Suppose $T: \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ is twice differentiable where $T(\sigma, \theta) = 0$. If $a = (a_1, \ldots, a_m)$ is a $\sqrt{n}$-consistent estimator of $\sigma = (\sigma_1, \ldots, \sigma_m)$, then, the $\hat{\theta}_n$ defined implicitly by $T(a, \hat{\theta}_n) = 0$ are $\sqrt{n}$-
consistent for estimating $\theta = (\theta_1, \ldots, \theta_m)$. Additionally, if the $a_i$ have asymptotic covariance $\Sigma$ then $\hat{\theta}_n$ satisfy $\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow N(0, \Sigma V^T)$, where $V = [\partial T(\alpha, \theta)/\partial \theta]^{-1} [\partial T(\alpha, \theta)/\partial \alpha]$.

Proof:

Let $\epsilon > 0$. By $\sqrt{n}$-consistency of the $a_i$, for all $n > n^*(\epsilon)$, for some $\delta$, $P(\sqrt{n} \mid \mid a - \alpha \mid > \delta) < \epsilon$. Let this delta neighborhood of $\alpha$ in $\mathbb{R}^p$ be called $A(\epsilon)$. Since $T$ is twice differentiable, by the implicit function theorem, for every point $p \in A$, there exists a unique $g(p) \in B$, some neighborhood of $\theta$, such that $T(p, g(p)) = 0$, and $g$ is differentiable.

Thus $T(a, g(a)) = 0$. Let $\hat{\theta}_n = g(a)$. Since $g$ is differentiable, we have

$$P(\sqrt{n} \mid \mid g(a) - g(\alpha) \mid \mid > \delta) < \epsilon, \text{ or}$$

$$P(\sqrt{n} \mid \mid \hat{\theta}_n - g(\alpha) \mid \mid > \delta) < \epsilon.$$

Now since $T(\alpha, \theta) = 0$, by uniqueness of $g$, $g(\alpha) = \theta$. Thus,

$P(\sqrt{n} \mid \mid \hat{\theta}_n - \theta \mid \mid > \delta) < \epsilon$ and the $\hat{\theta}_n$ are $\sqrt{n}$-consistent.

Now, expand $g$ around $g(\alpha) = \theta$. For all $a \in A$,

$$g(a) = g(\alpha) + (a - \alpha) \frac{g'(\alpha)}{2} \mid a - \alpha \mid + O_p (\mid \mid a - \alpha \mid \mid).$$
We can rewrite this as

$$\hat{\theta}_n = \theta + (a - \alpha) \, g'(\alpha) + O_p (||a - \alpha||).$$

Hence,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}(a - \alpha) \, g'(\alpha) + O_p (\sqrt{n}||a - \alpha||).$$

By the implicit function theorem, (see eg. Spivak (1965) Theorem 2-12),

$$0 = [\partial T(a, \theta)/\partial \alpha] + [\partial T(a, \theta)/\partial \theta] \times [\partial g(\alpha)/\partial \theta].$$

Therefore,

$$g'(\alpha) \equiv \partial g(\alpha)/\partial \alpha$$

$$= - [\partial T(a, \theta)/\partial \theta]^{-1} \, [\partial T(a, \theta)/\partial \alpha].$$

And thus

$$\sqrt{n}(a - \alpha) \to N(0, \Sigma),$$

implies that

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \Sigma \Sigma^T),$$
where,

\[ V = - \left[ \frac{\partial \mathcal{T}(\alpha, \theta)}{\partial \theta} \right]^{-1} \left[ \frac{\partial \mathcal{T}(\alpha, \theta)}{\partial \alpha} \right]. \]

QED.

The moment estimators \( \widehat{\theta}_n \) defined implicitly are therefore also \( \sqrt{n} \)-consistent for estimating the parameters of the mixture, \( \theta \). This is proved in Proposition 3.2.1. We will investigate the possibility of using these and other starting values for the EM algorithm in the next chapter.
Chapter 3

§3.1 Introduction

We have seen that given a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$, of $\theta$ for the mixture problem (five or seven parameter), we can obtain an asymptotically efficient estimator of $\theta$. This is achieved by using $\hat{\theta}_n$ as a starting point for the EM algorithm $\{\theta^{(p)}\}_{p \geq 0}$. The limit point $\theta^* = \lim_{p \to \infty} \{\theta^{(p)}\}_{p \geq 0}$, where $\theta^{(0)} = \hat{\theta}_n$ is then asymptotically efficient. In this Chapter, we turn our attention to various proposals for such a $\sqrt{n}$-consistent starting point $\hat{\theta}_n$. It is here that the five and seven parameter problems need to be discussed separately. We first discuss (§3.2) the five parameter, mixture of normals problem. Here we show that the method of moments (MM), minimum distance (MD) and Bayes methods produce $\sqrt{n}$-consistent estimators for $\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, at least in theory. We discuss the practical problems of implementing and using each estimator as a starting point for the EM algorithm sequence. We also discuss the moment and minimum distance estimators as competitors of the EM estimator in their own right.

In the section §3.3, we consider the seven parameter, mixture of regressions problem. We find that none of the methods considered for the five parameter case can be
implemented for the regression case. Instead, we propose an ad hoc procedure for finding an EM starting point, based on "binning" the x-values into three bins. We show that it is in fact $\sqrt{n}$-consistent for estimating $\theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, -\sigma_1^2, \sigma_2^2)$, and thus that the limit of the resulting EM algorithm sequence will be asymptotically efficient for all seven parameters. The central contribution of this chapter is the presentation of this binning procedure and the proof of its $\sqrt{n}$-consistency.

§3.2 Mixture of Normals

We reviewed the history of the five parameter problem in Chapter 1. In this section, we discuss in turn, the methods of moments (MM) estimator, the minimum distance (MD) estimator and the Bayes estimator. For each, we show $\sqrt{n}$-consistency and discuss its implementation.

§3.2.1 Method of Moments

The method of moments solution to the mixture of normals problem was introduced in §1.2. Let $a_i$ denote (as in (1.2.2)) the $i^{th}$ central sample moment of the mixture. The $a_i$ are $\sqrt{n}$-consistent for $\alpha_i$, the $i^{th}$ central moment of the mixture (Serfling (1980) Theorem 2.2.1B). Let $\hat{\theta}_n$ denote the estimator of $\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, obtained from a sample of
size $n$, by equation (1.2.4). Then $\hat{\theta}_n$ is an implicit function of the $a_i$ which are $\sqrt{\mathcal{M}}$-consistent for $\sigma_i$. Thus $\hat{\theta}_n$ is $\sqrt{\mathcal{M}}$-consistent for $\theta$ by Proposition 2.4.1. Formally, we have

**Proposition 3.2.1**

Let $\hat{\theta}_n$ denote the estimators of $\theta = (\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ defined in equation (1.2.4), where $\lambda \in (0, 1)$ and $\sigma_1 > 0$. Then $\hat{\theta}_n$ is $\sqrt{\mathcal{M}}$-consistent for $\theta$. In addition, $\sqrt{\mathcal{M}} (\hat{\theta}_n - \theta)$ is asymptotically normal with mean 0 and variance $\Sigma \Sigma^T$ where $\Sigma = [\sigma_{ij}]$, $\sigma_{ij} = \sigma_{i+j} - \sigma_i \sigma_j$, and $\Sigma$ is the inverse of the 5x5 matrix of partial derivatives $[\partial T_i/\partial \theta_j]$, where $T: \mathbb{R}^5 \times \mathbb{R}^5 \to \mathbb{R}^5$, $T(\alpha, \theta) = 0$, is defined in equation (1.2.4). $T_i$ ($i=1, \ldots, 5$) denotes the $i$th coordinate function of $T$ (the $i$th line of equation (1.2.2)).

**Proof:**

The vector $a = (a_1, \ldots, a_5)$ is asymptotically normal with mean vector $\alpha$ and covariance $\Sigma$ (as above) by a direct application of the Lindeberg-Lévy Central Limit Theorem. (See e.g. Serfling (1980) p. 28). Let $T: \mathbb{R}^5 \times \mathbb{R}^5 \to \mathbb{R}^5$ be the function defined by subtracting the right hand side of equation (1.2.2) from the left. That is, $T(\alpha, \theta) = 0$. The estimator $\hat{\theta}_n$ is defined implicitly from the equation

39
\( T(a, \theta_n) = 0 \). It is easy to verify that \( \partial^2 T/\partial \theta^2 \) exists for \( \theta \) in the interior of \( \Omega \) (i.e. \( \lambda \in (0, 1), \sigma_i > 0 \)). Thus by Proposition 2.4.1, \( \hat{\theta}_n \) is \( \sqrt{n} \)-consistent for \( \theta \), and \( \sqrt{n} (\hat{\theta}_n - \theta) \) is asymptotically normal with mean 0 and variance \( \mathbf{VEV}^T \), where \( \mathbf{V} = [\partial T/\partial \theta]^{-1} \times [\partial T/\partial \alpha] \). Finally, from equation (1.2.2) it is seen that \( [\partial T/\partial \alpha] \) is the identity matrix, and thus \( \mathbf{V} = [\partial T/\partial \theta]^{-1} \).

QED.

By this theorem, asymptotic variances can be explicitly computed for the MM estimator \( \hat{\theta}_n \). This was carried out using MACSYMA on a VAX 750. The resulting expressions for the asymptotic covariance matrix was a closed form expression in terms of \( \theta \) and \( \alpha \). The expressions were not particularly informative. In addition, this estimator will be considered only as a starting point for the EM and not as a competitor. Thus, the asymptotic covariance matrix is not of particular interest, and for these reasons, is not included here.

Problems of Implementation

To find the method of moments estimate \( \hat{\theta}_n \), we must solve the ninth order equation (1.2.3). Specifically, we must find the single negative real root of (1.2.3) if it exists. (See Cohen (1969)). We have used Newton's method to find this
solution numerically. If a single negative real root is found, the parameter estimates are found via (1.2.4), and checked to see if they are feasible (i.e. $\lambda \in (0,1)$ and $\sigma_i > 0$ for $i=1,2$). If no feasible, negative real roots are found, there is no MM solution. If more than one negative root yields a feasible solution, a choice must be made among the solutions. This corroborates the findings of previous authors, specifically Quandt and Ramsey (1978) and Woodward et al (1984). In either case, we are unable to use a MM estimator. An alternative in these cases for our simulation, is the naive estimator

$$(3.2.1) \quad \hat{\theta} = (0.5, \bar{x} - s/2, \bar{x} + s/2, s/2, s/2),$$

where $\bar{x}$ and $s$ denote the sample mean and standard deviation of the mixed sample. Fowlkes (1978) and Woodward et al (1984) suggest several other possibilities. For all of these alternative estimators, however, $\sqrt{n}$-consistency is not guaranteed. As will be discussed in Chapter 4, for fairly high overlap situations, in up to 30% of the runs, it was not possible to find a MM estimate. However, for the overlap situations considered by Woodward et al (overlaps of .10 and .03), the MM solution was found in over 90% of the cases. This is markedly more frequent than Quandt and Ramsey (1978)
reported. Possible explanations are given in Chapter 4. When the MM estimator was not found, the naive estimator (3.2.1) was used as the EM starting point. Since this estimator is not $\sqrt{n}$-consistent, the resulting EM algorithm limit is not necessarily asymptotically efficient. Fortunately, we found that in virtually all runs where the MM estimator was found, the naive estimator resulted in the same EM algorithm limit. This is discussed in detail in Chapter 4. Similar insensitivity of the EM algorithm to various "reasonable" starting points was found by Woodward et al (1984). Thus, while starting values can be critical, (see especially Fowlkes (1978)), in our samples we found that the naive estimator worked well in the sense that it resulted in the same EM algorithm limit as the MM estimator, albeit in more steps. In practice, for a single data set, no doubt one would want to start the EM algorithm from several starting values. If different estimates resulted, one would have to base the choice of which EM algorithm limit to use on data analytic considerations.

We have, up to now, ignored questions of resistance to spurious data points and of robustness of efficiency. Since the MM estimator is based on five sample moments, one spurious data value (far from $\mu_1$ and $\mu_2$) could influence the MM estimate by an arbitrarily large amount. This is cer-
tainly a problem for the MM estimator and warrants further consideration. In our simulation study, however, we restricted our samples to mixtures where the components were in fact $N(\mu_i, \sigma_i^2)$, thus skirting the issue of resistance against outliers and robustness of efficiency.

We conclude that the MM solution, while $\sqrt{n}$-consistent, has limitations. First, equation (1.2.3) may not have exactly one negative real solution which results in a feasible estimate of $\Theta$. Secondly, if an MM estimate is found, it is subject to arbitrarily large influence from outlying data. Finally, in most situations, the naive estimator (3.2.1) while not $\sqrt{n}$-consistent, results in the same EM algorithm limit. The advantage of the MM method is that when the MM estimate is found, the EM algorithm converges more quickly from it than from the naive estimator. In addition, the resulting EM algorithm limit is necessarily asymptotically efficient by Proposition 2.3.4. More sophisticated numerical root searches may help with the first limitation. We found a significant decrease in the number of failures of the MM procedure in comparison with a previous study (Quandt and Ramsey (1978), by re-examining the case of multiple negative real roots. Another approach would be to numerically solve the system (1.2.2) as a whole, rather than solving the ninth order equation (1.2.3). The second,
resistance, could be helped perhaps by "robustifying" each sample moment. This approach was not followed. We return to the question of finding an MM estimator for the five parameter problem in §3.3 when we require initial five parameter estimates in the regression problem.

§3.2.2 Minimum Distance Estimators

Minimum distance estimators have been the subject of numerous articles in the statistical literature over the past few years. As evidence of this, see the recent bibliography of Parr (1981). For a fuller theoretical discussion of minimum distance methods we refer the reader to Millar (1981). Briefly, let \( \{F_\theta, \theta \in \Omega\} \) be a parametric family of distributions on the line, where \( \Omega \) is an open subset of \( \mathbb{R}^d \). Let the sample distribution of \( n \) i.i.d. observations \( x_i \) be denoted by \( F_n \). Consider a distance \( d(F,G) \) defined on the space of distribution functions. Roughly, a minimum distance estimator for \( \theta \) is an element of \( F_\theta, F_0 \) which lies closest to \( F_n \) in terms of \( d \). Recently, Woodward et al (1984) have investigated a minimum distance estimator for the mixture of normals situation. They have used the Cramér-von Mises distance:

\[
d(F,G) = \int [F(x) - G(x)]^2 \, dF(x)
\]
This leads to the following definition (following Woodward et al (1984)).

**Definition:** A minimum Cramér-von Mises distance estimator of $\theta$ is any value $\theta_n \in \Omega$ such that:

$$d(F_n, F_{\theta_n}) \leq \inf_{\theta} d(F_n, F_{\theta}) + 1/n$$

For a discussion of other possible distances see Millar (1981) and Pollard (1981). In the present context, let $\{F_{\theta}\}$ be the family of mixtures of two Normal distributions,

$$F_{\theta}(x) = \int_{-\infty}^{x} (\lambda f_1(t) + (1 - \lambda) f_2(t)) \, dt$$

where

$$f_j(x) \text{ is } N(\mu_j, \sigma_j^2) \text{ and } \theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2).$$

Woodward et al (1984) (p. 591) show that the Cramér-von Mises distance estimator can be iteratively found by finding $F_{\theta}$ which minimizes:

$$(3.2.2) \quad \frac{1}{12n} + \sum \left[F_{\theta}(x_{(i)}) - (i - 0.5/n)\right]^2$$

45
over all \( \theta \), where \( x(i) \) is the \( i \)th order statistic.

Let

\[
\eta_i(u) = \frac{\partial F(x)}{\partial \theta_i} \bigg|_{x = F^{-1}(u)}
\]

Define the 5x5 symmetric matrices \( A \) and \( B \) by

\[
A = \{a_{ij}\}, \quad B = \{b_{ij}\}
\]

where

\[
a_{ij} = \int_0^1 \eta_i(u)\eta_j(u) \, du,
\]

and

\[
b_{ij} = \int_0^1 \int_0^1 \{\min(u, v) - uv\} \eta_i(u) \, \eta_j(v) \, dudv.
\]

Then, Theorem 2 of Woodward et al (1984) states that the Cramér-von Mises estimator, \( \hat{\theta}_n \), is asymptotically normal with

\[
\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow N(0, A^{-1}BA^{-1}).
\]

Thus, we have immediately that:

**Proposition 3.2.2**

The Cramér-von Mises estimator defined as the element \( \hat{\theta}_n \) of \( \Omega \) which minimizes (3.2.2) is \( \sqrt{n} \)-consistent for \( \theta \).

**Proof:** Immediate from asymptotic normality.

QED.
Problems of Implementation

The minimum distance estimator considered in our simulation study is the estimator $\hat{\theta}_n$ for which $F_{\hat{\theta}_n}$ numerically minimizes (3.2.2). We have used a general minimization algorithm for minimizing a function of five variables based on the Davidon-Fletcher-Powell algorithm. Specifically, we computed numerical gradients in all directions at each stage $p$. The algorithm then took steps in the direction of the product of the gradient vector with a numerically updated Hessian matrix, until the function increased by an amount less than $10^{-5}$. The step sizes were determined from the performance of the previous stage. The Hessian matrix was upgraded after each stage. This algorithm was modified from the Davidon-Fletcher-Powell algorithm as described by Kennedy and Gentle (1980).

The problems of the MD estimator are twofold. First, the algorithm needs a starting point of its own. For this, we used the same starting point used for the EM algorithm. That is, we used the MM estimator if found, otherwise, the naive estimator (3.2.1). The second problem is computation time. The five dimensional search took nearly ten minutes on an IBM PC-AT (with 80287 math co-processor) for a sample of size 100. While this time could be reduced considerably on a much larger computer, the minimization search would still
involve an order of magnitude more CPU time than the convergence of the EM algorithm. Thus, while we will consider the MD estimator as a possible competitor to the EM estimator, we do not consider it as a possible starting point for the EM algorithm.

§3.2.3 Bayes Estimators

Under fairly broad regularity conditions, Bayes estimators are $\mathbb{M}$-consistent (see for example Diaconis and Freedman (1984)). We will consider a Bayes estimator for the mixture of normals case. Ferguson (1983) has proposed the following prior distribution on $\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$:

(a) $\lambda, \mu_1, \mu_2, \sigma_1^2$ and $\sigma_2^2$ are mutually independent

(b) $\lambda \sim \text{Beta}(M, 1)$ i.e. $f(\lambda) = M^{M-1} \Gamma[0, 1] (\lambda)$

(c) $(\mu_1, \sigma_1^2)$, $(\mu_2, \sigma_2^2)$ are i.i.d. with common distribution equal to the inverse gamma-normal conjugate prior for the normal distribution. That is, $\rho_i = 1/\sigma_i^2$ has the Gamma distribution, $\Gamma(\alpha, 2/\beta)$:

$$f(\rho_i) = [1/\Gamma(\alpha)](\beta/2)^{\alpha} e^{-\beta\rho_i/2} \rho_i^{\alpha-1} I(0, \infty) (\rho),$$

and given $\rho_i$,

$$\mu_i \sim \mathcal{N}(\mu, 1/\rho_i \tau).$$

48
Then, conditional on these parameters, the data have density:

\[
f(y_1, y_2, \mu_1, \mu_2, \rho_1, \rho_2, \lambda) = 
\lambda \left( \rho_1 \right)^{1/2} \frac{2\pi}{2} \exp \left\{ -\rho_1/2 \left( y - \mu_1 \right)^2 \right\} 
(1-\lambda) \left( \rho_2 \right)^{1/2} \frac{2\pi}{2} \exp \left\{ -\rho_2/2 \left( y - \mu_2 \right)^2 \right\}.
\]

The joint density

\[
f(y, \mu_1, \mu_2, \rho_1, \rho_2, \lambda) = 
K \ M \lambda^{M-1} \exp \left\{ -(\rho_1+\rho_2) \beta/2 \right\} \rho_1^{a-1} \rho_2^{a-1} 
x \left( \rho_1 \rho_2 \right)^{1/2} \exp \left\{ -(\rho_1+\rho_2) \left( \mu_1 - \mu \right)^2 \right\} 
\times \prod_{i=1}^{n} \left\{ \lambda \phi(\sqrt{\beta} (y_i - \mu)) + (1-\lambda) \phi(\sqrt{\beta} (y_i - \mu)) \right\}
\]
where \( \phi \) denotes the standard normal density. Now,

\[
\prod_{i=1}^{n} \left\{ \lambda \phi(\sqrt{\beta} (y_i - \mu)) + (1-\lambda) \phi(\sqrt{\beta} (y_i - \mu)) \right\}
\]

\[
= \sum_{j=0}^{n} \lambda^j (1-\lambda)^{n-j} \prod_{i \in S_j} \phi(\sqrt{\beta} (y_i - \mu)) \phi(\sqrt{\beta} (y_i - \mu))
\]
where the sum is over all sets of indices \( S_j \) contained in the set \( \{1, \ldots, n\} \).

**Problems of Implementation**

This density is impossible to evaluate for all but very small \( n \). Ferguson (1983) has proposed a Monte Carlo method for sampling the set of indices and thus approximating the
density, but this does not seem reasonable for the sample sizes which we will encounter. In addition, there is no guarantee that the resulting estimators will be $\sqrt{N}$-consistent. For these reasons, we will not pursue Bayes estimators for either the mixture of normals or the mixture of regression case.

§3.2.4 Conclusions

We have shown that the method of moments, minimum distance and Bayes procedures provide $\sqrt{N}$-consistent estimators for $\theta = (\lambda, \nu_1, \nu_2, \sigma_1^2, \sigma_2^2)$ in theory. Practical considerations result in the use of only the MM estimator and a naive estimator (3.2.1) as possible starting points for the EM algorithm, and the MD estimator as a possible competitor to the EM algorithm estimator. While questions of resistance and robustness cast some doubts on the reliability of the MM estimator in practice, the EM algorithm seems fairly robust to starting values.

Previous authors (notably Woodward et al (1984)), have conducted Monte Carlo experiments to compare MD, maximum likelihood and other ad hoc estimators in the five parameter case. Their focus was to study the improvement of the performance of the MD estimator in comparison with the EM estimator in non-normal situations. Our need to study the
five parameter case is based on our need for an initial five parameter estimate in the regression case. Our focus for the five parameter case will be to study the relative efficiency of the EM estimator to the Cramér-Rao lower bounds and to the MD estimator in small samples. We are also concerned with qualitatively studying robustness to starting values, due to the problems with the MM estimator. We also include high overlap conditions not previously studied (details in Chapter 4). We use sample size \( n = 100 \) for comparison with other studies.

§3.3 Mixture of Regressions

We now consider data \( (x_1, y_1) \ i=1, \ldots, n \) where as in equation (1.1.2),

\[
\begin{cases}
\alpha + \beta_1 x_1 + \epsilon_1 & \text{with probability } \lambda \\
\alpha + \beta_2 x_1 + \epsilon_2 & \text{with probability } 1 - \lambda
\end{cases}
\]  

(3.3.1)

where \( \epsilon_{ji} \) are independent \( N(0, \sigma_j^2) \); \( j=1,2; \ i=1,\ldots,n \). A scatterplot of data with \( \theta = (0.5, 0, 1, 1, -1, 1^2, 2^2) \) is shown in Figure 3.1. We will assume throughout that the \( x_i \) satisfy condition (R1) of Chapter 2. We present an ad hoc procedure for finding a \( \sqrt{\mathcal{H}} \)-consistent estimator for \( \theta = (\lambda, \sigma_1, \sigma_2, \beta_1, -\beta_2, \sigma_1^2, \sigma_2^2) \). The EM algorithm sequence starting at this
estimator will be asymptotically efficient by proposition 2.3.6. We consider using method of moments, minimum distance and Bayes methods to provide a $\sqrt{n}$-consistent starting point, but conclude that none can be implemented in practice.

![Scatterplot of $y$ vs $x$](image)

**Figure 3.1 Scatterplot of Simulated Data**

$\theta = (.5, 0, 1, 1, -1, .1^2, .2^2)$
§3.3.1 A $\sqrt{n}$-consistent starting point

In this section, we propose a method for obtaining a starting point $\hat{\theta}_n$ for the EM algorithm. We show that $\hat{\theta}_n$ is $\sqrt{n}$-consistent for $\theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ and thus by Proposition 2.3.6, the limit $\Theta_n$ of the EM algorithm sequence $\{\theta(p)_n\}_{p \geq 0}$ where $\theta(0)_n = \hat{\theta}_n$ will be asymptotically efficient. We describe this method heuristically before proceeding to prove our results.

The procedure is started by splitting the $x$-axis into three disjoint, connected intervals, $I(j)$, $j = 1, 2, 3$, which we shall call bins. We choose the bins so that each contains a fixed $100p\%$ of the $N x_i$ points, resulting in $k = pN$ points in each bin. We now ignore the dependence of the response $y$ on $x$, and estimate the five parameters of the responses $y_i$ in each bin as if they were from a one-dimensional mixture of normals. That is, we replace the $(x_i, y_i)$ pairs by the points $(\bar{x}, y_i)$ in each bins. This is shown in Figure 3.2. The five parameters we denote by $\theta(j) = (\lambda, \mu_1(j), \mu_2(j), \sigma_1^2(j), \sigma_2^2(j))$. Here we have chosen $I(1) = [0.0, 0.2)$, $I(2) = (0.8, 1.0]$ and $I(3) = (0.6, 0.8]$. Each contains $k = 20$ points. By the results of Chapter 2, we can find $\sqrt{k}$-consistent estimates for each $\theta(j)$.

Consider the first bin, $[0.0, 0.2)$ in Figure 3.2. The
means that we have estimated in this bin we denote \( \hat{\mu}_1^{(1)} \) and \( \hat{\mu}_2^{(1)} \). The lower mean, \( \mu_1^{(1)} \), is equal to say \( \alpha_1 + \beta_1 \bar{x}^{(1)} \), while the other, \( \mu_2^{(1)} \) is equal to \( \alpha_2 + \beta_2 \bar{x}^{(1)} \). By estimating these means again in a second interval, we obtain \( \hat{\mu}_1^{(2)} \) and \( \hat{\mu}_2^{(2)} \). (See figure 3.2). Again, one of these estimates will estimate \( \alpha_1 + \beta_1 \bar{x}^{(2)} \), while the other will estimate \( \alpha_2 + \beta_2 \bar{x}^{(2)} \). We now have four estimates \( \hat{\mu}_1^{(1)}, \hat{\mu}_2^{(1)}, \hat{\mu}_1^{(2)}, \text{ and } \hat{\mu}_2^{(2)} \) and four parameters \( \alpha_1, \alpha_2, \beta_1 \text{ and } \beta_2 \). In our example the four mean estimates are \((.12, .94, .01, .92)\). These are indicated in Figure 3.2.

The problem is that we have two choices. We could associate both the means which are larger and the means which are smaller - resulting in a pair of lines which do not cross between the two intervals, or associate the larger mean in one with the smaller mean in the other and vice versa - resulting in a pair of lines which do cross in one of the intervals. The intercepts and slopes of the first pair, which we will refer to as the "straight" pair are denoted by \((\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2)\). In our example, via (3.3.4) and (3.3.6), \((\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2) = (.13, .94, -.14, -.02)\). This pair is shown in Figure 3.3. The intercepts and slopes of the other pair, which we will refer to as the "crossed" pair are denoted by \((\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2)\) = \((.03, 1.05, .99, -1.15)\) in our example. This pair of lines is shown in figure 3.4.
Figure 3.2 Three Bins.

Bin 1 = [0.0, 0.2); Bin 2 = (0.8, 1.0]; Bin 3 = (0.6, 0.8]
Figure 3.3

The straight pair.
To choose between these two pairs of lines, we will use the estimates of $\mu_1^{(3)}$ and $\mu_2^{(3)}$ from the third bin. For each of the four lines, we find the predicted value of $y$ at $\bar{x}^{(3)}$. For the straight pair, denote the smaller predicted $y$ value by $y_{1s}$ and the larger by $y_{2s}$. Similarly define $y_{1c}$ and $y_{2c}$. (For example, $y_{1s} = \min(\hat{a}_{1s} + \hat{b}_{1s} \bar{x}^{(3)}, \hat{a}_{2s} + \hat{b}_{2s} \bar{x}^{(3)})$. In our example, $y_{1s} = .04$, $y_{2s} = .93$, $y_{1c} = .24$ and
$y_{2c} = .72$. See Figures 3.3 and 3.4. Now compute the
distances from $\hat{\nu}_1^{(3)}$ to $y_{1s}$ and $y_{1c}$, and from $\hat{\nu}_2^{(3)}$ to $y_{2s}$
and $y_{2c}$. Define the total distance from $\nu_1^{(3)}$ and $\nu_2^{(3)}$ to
each pair by:

$$d_s = | \hat{\nu}_1^{(3)} - y_{1s} | + | \hat{\nu}_2^{(3)} - y_{2s} |,$$

and

$$d_c = | \hat{\nu}_1^{(3)} - y_{1c} | + | \hat{\nu}_2^{(3)} - y_{2c} |.$$

We will choose the pair which has the smaller of $d_s$ and $d_c$.
This is shown in Lemma 3.3.3 to be asymptotically the correct
choice. In our example, since $\hat{\nu}_1^{(3)} = .14$ and $\hat{\nu}_2^{(3)} = .66$,
$d_s = .10 + .27 = .37$ while $d_c = .10 + .06 = .16$. Thus, the
crossed pair is chosen, and the starting estimates for
$(\alpha_1, \alpha_2, \beta_1, \beta_2)$ are $(\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c}) = (.03, 1.05, .99, -1.15)$.

For $\lambda$, we will use the $\sqrt{N}$-consistent estimator of $\lambda$ from
any of the $\theta^{(j)}$. For the variances, we refer the reader to
Lemmas 3.3.4 and 3.3.5. We now proceed to show in a series
of Lemmas that this binning procedure results in a $\sqrt{N}$-
consistent estimator $\hat{\theta}_n$, for the seven parameter $\theta =
(\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$. Then, we summarize the argument in
Proposition 3.3.6, where we show that the EM limit $\theta^*_n$ which
starts at $\hat{\theta}_n$ will be asymptotically efficient.
Lemma 3.3.1

Let \((x_i, y_i)\) \((i=1, \ldots, n)\) satisfy the model of equation (3.3.1). Let I be an interval on the x-axis containing k of the \(x_i\). Assume that in I the regression lines do not cross. That is, \(\max (\alpha_1 + \beta_1 x_i) < \min (\alpha_2 + \beta_2 x_i)\) on I. Let \(\theta_n\) be a \(\sqrt{n}\)-consistent estimator of \(\theta = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)\) where \(y_i\) is considered to be from a mixture,

\[
(3.3.2) \quad y_i = \begin{cases} 
\mu_1 + e'_{1i} & \text{with probability } \lambda \\
\mu_2 + e'_{2i} & \text{with probability } 1 - \lambda.
\end{cases}
\]

where the \(e'_{ji}\) are independent with \(E(e'_{ji}) = 0\) and \(\text{Var}(e'_{ji}) = \sigma_j^2\), \(j=1,2; i=1,\ldots,k\). Then \(\theta_n\) is also a \(\sqrt{n}\)-consistent estimator, for \(\alpha_j + \beta_j \bar{x}\), \(j = 1, 2\), where \(\bar{x} = \Sigma x_i / k\) (sum over \(x_i \in I\)).

Proof:

From (3.3.1),

\[
(3.3.1) \quad y_i = \begin{cases} 
\alpha_1 + \beta_1 x_i + e'_{1i} & \text{with probability } \lambda \\
\alpha_2 + \beta_2 x_i + e'_{2i} & \text{with probability } 1 - \lambda
\end{cases}
\]

Thus,

59
\[ y_i = \begin{bmatrix} \alpha_1 + \beta_1 \bar{x} + \beta_1 (x_i - \bar{x}) + \epsilon_{1i} \quad \text{w. p. } \lambda \\ \alpha_2 + \beta_2 \bar{x} + \beta_2 (x_i - \bar{x}) + \epsilon_{2i} \quad \text{w. p. } 1 - \lambda \end{bmatrix} \]

where \( \epsilon_{ji} \sim N(0, \sigma_j^2) \) and \( \bar{x} = \sum x_i / k \).

Define \( \mu_j = \alpha_j + \beta_j \bar{x} \), and \( \epsilon'_{ji} = \beta_j (x_i - \bar{x}) + \epsilon_{ji} \).

Then,

\[ y_i = \begin{bmatrix} \mu_1 + \epsilon'_{1i} \quad \text{with probability } \lambda \\ \mu_2 + \epsilon'_{2i} \quad \text{with probability } 1 - \lambda \end{bmatrix} \]

where \( \epsilon'_{ji} \) are independent; \( \Sigma(\epsilon'_{ji}) = 0; \ \text{Var}(\epsilon'_{ji}) = \sigma_j'^2 \).

This is now a five parameter mixture problem, with \( \theta = (\lambda, \mu_1, \mu_2, \sigma_1'^2, \sigma_2'^2) \). Although the \( \epsilon' \) are no longer normal, the Method of Moments still produces a \( \sqrt{k} \)-consistent estimator for the true moments of the \( y_i \). By proposition 2.4.1, the implicit solutions are \( \sqrt{k} \)-consistent for \( \theta \). Since \( k = pN \), a fixed percentage of \( N \), \( O(k) = O(n) \), and these estimators are \( \sqrt{n} \)-consistent. The means \( \mu_j \) which are estimated are \( \alpha_j + \beta_j \bar{x} \). The variances \( \sigma_j'^2 \) are larger than the actual residual variance \( \sigma_j^2 \) since they contain the term \( \beta_j(x_i - \bar{x}) \). Thus the variances do not directly correspond, but \( \theta_K \) does provide a \( \sqrt{n} \)-consistent estimator for \( (\lambda, \alpha_1 + \beta_1 \bar{x}, \alpha_2 + \beta_2 \bar{x}, \sigma_1'^2, \sigma_2'^2) \).
For the present, we will ignore the variance estimates, and consider only \((\lambda, \mu_1(j), \mu_2(j))\) in each interval \(I(j)\). Let \(I(1)\) and \(I(2)\) be two intervals containing \(k\) of the \(x\)-values. Let \(\hat{\theta}(1)\) and \(\hat{\theta}(2)\) be the \(\sqrt{n}\)-consistent estimators of \(\theta(3.3.3)\)

\[
\hat{\theta}(1) = (\lambda, \mu_1(1), \mu_2(1)),
\]

and

\[
\hat{\theta}(2) = (\lambda, \mu_2(2), \mu_2(2)).
\]

We have the two sets of estimators. The "straight" pair of lines estimators are:

\[
\hat{\beta}_{1s} = (\hat{\mu}_1(2) - \hat{\mu}_1(1))/(\bar{x}(2) - \bar{x}(1)),
\]

\[
\hat{\beta}_{2s} = (\hat{\mu}_2(2) - \hat{\mu}_2(1))/(\bar{x}(2) - \bar{x}(1))
\]

which would be the correct choice if the true regression lines did not cross between the intervals, and the "crossed" pair of lines estimators:

\[
\hat{\beta}_{1c} = (\hat{\mu}_2(2) - \hat{\mu}_1(1))/(\bar{x}(2) - \bar{x}(1)),
\]

\[
\hat{\beta}_{2c} = (\hat{\mu}_1(2) - \hat{\mu}_2(1))/(\bar{x}(2) - \bar{x}(1))
\]

which are correct if the lines do cross.

The corresponding intercept estimators are:
\[\hat{\alpha}_{1s} = \hat{\nu}_1(1) - \beta_{1s} \bar{X}(1)\]
\[\hat{\alpha}_{2s} = \hat{\nu}_2(1) - \beta_{2s} \bar{X}(1)\]

and

\[\hat{\alpha}_{1c} = \hat{\nu}_1(1) - \beta_{1c} \bar{X}(1)\]
\[\hat{\alpha}_{2c} = \hat{\nu}_2(1) - \beta_{2c} \bar{X}(1)\]

The two estimated pairs of lines are shown in Figures 3.3 and Figure 3.4. We now show that one set of estimators is \(\sqrt{m}\)-consistent.

**Lemma 3.3.2**

Under the conditions of Lemma 3.3.1, one set of estimators \((\hat{\alpha}_{1s}, \hat{\alpha}_{2s}, \hat{\beta}_{1s}, \hat{\beta}_{2s})\) or \((\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c})\) defined above is \(\sqrt{m}\)-consistent for \((\alpha_1, \alpha_2, \beta_1, \beta_2)\).

**Proof:**

Assume WLOG that the true parameters \((\alpha_1, \alpha_2, \beta_1, \beta_2)\) are such that the lines do not cross between \(I(1)\) and \(I(2)\). By definition, \(\nu_i(j) = \alpha_i + \beta_i \bar{X}(j)\). By Lemma 3.3.1, the \(\sqrt{m}\)-consistent MM estimator \(\hat{\nu}_1(1)\) is \(\sqrt{m}\)-consistent for either \(\alpha_1 + \beta_1 \bar{X}(1)\) or \(\alpha_2 + \beta_2 \bar{X}(1)\). Since, by assumption, the true lines do not cross, this must be \(\alpha_1 + \beta_1 \bar{X}(1)\). Thus,
\[
\hat{\beta}_{2s} = \frac{(\hat{\nu}_2(2) - \hat{\nu}_2(1))}{(\bar{x}(2) - \bar{x}(1))} \\
= \frac{(\nu_2(2) - \nu_2(1))}{(\bar{x}(2) - \bar{x}(1))} + O_p(1/\sqrt{n}) \\
= \frac{(\alpha_2 + \beta_2 \bar{x}(2) - (\alpha_2 + \beta_2 \bar{x}(1))}{(\bar{x}(2) - \bar{x}(1))} + O_p(1/\sqrt{n}) \\
= \beta_2 + O_p(1/\sqrt{n}),
\]

implies that \(\hat{\beta}_{2s}\) is \(\sqrt{n}\)-consistent for \(\beta_2\). Similarly, \(\hat{\beta}_{1s}\), \(\hat{\alpha}_{2s}\), and \(\hat{\alpha}_{1s}\) are \(\sqrt{n}\)-consistent for their respective parameters. If the true lines do cross, then \(\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}\) and \(\hat{\beta}_{2c}\) are \(\sqrt{n}\)-consistent by a similar argument.

GED.

We are now left with two problems. The first is how to choose which set of lines provides the \(\sqrt{n}\)-consistent estimates of \((\lambda, \alpha_2, \alpha_2, \beta_2, \beta_2)\). The second is to estimate the variances \(\sigma_j^2\) of the regression errors \(e_{ij}\). Recall that the variances estimated in Lemma 3.3.1 include the sum of squares due to regression, and are thus too large.

The solution to the first problem is to estimate the means of the \(y\)'s in a third interval \(I(3)\), by \(\hat{\nu}_1(3)\) and \(\hat{\nu}_2(3)\). The straight pair predicts two values for \(y\) at \(\bar{x}(2)\): \(y_{1s} = (\hat{\alpha}_{1s} + \hat{\beta}_{1s} \bar{x}(3))\) and \(y_{2s} = (\hat{\alpha}_{2s} + \hat{\beta}_{2s} \bar{x}(3))\), while the crossed lines predict them at
\[ Y_{1c} = \min(\hat{\alpha}_{1c} + \hat{\beta}_{1c} \bar{X}(3), \hat{\alpha}_{2c} + \hat{\beta}_{2c} \bar{X}(3)) \]

and

\[ Y_{2c} = \max(\hat{\alpha}_{1c} + \hat{\beta}_{1c} \bar{X}(3), \hat{\alpha}_{2c} + \hat{\beta}_{2c} \bar{X}(3)). \]

We choose the straight or crossed pair based on the distance from the mean estimates \( \hat{\mu}_j(3) \) to these two pairs of lines. Explicitly, define the distance of the straight pair of lines to the estimated means as

\[ d_s = d_{s1} + d_{s2}, \]

where

\[ (3.3.8) \quad d_{s1} = |y_{1s} - \hat{\mu}_1(3)| \]

and

\[ d_{s2} = |y_{2s} - \hat{\mu}_2(3)|. \]

Define \( d_c = d_{c1} + d_{c2} \) where \( d_{c1} \) and \( d_{c2} \) are defined by replacing \( s \) with \( c \) in (3.3.8). We use the \( s \) or \( c \) estimates based on the smaller of \( d_s \) or \( d_c \). This results (with probability tending to 1 as \( n \to \infty \)) in the correct choice. This is proved in the following lemma.

**Lemma 3.3.3**

Under the conditions of Lemma 3.3.1, choose three intervals \( I(j), j=1,2,3 \). Let \( (\hat{\alpha}_{1s}, \hat{\alpha}_{2s}, \hat{\beta}_{1s}, \hat{\beta}_{2s}) \) and \( (\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c}) \) be the estimates defined from (3.3.4) through (3.3.7) from intervals \( I(1) \) and \( I(2) \). Let \( \hat{\mu}_1(3) \) and \( \hat{\mu}_2(3) \) be the estimates from the mixture model (3.3.2) in \( I(3) \). Then
the set of estimates which minimizes \((d_s, d_c)\) defined in (3.3.8) provides \(\sqrt{n}\)-consistent estimates of \((\alpha_1, \alpha_2, \beta_1, \beta_2)\) with probability \(\to 1\) as \(n \to \infty\).

Proof:

By Lemma 3.3.2, we know that one set of estimates is \(\sqrt{n}\)-consistent for \((\alpha_1, \alpha_2, \beta_1, \beta_2)\). We must show, therefore, that with \(p \to 1\) as \(n \to \infty\), the pair of lines minimizing \((d_s, d_c)\) is the correct pair.

Assume without loss of generality that the straight lines are the correct pair. Then, by Lemma 3.3.2

\[
\hat{\alpha}_{js} = \alpha_j + O_p(1/\sqrt{n})
\]

and

\[
\hat{\beta}_{js} = \beta_j + O_p(1/\sqrt{n}) \quad (j=1, 2).
\]

Since the straight lines are correct, the crossed lines estimate different (i.e. wrong) slopes and intercepts. We denote these by \(\alpha'_j\) and \(\beta'_j\) \((j=1, 2)\). Thus

\[
\hat{\alpha}_{jc} = \alpha'_j + O_p(1/\sqrt{n})
\]

and

\[
\hat{\beta}_{jc} = \beta'_j + O_p(1/\sqrt{n}) \quad (j=1, 2).
\]

Let \(m\) be defined by:
(3.3.9) \[ m = (\alpha' + \beta' \bar{x}(3)) - (\alpha + \beta \bar{x}(3)) \]

By assumption \( \alpha \neq \alpha' \), and thus without loss of generality we can assume that \( m > 0 \).

We now compare the two distances \( d_c \) and \( d_s \). For simplicity, we consider only the first terms of \( d_s \) and \( d_c \). This makes the following argument notationally simpler, without changing the essence of the argument.

\[
d_{s1} = |\hat{\mu}_{1}^{(3)} - (\hat{\alpha}_{1s} + \hat{\beta}_{1s} \bar{x}(3))| \\
= |\hat{\mu}_{1}^{(3)} - (\alpha_{1} + \beta_{1} \bar{x}(3)) + (\alpha_{1} + \beta_{1} \hat{x}(3)) - (\hat{\alpha}_{1s} + \hat{\beta}_{1s} \bar{x}(3))| \\
\leq 0_p(1/\sqrt{n}) + 0_p(1/\sqrt{n}) = 0_p(1/\sqrt{n})
\]

While,

\[
d_{c1} = |\hat{\mu}_{1}^{(3)} - (\hat{\alpha}_{1c} + \hat{\beta}_{1c} \bar{x}(3))| \\
= |\hat{\mu}_{1}^{(3)} - (\alpha'_{1} + \beta'_{1} \bar{x}(3)) + (\alpha'_{1} + \beta'_{1} \hat{x}(3)) - (\hat{\alpha}_{1s} + \hat{\beta}_{1s} \bar{x}(3))| \\
= |\hat{\mu}_{1}^{(3)} - (m + \alpha_{1} + \beta_{1} \bar{x}(3)) + 0 p (1/\sqrt{n})| \\
= |m + 0 p (1/\sqrt{n})| \\
= m + 0_p(1/\sqrt{n})
\]

Since \( m \) is strictly positive, \( P(d_{s1} > d_{c1}) \to 0 \) as \( n \to \infty \).
Thus the correct pair of lines has smaller total distance (3.3.9) with probability 1 as \( n \to \infty \).

QED.

We must now adjust the variance estimators to get \( \sqrt{n} \)-consistent estimators of \( \sigma_1^2 \) and \( \sigma_2^2 \), rather than the "enlarged" variances \( \sigma_1'^2 \) and \( \sigma_2'^2 \) of lemma 3.3.1. To do this, we consider any one of the intervals \( I(j) \). We have a \( \sqrt{n} \)-consistent estimator \( \hat{\theta}(j) \) of the five parameters \( \theta(j) = (\lambda, \alpha_1 + \beta_1 \bar{x}(j), \alpha_2 + \beta_2 \bar{x}(j), \sigma_1'^2, \sigma_2'^2) \). We start the EM algorithm at this starting point. The estimators of \( \sigma_j'^2 \) (\( j=1,2 \)) become:

\[
\hat{\sigma}_j'^2 = \frac{\sum (y_i - \mu_{*j})^2 w_{*j}}{\sum w_{*j}}
\]

The * indicates that these are the limit estimators provided by the EM algorithm limit. This is too large because the term \( \mu_{*j} \) is constant over the interval. Rather than \( \mu_{*j} \), we should subtract \( \hat{\alpha}_j + \hat{\beta}_j x_1 \) from each \( y_i \). If we substituted the EM limit estimators \( \alpha_{*j} + \beta_{*j} x_1 \) instead of the mean estimators \( \mu_{*j} \), we would have the EM estimators for the correct model (3.3.1) instead of the model (3.3.2). That is, we would have corrected the variance for the regression. We do not, however, know the EM estimators for the slopes and intercepts. Instead, we will substitute the \( \sqrt{n} \)-consistent
slopes and intercept estimators from Lemmas 3.3.2 and 3.3.3. This variance estimator will no longer be the EM estimator, but will be $\sqrt{n}$-consistent for the correct variance. This is stated in Lemma 3.3.5. We prove a version which is notationally simpler in Lemma 3.3.4.

**Lemma 3.3.4**

Let $y_i$ be a from a mixture of normals, as in (2.2.2). Let

\[(3.3.10) \quad \sigma^2_{\star j} = \frac{\sum (y_i - \mu^\star_j)^2 W^\star_{ji}}{\sum W^\star_{ji}}\]

be the EM algorithm estimator of $\sigma_j^2$ from (2.2.13). Let

\[(3.3.11) \quad \hat{\sigma}_{jn}^2 = \frac{\sum (y_i - \hat{\mu}_{jn})^2 W^\star_{ji}}{\sum W^\star_{ji}}\]

be identical to (3.3.10) except with a $\sqrt{n}$-consistent estimator $\hat{\mu}_{jn}$ substituted for the EM algorithm limit estimator $\mu^\star_j$. Then $\hat{\sigma}_{jn}^2$ is $\sqrt{n}$-consistent for $\sigma_j^2$.

**Proof:**

68
\[ \hat{\sigma}_{jn}^2 = \frac{\Sigma (y_i - \hat{\nu}_{jn})^2 w^*_{ji}}{\Sigma w^*_{ji}} \]

\[ = \frac{\Sigma (y_i - \mu^*_{j})^2 w^*_{ji}}{\Sigma w^*_{ji}} \]

\[ + \frac{\Sigma (y_i - \mu^*_{j})(\mu^*_{j} - \hat{\nu}_{jn}) w^*_{ji}}{\Sigma w^*_{ji}} \]

\[ + \frac{\Sigma (\mu^*_{j} - \hat{\nu}_{jn})^2 w^*_{ji}}{\Sigma w^*_{ji}} \]

\[ = \sigma_j^2 + o_p(1/\sqrt{\pi}) + \]

\[ + \frac{2 \Sigma (y_i - \mu^*_{j}) w^*_{ji}}{\Sigma w^*_{ji}} \]

\[ + \frac{\Sigma w^*_{ji}}{\Sigma w^*_{ji}} \]

\[ = \sigma_j^2 + o_p(1/\sqrt{\pi}) + o_p(1/\sqrt{\pi}) \times 20_p(1/\sqrt{\pi}) + \]

\[ 0_p(1/n) \]

\[ = \sigma_j^2 + o_p(1/\sqrt{\pi}) \]

\text{QED.}
Lemma 3.3.5

Let \((x_i, y_i)\) be a from a mixture of regressions, as in (3.1.1).

Let

\[
\sigma_j^2 = \frac{\sum(w^{*}_{ij})(y_i - (\alpha^{*}_j + \beta^{*}_j x_i))^2}{\sum w^{*}_{ij}}
\]

be the EM algorithm estimator of \(\sigma_j^2\) from (2.2.14). Let

\[
\hat{\sigma}_{jn}^2 = \frac{\sum(w^{*}_{ij})(y_i - (\hat{\alpha}_{jn} + \hat{\beta}_{jn} x_i))^2}{\sum w^{*}_{ij}}
\]

be identical to (3.3.12) with \(\sqrt{n}\)-consistent estimators \(\hat{\alpha}_{jn}\) and \(\hat{\beta}_{jn}\) substituted for the EM algorithm limit estimators \(\alpha^{*}_j\) and \(\beta^{*}_j\). Then \(\hat{\sigma}_{jn}^2\) is \(\sqrt{n}\)-consistent for \(\sigma_j^2\), \(j = 1, 2\).

Proof: Essentially the same as the proof of Lemma 3.3.4.

QED.

The binning procedure, the proof of its \(\sqrt{n}\)-consistency, and thus the asymptotic efficiency of the EM algorithm are summarized in the following proposition.
Proposition 3.3.6
Let \((x_i, y_i) \ i=1, \ldots, n\) be independent such that \(x_i\) are non-random satisfying (R1) and given \(x_i, y_i\) follows the mixture of regressions model (3.3.1). Let \(I(j), j=1, 2, 3\) be non-overlapping connected intervals on the x-axis, such that exactly \(k\) of the \(x_i\) are in each interval. Also, if \(\beta_1 \neq \beta_2\), assume that the crosspoint \(x_c\) defined by:

\[
x_c = \frac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1},
\]

is not contained in any of the \(I(j)\).

For each \(I(j)\), let \(\theta(j)\) be a \(\sqrt{n}\)-consistent estimator of the three parameters:

\[
\theta(j) = (\lambda, \mu_1(j), \mu_2(j)),
\]

where \(\mu_1(j) = \alpha_1 + \beta_1 \bar{x}(j), (j=1, 2, 3; i=1, 2)\). Let \((\hat{\alpha}_1 s, \hat{\alpha}_2 s, -\hat{\beta}_1 s, \hat{\beta}_2 s)\) and \((\hat{\alpha}_1 c, \hat{\alpha}_2 c, \hat{\beta}_1 c, \hat{\beta}_2 c)\) be the estimators defined from (3.3.4) through (3.3.7) from intervals \(I^{(1)}\) and \(I^{(2)}\). Let \(\hat{\mu}_1(3)\) and \(\hat{\mu}_2(3)\) be the mean estimates from \(\theta(3)\). Compute the distances (3.3.8) and choose the estimates \((\hat{\alpha}_1 s, \hat{\alpha}_2 s, \hat{\beta}_1 s, \hat{\beta}_2 s)\) if \(d_s \leq d_c\). Otherwise choose \((\hat{\alpha}_1 c, \hat{\alpha}_2 c, \hat{\beta}_1 c, \hat{\beta}_2 c)\). Denote the chosen set of estimates simply by \((\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2)\).
Now, in $I(1)$, let $W^*_{1i}$ and $W^*_{2i}$ be the EM algorithm limit estimators of the group memberships of the $y_i$ in $I(1)$, from (2.2.12). Let

$$
\hat{\sigma}_j^2 = \frac{\sum (y_i - (\hat{\alpha}_{jn} + \hat{\beta}_{jn} x_i))^2 W^*_{ji}}{\sum W^*_{ji}}
$$

be the estimators of $\sigma_j^2$ from (3.3.13). Define $\hat{\theta}_n$ to be the estimator

$$
\hat{\theta}_n = (\hat{\lambda}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2).
$$

Now consider the EM algorithm sequence $\{\theta^{(p)}\}$ (via (2.2.14)) starting at $\hat{\theta}_n$. That is, $\theta^{(0)} = \hat{\theta}_n$. Let $\theta^*_n = \lim_{p \to \infty} \theta^{(p)}$. Then $\theta^*_n$ is asymptotically efficient for estimating $\theta$.

Proof: (Outline)

By lemma 3.3.1, the $\theta^{(j)}$, $j=1,2,3$ are each $\sqrt{n}$-consistent for $(\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2)$. By Lemma 3.3.3, the correct pair of lines is chosen with $P \to 1$ as $n \to \infty$. By Lemma 3.3.5, the variance estimators $\hat{\sigma}_j^2$ are $\sqrt{n}$-consistent for $\sigma_j^2$ ($j=1,2$). Therefore, $\hat{\theta}_n$ above is $\sqrt{n}$-consistent for $\theta$ and by Proposition 2.3.6, $\theta^*_n$ is asymptotically efficient.

QED.
Problems of Implementation

We have shown that given the choice of any three intervals satisfying the conditions of Lemma 3.5.2, the binning procedure outlined in the previous section will produce a $\sqrt{n}$-consistent estimator of $\theta = (\lambda, \alpha_2, \sigma_2, \beta_2, \gamma_2, \sigma_2', \sigma_2^2)$ for the mixture of regressions problem. By lemma 3.5.1, one must assume that the true regression lines do not cross within the interval $I(J)$. This can be accomplished by estimating the slopes and intercepts from several intervals and estimating the position of the cross point $x_c$. Then the intervals used must not contain $x_c$. This will ensure that the cross point is not within any of the intervals considered with probability tending to 1 as $n$ increases. In practice, we found that ignoring this consideration while changing the starting values to some extent, does not change the EM algorithm limits greatly. For our simulations and data analysis, we often use the lower and upper thirds of the data to obtain the crossed and straight parameter estimates from equations (3.3.4) - (3.3.7). We then used the middle interval to choose between the two sets of estimates. This caused the "wrong" sets of starting lines to be chosen up to 37% of the time, but resulted in a spurious EM algorithm limit only once out of 300 simulations. For a particular
data set, one might want to estimate the cross point first and avoid intervals containing it. We will do this for the subject we will discuss in detail in Chapter 5.

§3.3.2 Other possible starting points

We conclude this chapter with a brief examination of how one might construct other $\sqrt{n}$-consistent estimators for $\theta$, based on the method of moments and minimum distance techniques. For both, we conclude that the problems of implementation at the present level of available computation are too severe to warrant implementation. For the sake of further research, we indicate how one might proceed.

Method of Moments

In order to generalize the method of moments technique used in §3.2.1 to the regression case, we will view the pairs $(x_i, y_i)$ as coming from a mixture (with mixing parameter $\lambda$) of bivariate normals, with different means, variances and covariances. That is, we will view the $x_i$ as random in order to use the moments of the data to estimator the parameters of the bivariate mixture. Specifically, we assume that $(x_i, y_i)$ are i.i.d with density
\[ f(x, y) = \lambda f_1(x, y) + (1 - \lambda) f_2(x, y), \]

with

\[ f_j(x, y) = N((\mu_{xj}, \mu_{yj}), \Sigma_j), \]

where \[ \Sigma_j = \begin{bmatrix} \sigma^2 & \sigma \rho & \sigma \\ \rho & \sigma_{yj} & \rho_j \\ \sigma & \rho_j & \sigma^2 \end{bmatrix} \]

We then associate the regression parameters from the model (3.3.1) via

\[ \alpha_j = \mu_{yj} - \mu_{xj} \rho_j \left( \frac{\sigma_{yj}}{\sigma_{xj}} \right) \]

and,

\[ \beta_j = \rho_j \left( \frac{\sigma_{yj}}{\sigma_{xj}} \right) \]

Thus, if we can find \( \sqrt{n} \)-consistent estimators for the eleven parameters of the mixed bivariate normal, we can determine the seven regression parameters via (3.3.15).

We will attempt to generalize the method of moments used by Pearson (1894) and Cohen (1967) for the five parameter case, described in §3.2 to the model (3.3.14). That is, we equate eleven sample moments of the data \((x_i, y_i)\) to functions of the eleven parameters of (3.3.14). We then find the regression estimators via (3.3.15).
We use the characteristic function of the bivariate normal to equate the moments of the mixed distribution with the moments of the two bivariate normals. We then substitute the sample moments from the data for the moments of the mixed distribution and solve for the eleven parameters in equation (3.3.14). We estimate $\Theta = (\lambda, \alpha_2, \alpha_2, \beta_2, \beta_2, \sigma_2^2, \sigma_2^2)$ from (3.3.15). This is precisely what was done for a mixture of normals in §3.2.1.

The first question is which eleven mixed moments to use. In the five parameter case, the natural choice was the first five moments. Recall that we used $\alpha_1, \ldots, \alpha_5$ where $\alpha_i$ was the $i$th central moment of the mixture. (See equation (1.1.2)). We equated these with functions of $\Theta = (\lambda, \nu_1, \nu_2, -\sigma_1^2, \sigma_2^2)$. In the present case one might naturally use

$$EX, \ EY, \ EX^2, \ EY^2$$

and perhaps

$$EXY, \ EX^2Y^2 \text{ and } EX^3Y^3.$$ 

To solve for $(\lambda, \nu_{X1}, \nu_{X2}, \nu_{Y1}, \nu_{Y2}, \sigma_{X1}, \sigma_{X2}, \sigma_{Y1}, \sigma_{Y2}, \rho_1, \rho_2)$, however, we need four more moments. One could use any of $EX^2Y, \ EXY^2, \ EX^3Y, \ EX^3Y^2, \ EX^2Y^3$ or even higher mixed moments. The choice is not clear. Any eleven sample moments could be used to solve the eleven equations for the eleven parameters. Because of implementation problems discussed below, we did not consider this choice further. We will, however,
indicate the procedure for the first seven moments and leave open the full specification for further research. Using characteristic functions, we find that

\[(3.3.16)\]

\[
\begin{align*}
E(X) &= \lambda \mu_{X1} + (1 - \lambda) \mu_{X2} \\
E(Y) &= \lambda \mu_{Y1} + (1 - \lambda) \mu_{Y2} \\
E(X - EX)^2 &= \lambda \sigma_{X1}^2 + (1 - \lambda) \sigma_{X2}^2 \\
E(Y - EY)^2 &= \lambda \sigma_{Y1}^2 + (1 - \lambda) \sigma_{Y2}^2 \\
E(XY - EXEY) &= \lambda \rho_1 + (1 - \lambda) \rho_2 \\
\sigma_X \sigma_Y \\
E(2X^2Y^2 - EX^2EY^2) &= \lambda \rho_1 (\sigma_{X1}^2 \sigma_{Y1} \rho_1 + 2 \mu_{X1} \mu_{Y1}) + (1 - \lambda) \rho_2 (\sigma_{X2}^2 \sigma_{Y2} \rho_2 + 2 \mu_{X2} \mu_{Y2}) \\
E(X^3Y^3 - EX^3EY^3) &= \lambda \rho_1 (\sigma_{X1}^3 \rho_1^2 + B_1 \rho_1 + C_1) + (1 - \lambda) \rho_2 (\sigma_{X2}^3 \rho_2^2 + B_2 \rho_2 + C_2)
\end{align*}
\]

where \(A_j = 2\sigma_{Xj}^2 \sigma_{Yj}^2; \ B_j = 6\mu_{Xj} \mu_{Yj} \sigma_{Xj} \sigma_{Yj};\) and \(C_j = 3(\mu_{Xj}^2 + \sigma_{Xj}^2)(\mu_{Yj}^2 + \sigma_{Yj}^2).\)

After choosing four additional moments and solving for four more functions of \((\lambda, \mu_{X1}, \mu_{X2}, \mu_{Y1}, \mu_{Y2}, \sigma_{X1}, \sigma_{X2}, \sigma_{Y1}, \sigma_{Y2}, \rho_1, \rho_2),\) we would follow the method of §3.2 and substitute the sample values for the corresponding values on the left hand side of (3.3.16) and solve for the parameters on the right hand side. We could attempt to solve these eleven equations...
algebraically as Pearson (1894) did in the five parameter case, or we could find the solutions numerically. That the resulting estimator would be \( \sqrt{n} \)-consistent as can be seen by generalizing the proof of Proposition 3.2.1.

**Problems of Implementation**

The problems of using the method of moments estimator described in the preceding section are threefold. First, as mentioned, a choice of which eleven moments to use must be made. There is no natural choice for the higher order moments \( \mathbf{E}^{j+k} \) \( j \neq k \). Secondly, after choosing the eleven moments and thus, via characteristic functions, the eleven functions of \( (\lambda, \mu_{x1}, \mu_{x2}, \mu_{y1}, \mu_{y2}, \sigma_{x1}, \sigma_{x2}, \sigma_{y1}, \sigma_{y2}, \rho_1, \rho_2) \), one must substitute the sample moments and solve for the estimators of \( (\lambda, \mu_{x1}, \mu_{x2}, \mu_{y1}, \mu_{y2}, \sigma_{x1}, \sigma_{x2}, \sigma_{y1}, \sigma_{y2}, \rho_1, \rho_2) \). In the five parameter case, Pearson (1894) was able to reduce the system of five equations to one ninth-order equation—(equation (1.2.2)). We used Newton's method to solve for the negative, real root of this equation numerically. In the regression situation, one is not guaranteed of finding a polynomial from which we can estimate the eleven parameters. The other option would be to solve the eleven equations (the first seven of which comprise (3.3.16)), numerically. This is equivalent to numerically minimizing an 11-dimensional
function, which for all but the largest computers is impractical. This is especially inhibiting when one keeps in mind that this estimator serves only as a starting point for the EM algorithm.

Finally, if the first two problems are overcome, robustness problems persist. Because high order moments are used, data points located far from the center will have arbitrarily high influence. From studies of robust correlation and robust variance (see e.g. Huber (1973) and (1981)), one must question using sample moments of the form $E x_1^j y_1^k$ for $j \geq 2$, $k \geq 2$.

In conclusion, we find that the method of moments as applied here is not, under present computing constraints, a feasible method for finding a $\sqrt{m}$-consistent starting point for the EM algorithm. We have discussed the method for completeness, and to aid in further research in the application of the method to this problem.

**Minimum Distance Estimation**

As in the method of moments section, we will consider the data $(x_i, y_i)$ to follow equation (3.3.14), a mixture of bivariate normals. We again consider the Cramér-von Mises distance where now $F, G : \mathbb{R}^2 \rightarrow \mathbb{R}$. The regularity conditions leading to Theorem 2 of Woodward et al (1984) remain valid.
for the family of bivariate Normal mixtures. Asymptotic normality of the Cramér-von Mises estimator is then guaranteed by result 5B of Millar (1981 p.88) which states that analogues of the main results (including asymptotic normality) continue to hold if $F_\theta$ is permitted to be a distribution on $\mathbb{R}^d$ instead of $\mathbb{R}$.

The formula for minimization becomes

$$(3.3.17) \quad \sum_i [F_\theta(x_i, y_i) - (i-.5)/n]^2$$

where the $(x_i, y_i)$ are ordered in increasing $F$. The sum is over all pairs, $i=1, \ldots, n$. The asymptotic normality of the estimator ensures that the estimator found by minimizing $(3.3.17)$ is $\sqrt{n}$-consistent for $\theta$.

Problems of Implementation

The central problem of implementing the minimization of $(3.3.17)$ is the same as the final problem discussed for the method of moments estimator. That is, we must again minimize a function of eleven parameters. For this reason, we are forced to conclude that the minimum distance estimator is not feasible as a starting point for the EM algorithm.

One might also apply minimum distance techniques directly to the model $(3.3.1)$ rather than the bivariate normal mixture $(3.3.14)$. While this is a possible avenue for further research, no reliable minimum distance methods
currently exist for model (3.3.1) (personal communication R. Beran (1985)).

§3.4 Conclusions

In the five parameter case, we have shown that two methods are available for finding a $\sqrt{n}$-consistent starting point for the EM algorithm: the method of moments and minimum distance method. The main problems of the MM method are data sensitivity and inability in some cases to find a root of (1.2.3) which will result in a feasible estimate. By contrast, the MD method requires its own starting point and involves an order of magnitude more computation to perform a five dimension minimization search than does the resulting EM algorithm. In practice, we have found that the naive estimator

$$\hat{\theta} = (0.5, \overline{x} - s/2, \overline{x} + s/2, s/2, s/2)$$

while not $\sqrt{n}$-consistent, usually results in the same EM algorithm limit as does the MM estimate starting point. Thus, in samples where no MM estimate exists or where outliers are suspected, we recommend using the naive estimator either as a check for the MM starting point or when the MM estimator cannot be found.

In the regression problem, we consider MM and MD methods and conclude that computation time and other problems
prohibit their implementation in practice. Instead, we have proposed an ad hoc procedure based on binning the data. We have shown that this procedure provides $\sqrt{N}$-consistent starting points for the EM algorithm which are relatively easy to implement.

In the next chapter, we present our simulation results and conclusions for the methods presented in this chapter.
Chapter 4

§ 4.1 Introduction

In this chapter we present the results of two Monte Carlo simulations. In the first, the performance of the maximum likelihood estimates as found by the EM algorithm is compared to the minimum distance estimates. Their empirical mean square error, sample covariance structure and other properties are compared to each other as well as to the information matrix. The asymptotic variances found via the information matrix are used to obtain asymptotic confidence intervals. In the second set of simulations, the EM algorithm is used to find maximum likelihood estimates in the mixture of regressions case. Because of the problems outlined in Chapter 3, there are no competitors with which to compare these estimators. Instead, they are evaluated by comparing their MSE and the variance-covariance matrix with their asymptotic counterparts. Asymptotic confidence intervals are also obtained for this case using the information matrix.

§ 4.2 Mixture of Normals Simulations

Computational Considerations

In this section we consider $y_i$, $i=1, \ldots, n$ from the model
\[ f(y_1) = \lambda f_1(y_1) + (1 - \lambda) f_2(y_1) \]

where

\[ f_j = \frac{1}{\sqrt{2\pi}} \sigma_j \exp \left\{ -\frac{1}{2} \left( \frac{y_1 - \mu_j}{\sigma_j} \right)^2 \right\}. \]

That is, the \( y_1 \) are an i.i.d sample of size \( n \) from a mixture of normals.

We compare the performance of the maximum likelihood estimates as found by the EM algorithm (referred to as the EM estimates) to the minimum distance estimates (MD estimates). A similar study has been recently performed by Woodward et al. (1984).

Behboodian (1972) has shown that the information matrix \( I(\theta) \) of a mixture (4.2.1) can be found by first finding the "scaled" information matrix \( J(\theta) \) which depends only on three parameters \( d, r \) and \( \bar{\sigma} \). Then \( I(\theta) = DJD \) where \( D = \text{diag}(1, 1/\sigma_1, -1/\sigma_2, 1/\sigma_1^2, 1/\sigma_2^2) \). The parameters \( d, r \) and \( \bar{\sigma} \) are defined as follows:

\[
\begin{align*}
    r &= \sigma_1 / \sigma_2 \\
    d &= (\mu_2 - \mu_1) / 2\bar{\sigma} \\
    \bar{\sigma} &= \sqrt{\sigma_1 \sigma_2}.
\end{align*}
\]

In our notation, \( d \) is always positive since \( \mu_2 > \mu_1 \). By defining \( \bar{\mu} = (\mu_1 + \mu_2)/2 \), and considering the transformation \( x = (y - \bar{\mu}) / \bar{\sigma} \), the normal mixture (4.2.1) is reduced to the standard normal
mixture with three positive parameters $\lambda, d$, and $r$. Given $\lambda$ and $r$, the parameter $d$ expresses the separation between the two mixtures.

Behboodian (1970) has shown that a sufficient condition for the mixture to be unimodal is $(\mu_2 - \mu_1) \leq 2\min(\sigma_1, \sigma_2)$. It is intuitively clear that the larger the separation, the easier it will be to estimate the five parameters $\theta = (\lambda, \mu_1, -\mu_2, \sigma_1^2, \sigma_2^2)$. Woodward et al. (1984) defined the concept of overlap to quantify this separation. Following Woodward et al. (1984), consider the following classification scheme for identifying the population of an observation $y_i$ from (4.2.1).

Classify an observation $y$ as

In population 1 if $y < y_C$

In population 2 otherwise

where $y_C$ is the unique solution to

$$\lambda f_1(y_C) = (1-\lambda) f_2(y_C).$$

(The $f_i$ are $N(\mu_i, \sigma_i^2)$ as in (4.2.1)).

**Definition:** The overlap of two normal densities is defined as the total probability of misclassification using the rule (4.2.2).

In our simulation, we will take $\mu_1 = 0$ and $\sigma_1 = 1$. Each
simulation situation is defined in terms of $r$, $\lambda$ and the overlap. For example, a $\lambda = .5$ mixture of two normals, one with $\mu_1 = 0$, $\sigma_1^2 = 1$ and the other at $\mu_2 = 1.94$, $\sigma_2^2 = 2$, corresponds to $\lambda = .5$, $r = 1/\sqrt{2}$, and overlap .10.

The goals of our simulations are different from Woodward et al (1984) for several reasons. First, they restricted their comparisons to estimation of the mixture parameter ($\lambda$) only. They regarded the other parameters as "nuisance parameters". Secondly, their main concern was to show the improvement of the MD estimates over the MLE (EM) estimates in non-normal situations, which we do not consider. In contrast, our goals are twofold:

- To compare the performance of the EM estimates in small samples to that predicted by asymptotic efficiency.

- To compare the performance of the EM estimates in small samples to the MD estimates, with data from normal mixtures.

By the results of Chapter 2, for the mixture of normals and regressions, MD estimates are asymptotically less efficient than EM estimates. We have included the MD estimates only for qualitative comparison purposes, to learn what to do in small samples.

For each set of parameter configurations, 100 samples of
size n=100 were generated from the appropriate mixture
distribution. All programs were written in C and carried out
on an IBM PC. A Uniform random variable was used to assign
the normals to one of the two mixture components with
probability λ.

We have followed the recommendation of Behboodian (1972)
in calculating the information matrices for normal mixtures.
Specifically, we have used Romberg's algorithm at 2^9 points
to numerically calculate each integral. For all parameter
combinations in common, our results agree to a least 4
decimal places with Behboodian (1972). From the diagonal
elements of I^{-1}(θ)/n, we obtain the Cramér-Rao lower bounds
for the variances of a sample of size n from (4.2.1). Con-
sider the matrix below for θ = (.5, 0, 1.05, 1, 1) with n=100.
This has overlap .30.

\[
\frac{1}{100} (I^{-1}(θ)) = \begin{bmatrix}
58.81 & 61.74 & 61.74 & 21.93 & -21.93 \\
61.74 & 65.19 & 64.49 & 23.39 & -22.67 \\
61.74 & 64.49 & 65.19 & 22.67 & -23.39 \\
21.93 & 23.39 & 22.67 & 8.62 & -7.82 \\
\end{bmatrix}
\]

Thus we see that for estimating λ for example, the lower
bound on the variance is 58.81. For the EM, which is
asymptotically efficient and normal, a 100(1-α)% asymptotic
confidence interval for λ is \( \hat{λ} \pm 2\alpha/2 (58.81)^{1/2} \). Thus a 95%
confidence interval for λ would be \( \hat{λ} \pm 15.04 \), which is
clearly absurd. Similarly a 95% confidence interval for $\nu_1$ is $\hat{\nu}_1 \pm 15.85$. To get reasonable confidence intervals, we need a sample size on the order of $10^5$. Thus, there is essentially no hope in estimating the parameters of a normal mixture with overlap = .30 for other than huge sample sizes. We nevertheless considered one high overlap situation to study the behavior of the estimators under these adverse conditions, and concluded that our suspicions were founded. The sample variances were of the same order of magnitude as predicted by the information matrix, and thus confidence intervals from the asymptotic variances would be too wide to be useful. As will be seen later, the situation is not nearly so bleak for smaller overlaps.

Based on the results of Chapter 3, it was desired to use the method of moments estimate as the starting value for the EM and MD searches. The method of moments estimate is obtained by numerically finding the negative real root of the nonic equation (1.2.3), and then solving for the parameter estimates using equation (1.2.4). It is then checked to see if it is a feasible solution, i.e., one for which $0 \leq \lambda_i$, and $\sigma_i^2 > 0$. Unfortunately, this procedure may fail for several reasons:

1) No negative root can be found
2) A negative root is found, but estimates are not feasible
3) More than one feasible negative root is found

Previous authors (see especially Quandt and Ramsey (1978)) have reported high frequency of failure for the method of moments estimator. We have found that the frequency is related to the overlap. For overlap = .30 it failed in 35% of the runs. For overlap = .10 it failed in 16.9% of the runs, while for overlap ≤ .03, it failed in only 6.3% of the runs. By contrast, Quandt and Ramsey (1978) reported 59.7% failures at overlap = .22, 28.6% failures at overlap = .09 and 2% failures at overlap = .01.

We are faced with two choices for the simulation. We can either use a different starting value for the EM and MD estimator when the MM does not provide one, or we can restrict our analysis to those simulations for which an initial MM estimate was found. The problem with the first alternative is that we have no other starting point candidates which are $\sqrt{M}$-consistent, and the resulting EM estimates will not necessarily be asymptotically efficient.

We have considered, however, the naive estimator (NE) defined by:

\begin{equation}
\hat{\theta} = (0.5, \bar{x} - s/2, \bar{x} + s/2, s/2, s/2)
\end{equation}
where \( \bar{x} \) and \( s \) denote the sample mean and standard deviation of the mixed sample. On the other hand, by considering only the simulations for which a MM estimate was obtained, we may introduce bias in the performance of the resulting EM and MD estimators. Fortunately, we found that in a sample of 100 simulations from different \( \theta \) configurations where both an MM and an NE starting point were found, the EM algorithm converged to the same final estimate in all 100 of the simulations. In addition, we compared the run length of the EM and the mean square errors for each parameter estimate of the EM in 50 simulations for which an MM estimate was found against 50 simulations for which an MM estimate was not. Although, the EM seemed to converge more quickly from the MM estimate, we found no differences (at \( \alpha = .05 \)) in any of these characteristics. Previous authors, (see especially Fowlkes (1977)) have discussed the importance of starting points for the EM. We have concluded, however, that the naive estimator (4.2.3) is "close enough" to the \( \sqrt{n} \)-consistent MM estimator to ensure EM convergence to the right local maximum. No doubt, in practice on a non-simulated data set, one would want to start the EM from several places to see if more than one local maximum exists. In summary, we have decided to restrict our attention to the simulations for which an MM starting point was found. Specifically, for
each $\theta$ configuration, we selected the first 100 runs for which the MM estimate was obtained.

A modified Davidon-Fletcher-Powell procedure was used to minimize the sample Cramér-von Mises distance over the five parameters, thus obtaining the MD estimates. Specifically, at each stage, the function take a variable step size in the direction of the product of the numerically computed gradient with a numerically updated Hessian matrix until the function no longer decreases. At the next stage, the gradients are recalculated, the Hessian updated, and the step size recalculated based on the performance at the last stage. This continues until both the gradient norm and the decrease in the function value fall below the convergence criterion, discussed below. Since we must have $0 \leq \lambda \leq 1$ and $\sigma_1^2 > 0$, it might be argued that a restricted minimization is required. (For example by introducing a barrier function on $\lambda$ and the $\sigma_1^2$). However, in over 1400 samples considered, only twice did the MD algorithm converge to an infeasible estimate ($\hat{\lambda} < 0$, $\hat{\lambda} > 1$, or $\hat{\sigma}_1^2 < 0$). Thus we conclude that the unrestricted minimization search suffices for our purpose.

For each sample, both the EM and MD estimates were obtained. A rather severe convergence criterion $\epsilon^*$ was used for each. For the EM, convergence continued until the
improvement in the sample likelihood was less than $\varepsilon^k$. For the MD, convergence continued until both the gradient norm and the decrease in the function were less than $\varepsilon^k$.

Previous authors have used $\varepsilon^k = 10^{-4}$, but it due to many "flat points" in the multidimensional likelihood function, and the function (3.2.2), we found that estimates often differed by 10% or more from the values obtained by using $\varepsilon^k = 10^{-4}$, if left to converge to $\varepsilon^k = 10^{-6}$. This, of course resulted in more steps for each.

It is known that the EM converges to a local maximum slowly while typically a Newton-Raphson search converges to a solution quadratically (see Wu (1983) and Kennedy and Gentle (1980)). While this seems a reason for preferring the MD estimates, this is misleading. The EM sometimes took over 1000 steps (maximum was ~5500), and averaged about 100 steps. The MD algorithm, on the other hand, took many fewer "steps", averaging about 30 steps per sample. However, each step for the MD algorithm was comprised of many more operations than the EM step. A typical sample of n=100 (excluding the high overlap case), took roughly 8 minutes of computing time (IBM PC-AT with 80287 math co-processor) to converge ($\varepsilon^k = 10^{-6}$) to both the EM and MD estimates. Of this time, more than 80% was spent computing the MD estimates. In short, the EM solution was, in general, much
quicker to find. This corroborates previous findings (see Wu (1983)).

Simulation Design

For comparison purposes we have based our simulation design around the one presented by Woodward et al. (1984). The design centers around 8 configurations of θ, a $2^3$ factorial design in the three parameters $λ$, $r$, and overlap. Specifically $λ = .25$ or .50, $r (\sigma_2/\sigma_1) = 1$ or $\sqrt{2}$, and overlap = .03 or .10. Figure 4.1 presents the values of the parameters for the 8 runs.

<table>
<thead>
<tr>
<th>#</th>
<th>Overlap</th>
<th>$r$</th>
<th>$λ$</th>
<th>$μ_1$</th>
<th>$μ_2$</th>
<th>$σ_1^2$</th>
<th>$σ_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.03</td>
<td>1</td>
<td>.50</td>
<td>0</td>
<td>3.75</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>.03</td>
<td>1</td>
<td>.25</td>
<td>0</td>
<td>3.60</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>.03</td>
<td>.717</td>
<td>.50</td>
<td>0</td>
<td>4.52</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>.03</td>
<td>.717</td>
<td>.25</td>
<td>0</td>
<td>4.46</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>.10</td>
<td>1</td>
<td>.50</td>
<td>0</td>
<td>2.56</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>.10</td>
<td>.717</td>
<td>.25</td>
<td>0</td>
<td>2.32</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>.10</td>
<td>.717</td>
<td>.50</td>
<td>0</td>
<td>3.06</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>.10</td>
<td>.717</td>
<td>.25</td>
<td>0</td>
<td>2.96</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 4.1 Configurations for Simulation

The results of the first eight simulation configurations are summarized in Figure 4.3. Each cell (representing one configuration) contains summary information on the performance of the EM estimates for all 5 parameters in comparison with the MD and MM estimates. In addition, it contains the percentage failure of the MM for that cell.
Specifically, the first row of the cell indicates, for each parameter \((\lambda, \nu_1, \nu_2, \sigma_1, \sigma_2)\) whether the MSE of the EM was significantly \((\alpha = .05)\) larger than the MSE of the MD estimate, indicated by a \(-\), significantly smaller than the MSE of the MD estimate, indicated by a \(+\), or not significantly different, indicated by a \(0\). The next line is percentage failure of the MM to find feasible estimates. Below that are three lines containing the MSEs for the EM, MD and MM estimates respectively (for each parameter). Finally, the last line indicated the percentage that the EM estimate was closer (in absolute value) to the parameter than the MD estimate. The format of these cells, shown in Figure 4.2, provides a summary of the above description.

<table>
<thead>
<tr>
<th>(\lambda)</th>
<th>(\nu_1)</th>
<th>(\nu_2)</th>
<th>(\sigma_1)</th>
<th>(\sigma_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE EM &lt; MSE MD?</td>
<td>(+, 0, -)</td>
<td>MSE EM &lt; MSE MM?</td>
<td>(+, 0, -)</td>
<td>% MM failed</td>
</tr>
<tr>
<td>MSE EM for each parameter</td>
<td>MSE MD</td>
<td>MSE MM</td>
<td>% EM closer than MD</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.2**

Format for performance summaries of Figure 4.3
We summarize the performance of all eight configurations in Figure 4.4. Here we list the number of parameters for which the MSE of the EM is significantly lower than the MD, the percentage of the MM, and the percentage that the EM was closer to the parameter than the MD, averaged over the 5 parameters. We see that the EM outperforms the MD in every configuration both in terms of MSE and in percentage closer to the parameter. The failure percentage of the MM increases as the overlap increases and as \( \lambda \) deviates from .5. The MM estimates of \( \lambda \) and the means seem to fare well in comparison with the EM estimates, while the EM always estimate the variances with lower MSE. This is not surprising given the nature of the MM estimate, which depends on 4th and 5th sample moments to estimate the variances.

The variances of the estimates (EM, MD and MM) are compared in Figure 4.5. The first row lists the relative efficiency of the EM to the Cramér-Rao lower bound for a sample of the same size. The second row lists the relative efficiency of the MD compared to the EM, and the last row lists the relative efficiency of the MM compared to the EM. We see that the EM does perform (in terms of variance) much as predicted by its asymptotic efficiency. Thus, using variances from the information matrix in building confidence intervals does not seem unreasonable. On again sees that
<table>
<thead>
<tr>
<th>$\sigma_1/\sigma_2 = 1$</th>
<th>$\sigma_1/\sigma_2 = \sqrt{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0 + +</td>
<td>0 0 0 0 + +</td>
</tr>
<tr>
<td>0 0 0 0 + +</td>
<td>+ + 0 0 + +</td>
</tr>
<tr>
<td>5.0%</td>
<td>8.3%</td>
</tr>
<tr>
<td>.360 .350 5.50 2.30 2.26</td>
<td>.413 .379 9.70 2.45 4.48</td>
</tr>
<tr>
<td>.430 .378 6.23 2.97 3.95</td>
<td>.430 .373 10.3 3.04 6.94</td>
</tr>
<tr>
<td>.358 4.63 5.85 13.2 9.17</td>
<td>.592 6.77 13.8 17.0 65.7</td>
</tr>
<tr>
<td>53 55 57 56 65</td>
<td>50 53 52 62 60</td>
</tr>
<tr>
<td>+ + + + + +</td>
<td>+ + + + +</td>
</tr>
<tr>
<td>7.0%</td>
<td>5.0%</td>
</tr>
<tr>
<td>.340 9.58 1.96 4.48 1.29</td>
<td>.459 13.3 5.79 6.66 3.53</td>
</tr>
<tr>
<td>1.02 10.5 3.52 25.1 2.58</td>
<td>.897 46.6 8.10 24.3 5.39</td>
</tr>
<tr>
<td>.659 24.8 3.11 44.9 8.21</td>
<td>.527 22.1 7.64 36.0 48.3</td>
</tr>
<tr>
<td>69 68 61 76 64</td>
<td>60 60 65 69 60</td>
</tr>
<tr>
<td>$\lambda = .5$</td>
<td></td>
</tr>
<tr>
<td>+ + 0 + +</td>
<td>+ 0 + + +</td>
</tr>
<tr>
<td>10.0%</td>
<td>10.8%</td>
</tr>
<tr>
<td>3.13 20.4 28.4 5.53 5.10</td>
<td>.232 18.2 31.5 6.84 10.6</td>
</tr>
<tr>
<td>4.05 28.1 28.4 10.9 9.77</td>
<td>.312 19.0 38.8 9.67 15.2</td>
</tr>
<tr>
<td>2.41 16.0 16.8 16.4 13.6</td>
<td>.141 11.9 26.9 12.5 56.8</td>
</tr>
<tr>
<td>63 69 62 65 63</td>
<td>69 64 62 61 66</td>
</tr>
<tr>
<td>+ + 0 + +</td>
<td>+ + 0 + +</td>
</tr>
<tr>
<td>31.6%</td>
<td>17.5%</td>
</tr>
<tr>
<td>5.60 82.7 10.6 18.2 6.05</td>
<td>4.65 54.8 32.6 13.0 10.6</td>
</tr>
<tr>
<td>9.14 111.7 12.1 27.5 11.4</td>
<td>7.45 81.0 29.3 26.0 21.0</td>
</tr>
<tr>
<td>4.27 31.7 174.9 66.2 38.5</td>
<td>3.34 39.9 24.4 34.1 26.1</td>
</tr>
<tr>
<td>68 73 65 68 66</td>
<td>66 58 58 64 66</td>
</tr>
<tr>
<td>$\lambda = .25$</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.3**

Performance of the estimators in eight simulations

96
\[
\begin{array}{|c|c|}
\hline
\sigma_1/\sigma_2 = 1 & \sigma_1/\sigma_2 = \sqrt{2} \\
\hline
\lambda = .5 & \lambda = .25 \\
5.0\% & 7.0\% \\
2 & 5 \\
57.2\% & 67.6\% \\
\hline
\end{array}
\]

\[\text{MSE(EM)} < \text{MSE(MD)}\]
\[\% \text{ EM closer than MD}\]

\[
\begin{array}{|c|c|}
\hline
\sigma_1/\sigma_2 = 1 & \sigma_1/\sigma_2 = \sqrt{2} \\
\hline
\lambda = .5 & \lambda = .25 \\
10.0\% & 31.6\% \\
4 & 4 \\
64.2\% & 68.0\% \\
\hline
\end{array}
\]

\[\text{MSE(EM)} < \text{MSE(MD)}\]
\[\% \text{ EM closer than MD}\]

\textbf{Figure 4.4}

Summary of the performance
<table>
<thead>
<tr>
<th>$\sigma_1/\sigma_2 = 1$</th>
<th>$\sigma_1/\sigma_2 = \sqrt{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = .5$</td>
<td>$\lambda = .25$</td>
</tr>
<tr>
<td>.994 .102 .649 .989 .860</td>
<td>.846 .869 .737 .701 .909</td>
</tr>
<tr>
<td>.839 .925 .883 .775 .574</td>
<td>.961 1.01 .943 .794 .626</td>
</tr>
<tr>
<td>1.01 .756 .940 .176 .247</td>
<td>.723 .582 .730 .086 .038</td>
</tr>
<tr>
<td>.972 1.07 1.14 .902 1.15</td>
<td>.658 .676 .765 .650 .746</td>
</tr>
<tr>
<td>.377 .259 .578 .102 .552</td>
<td>.569 .316 .764 .122 .816</td>
</tr>
<tr>
<td>.547 .391 .656 .036 .159</td>
<td>.950 .649 .841 .084 .060</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\sigma_1/\sigma_2 = 1$</th>
<th>$\sigma_1/\sigma_2 = \sqrt{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = .5$</td>
<td>$\lambda = .25$</td>
</tr>
<tr>
<td>.937 .981 .706 1.05 1.13</td>
<td>1.05 .680 1.31 .567 1.75</td>
</tr>
<tr>
<td>.774 .725 1.01 .479 .551</td>
<td>.740 .962 .781 .682 .551</td>
</tr>
<tr>
<td>1.34 1.27 1.82 .239 .382</td>
<td>1.70 1.56 1.28 .045 .071</td>
</tr>
<tr>
<td>1.01 1.37 1.50 .506 .675</td>
<td>.832 .969 .885 .827 1.77</td>
</tr>
<tr>
<td>.613 .741 .880 .662 .523</td>
<td>.649 .771 1.12 .421 .476</td>
</tr>
<tr>
<td>1.31 2.61 .061 .275 .157</td>
<td>1.86 1.64 1.77 .182 .131</td>
</tr>
</tbody>
</table>

Figure 4.5

Relative efficiencies of EM, MD, and MM estimators
the EM in general outperforms the MD estimate in nearly all combinations of parameters and configurations (38 out of 40). The MM estimate again seems a possible competitor to the EM for estimating $\lambda$, $\mu_1$ and $\mu_2$, but is very inefficient for estimating the variances.

In conclusion, the EM seems to provide reasonable estimates of all 5 parameters under the configurations studied. Its asymptotic efficiency is reflected in its performance relative to the calculated Cramér-Rao lower bound. The MD estimate, while perhaps performing better under non-normal conditions (see Woodward et al (1984)), does not perform as well as the EM under a true mixture of normals. The MM provides estimates of $\lambda$, $\mu_1$ and $\mu_2$ which compare with the EM in terms of MSE, but performs much worse in estimating the variances $\sigma_1^2$ and $\sigma_2^2$. In addition, while starting points are important, and the asymptotic efficiency of the EM is guaranteed only when starting at a $\sqrt{n}$-consistent starting point, under the configurations above, the naive estimator provided a starting point which resulted in the same EM estimate in nearly all cases. One suspects that this would not be the case for higher overlap, or for variance ratios far from one (see for example, Hathaway (1985)).

For the regression problem, the combination of the MM
and the EM will be used to obtain starting points in each third of the regression as described in Chapter 3. From the results of this simulation, we have learned that under reasonable mixture conditions, these starting points should perform adequately. This will be tested by the simulation in the next section.

§ 4.3 Mixture of Regression Simulation

Computational Considerations

In this section we consider data \((x_i, y_i) \) \(i = 1, \ldots, n\) where

\[
(4.3.1) \quad y_i = \begin{cases} \alpha + \beta x_i + \epsilon_i & \text{with probability } \lambda \\ \alpha_2 + \beta_2 x_i + \epsilon_2 & \text{with probability } (1-\lambda), \end{cases}
\]

and the \(\epsilon_{ji}\) are independent, \(N(0, \sigma^2_j)\). As discussed in Chapter 3, MM estimates and MD estimates for \(\theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)\) are not presently feasible in the mixture of regressions model \((4.3.1)\). This is due to the high dimensionality of the search involved in the numerical minimization of the respective functions. Because of this, the EM was the only estimate considered. The binning procedure (§3.3) was used to generated starting values,
providing the $\sqrt{n}$-consistency needed to ensure asymptotic efficiency of the EM estimator.

For this simulation, programs were written in APL and run on an IBM PC-AT with 80287 numeric co-processor. At each $x_i$, a uniform random variable was generated to assign $y_i$ to one of the regression lines in (4.3.1).

To calculate the information matrices we have again used Romberg's algorithm at $2^9$ points. The information is dependent on the $x$-value, and thus the total information from a random sample $(x_i, y_i)$ $i = 1, \ldots, n$ is the sum of the information at each $x_i$. This requires the numerical integration of 20 integrals for each $x_i$, for a total of 2000 integrals from a sample of size 100. This took approximately 5 hours on the IBM PC-AT. We will use the information matrix to obtain asymptotic confidence intervals for the parameter estimates, as in the 5 parameter case.

**Simulation Design**

Because this simulation study is not a comparison between estimators, we have limited the scope of the study to three cases. The first is designed to resemble the data mentioned in Chapter 1, and analyzed in Chapter 5. Specifically, we set $\theta = (0.5, 2, 0, 0, 1, 0.05^2, 0.03^2)$, where $x_i$ ranges from 1.5 to 3.0. A scatterplot of one of the simulated data
sets with is shown in Figure 4.6. This has small overlap in most of the x range. For comparison, we also investigated two higher overlap situations, θ = (.33, 0, 1, 1, -1, 1^2, 2^2), (Figure 4.7), and θ = (.5, 0, 1, 1, -1, .2^2, .3^2) (Figure 4.8), with x ranging from 0 to 1. For each configuration, there were 100 replications of 100 (x, y) data points. For

![Figure 4.6](image)

**Figure 4.6**

Scatterplot of data with θ = (.5, 2, 0, 0, 1, .05^2, .03^2)
convenience, the $x$-values were taken to be equally spaced in its range. Note that in all three situations, the true regression lines do cross within the range of the $x$'s. Thus, the binning procedure should choose the crossed pair to start the EM.

![Graph showing data points with labels](image)

**Figure 4.7**

Scatterplot of data with $\theta = (.33, 1, 0, 1, -1, .1^2, .2^2)$
Figure 4.8

Scatterplot of data with \( \theta = (0.5, 1, 0, 1, -1, 2^2, 3^2) \)

Simulation Analysis

In this section we compare the performance of the EM estimates of \( \theta \) in the three simulation configurations to the
asymptotic performance predicted by the results of Chapter 3. We also compare the results to the maximum likelihood estimates which would have been obtained given knowledge of the group membership. This will give a qualitative comparison to "ideal" performance.

A first concern is how well the binning estimates did in choosing the correct pair of lines, and how robust the EM was in finding the correct maximum when they did not.

Configuration 1: \( \theta = (0.5, 2.0, 0.0, 1.05^2, 0.03^2) \);
\( x_1 \) ranges from 1.5 to 3.0. (See Figure 4.6)

The binning procedure chose the correct (i.e. the crossed lines) 92% of the time. However, regardless of which starting point was used (crossed or straight), the EM chose the same final crossed estimates 100% of the time.

Configuration 2: \( \theta = (0.33, 0.1, 1.0, -1.1^2, 0.2^2) \);
\( x_1 \) ranges from 0.0 to 1.0. (See Figure 4.7)

The binning procedure chose the correct (i.e. the crossed lines) 91% of the time. Again, regardless of starting point, the EM choose the same final
crossed estimates 100% of the time.

Configuration 3: θ = (.5, 0, 1, 1, -1, 1.2^2, 3^2);
θ_i ranges from 0.0 to 1.0. (See Figure 4.8)
The binning procedure chose the correct
(i.e. the crossed lines) only 63% of the
time. The EM always chose a crossed
solution. However in 14% of the runs,
the EM estimate (while crossed) was
different for the two starting points.

Thus we conclude that while the binning procedure does
not choose the correct starting lines 100% of the time, the
EM is quite robust with respect to the starting point for
these configurations. A possible explanation for the
failure of the binning procedure to choose with 100%
accuracy is that we have violated the condition of Prop-
osition 3.3.6 in which we require that the cross point not
be contained in any of the three intervals. We took for
convenience the first, middle and last thirds of the data
for our three bins. Since the true lines do cross in one of
these intervals, we have violated this assumption in every
case. As mentioned in Chapter 3, a better procedure would
perhaps be to estimate the cross point, and to choose three

106
bins which do not contain this point. This was not done for our simulations, but will be done for the data in Chapter 5.

For each configuration, the information matrix was obtained. Each is listed below with the sample covariance matrix of the 100 runs listed below it for comparison. For each configuration, we examine the ratios of the asymptotic variances (the diagonal of the inverse of the information matrix) to the sample variances. We also compare both these variances with the expected variances if group membership of the \( y_i \) were known. We shall refer to these variances as the "full information" variances. From configuration 2, 33 of the \((x, y)\) points (on the average) will be from line 1, with \( \alpha_1=0, \beta_4=1 \) and \( \sigma^2=.1 \) while the remaining 67 will be from line 2: \( \alpha_2=1, \beta_2=-1 \) and \( \sigma^2=.2 \). Using the usual least squares estimates, the "full information" covariance matrix of \( \hat{\alpha}_j \) and \( \hat{\beta}_j \) is \( \sigma_j^2 (X'X)^{-1} \), where \( X \) is the usual design matrix of ones and \( x_i \), for the \( x_i \) belonging to line \( j \) \((j=1, 2)\). To calculate these covariance matrices, we take \( \lambda \) (or \( 1-\lambda \)) of the \( x_i \) randomly to form \( X \). Thus the matrices are only one realization of the full information, but are, however, informative for comparative purposes. Summaries of the performance of the EM estimator in each configuration follows.
Configuration 1: $\theta = (0.5 2 0 0 1, 0.05^2, 0.03^2)$

$I^{-1}(\theta)$:

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\sigma_1^2$</th>
<th>$\sigma_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00250</td>
<td>-3.13E-12</td>
<td>-1.84E-12</td>
<td>0</td>
<td>0</td>
<td>9.31E-13</td>
<td>-3.22E-13</td>
</tr>
<tr>
<td>0.00141</td>
<td>3.16E-11</td>
<td>-0.000602</td>
<td>-1.35E-11</td>
<td>-4.10E-15</td>
<td>1.96E-13</td>
<td></td>
</tr>
<tr>
<td>0.000507</td>
<td>-1.35E-11</td>
<td>-0.000217</td>
<td>-3.33E-13</td>
<td>-2.31E-13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.000267</td>
<td>5.98E-12</td>
<td>0</td>
<td>0</td>
<td>2.50E-7</td>
<td>-5.81E-14</td>
<td></td>
</tr>
<tr>
<td>9.60E-5</td>
<td>0</td>
<td>0</td>
<td>3.24E-8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sample Covariance Matrix:

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\sigma_1^2$</th>
<th>$\sigma_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00300</td>
<td>-0.000234</td>
<td>8.07E-5</td>
<td>0.000103</td>
<td>-1.43E-5</td>
<td>3.96E-7</td>
<td>9.85E-8</td>
</tr>
<tr>
<td>0.000608</td>
<td>5.43E-5</td>
<td>-0.000257</td>
<td>-2.54E-5</td>
<td>2.57E-7</td>
<td>2.56E-7</td>
<td></td>
</tr>
<tr>
<td>0.00171</td>
<td>-5.98E-6</td>
<td>-0.000712</td>
<td>5.36E-7</td>
<td>-8.27E-8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.000111</td>
<td>3.42E-6</td>
<td>-1.29E-7</td>
<td>4.20E-8</td>
<td>1.68E-8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.000304</td>
<td>-3.39E-7</td>
<td>-4.44E-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The standard deviations from the information matrix, the sample covariance matrix and full information are:

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\sigma_1^2$</th>
<th>$\sigma_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Info</td>
<td>0.0500</td>
<td>0.0375</td>
<td>0.0225</td>
<td>0.0163</td>
<td>0.00980</td>
<td>0.000500</td>
</tr>
<tr>
<td>Sample</td>
<td>0.0548</td>
<td>0.0247</td>
<td>0.0414</td>
<td>0.0106</td>
<td>0.0174</td>
<td>0.000205</td>
</tr>
<tr>
<td>Full</td>
<td>0.0375</td>
<td>0.0222</td>
<td>0.0163</td>
<td>0.00979</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For this configuration, we see that the asymptotic variances are nearly the same as the full information variances. This is due to the extremely low overlap of the configuration. The only curious feature of the comparison is that for $\alpha_1$, $\beta_1$ and $\sigma_1$, the simulation variances are significantly ($\alpha = .05$) less than the predicted asymptotic variances. This may be due to the sample correlation of the estimates which in general are more correlated than predicted by the information matrix. This behavior was not seen in the next two configurations.

Configuration 2: $\theta = (.33 0 1 1 -1 .1^2 .2^2)$

$I^{-1}(\theta) =$

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\sigma_1^2$</th>
<th>$\sigma_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00222</td>
<td>4.28E-6</td>
<td>4.90E-6</td>
<td>0</td>
<td>0</td>
<td>1.17E-6</td>
<td>-3.10E-5</td>
</tr>
<tr>
<td>0.00125</td>
<td>1.63E-5</td>
<td>-0.0016</td>
<td>-2.38E-5</td>
<td>-1.13E-6</td>
<td>-2.26E-6</td>
<td>-3.22E-6</td>
</tr>
<tr>
<td>0.00244</td>
<td>-2.38E-5</td>
<td>-0.00365</td>
<td>1.16E-6</td>
<td>-3.22E-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00368</td>
<td>4.71E-5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00722</td>
<td>6.40E-6</td>
<td>-7.01E-7</td>
<td>4.98E-5</td>
<td>109</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample Covariance Matrix:

<table>
<thead>
<tr>
<th>λ</th>
<th>α1</th>
<th>α2</th>
<th>β1</th>
<th>β2</th>
<th>σ₁²</th>
<th>σ₂²</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00346</td>
<td>0.000493</td>
<td>0.000498</td>
<td>-0.000801</td>
<td>-0.000556</td>
<td>6.54E-5</td>
<td>-8.20E-5</td>
</tr>
<tr>
<td>0.00172</td>
<td>0.000341</td>
<td>-0.00267</td>
<td>-0.000683</td>
<td>1.51E-5</td>
<td>-1.50E-5</td>
<td></td>
</tr>
<tr>
<td>0.00226</td>
<td>-0.000952</td>
<td>-0.00330</td>
<td>1.21E-5</td>
<td>-7.46E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00571</td>
<td>0.00175</td>
<td>-5.33E-5</td>
<td>6.92E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00714</td>
<td>-7.19E-6</td>
<td>8.41E-6</td>
<td>-7.90E-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8.31E-6</td>
<td>7.46E-5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The standard deviations from the information matrix, the sample covariance matrix and full information are:

<table>
<thead>
<tr>
<th>λ</th>
<th>α1</th>
<th>α2</th>
<th>β1</th>
<th>β2</th>
<th>σ₁²</th>
<th>σ₂²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Info</td>
<td>.0471</td>
<td>.0353</td>
<td>.0494</td>
<td>.0607</td>
<td>.0850</td>
<td>.00253</td>
</tr>
<tr>
<td>Sample</td>
<td>.0588</td>
<td>.0414</td>
<td>.0476</td>
<td>.0756</td>
<td>.0845</td>
<td>.00288</td>
</tr>
<tr>
<td>Full</td>
<td>.0350</td>
<td>.0486</td>
<td>.0601</td>
<td>.0841</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For this configuration, again a relatively low overlap configuration, the full information variances are comparable to the predicted variances. The sample variances are in general slightly larger than predicted, but none are significantly so (overall α = .05).
Configuration 3: \( \theta = (.5 \ 0 \ 1 \ 1 \ -1 \ .2^2 \ .3^2) \)

\[ I^{-1}(\theta) = \]

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \alpha_1 )</th>
<th>( \alpha_2 )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \sigma_1^2 )</th>
<th>( \sigma_2^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00318</td>
<td>0.000499</td>
<td>0.000848</td>
<td>0</td>
<td>0</td>
<td>0.000157</td>
<td>-0.000430</td>
</tr>
<tr>
<td>0.00419</td>
<td>0.00126</td>
<td>-0.00578</td>
<td>-0.00131</td>
<td>0.000127</td>
<td>-0.000288</td>
<td></td>
</tr>
<tr>
<td>0.00962</td>
<td>-0.00131</td>
<td>-0.0134</td>
<td>0.000181</td>
<td>-0.000556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0114</td>
<td>0.00260</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0265</td>
<td>0.000110</td>
<td>-8.28E-5</td>
<td>0.000627</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sample Covariance Matrix:

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \alpha_1 )</th>
<th>( \alpha_2 )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \sigma_1^2 )</th>
<th>( \sigma_2^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0122</td>
<td>0.00495</td>
<td>0.0110</td>
<td>-0.00959</td>
<td>-0.0207</td>
<td>0.00116</td>
<td>-0.00199</td>
</tr>
<tr>
<td>0.00687</td>
<td>0.00594</td>
<td>-0.0114</td>
<td>-0.0111</td>
<td>-0.0364</td>
<td>0.000624</td>
<td>-0.000893</td>
</tr>
<tr>
<td>0.0219</td>
<td>-0.00110</td>
<td>-0.00364</td>
<td>-0.00106</td>
<td>-0.00239</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0228</td>
<td>0.0216</td>
<td>-0.000123</td>
<td>0.00167</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0717</td>
<td>-0.00210</td>
<td>0.00449</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.000229</td>
<td>-0.000234</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.000754</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The standard deviations from the information matrix, the sample covariance matrix and full information are:

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \alpha_1 )</th>
<th>( \alpha_2 )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \sigma_1^2 )</th>
<th>( \sigma_2^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Info</td>
<td>0.0565</td>
<td>0.0640</td>
<td>0.0981</td>
<td>0.1069</td>
<td>0.1628</td>
<td>0.0105</td>
</tr>
<tr>
<td>Sample</td>
<td>0.1104</td>
<td>0.0829</td>
<td>0.1479</td>
<td>0.1511</td>
<td>0.2677</td>
<td>0.0151</td>
</tr>
<tr>
<td>Full</td>
<td>0.0559</td>
<td>0.0874</td>
<td>0.0980</td>
<td>0.1473</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For this configuration, we see that the asymptotic standard deviations are about 25% higher than the full information standard deviations, reflecting the comparatively high overlap of this configuration. Five of the sample standard deviations are significantly larger than the predicted standard deviations the exceptions being $\sigma_1$ and $\sigma_2$. (overall $\alpha = .05$)

Conclusions

Three simulations have been presented to see how the combination of binning starting value and EM estimation works in small data sets. The parameters of these simulations were chosen to reflect a very low overlap (corresponding to the data examined in Chapter 5), moderate overlap and high overlap.

The binning procedure, as expected, chooses the correct pair of starting lines with increasing frequency as the overlap decreases. As we have mentioned, the binning procedure should be modified to insure that the cross point does not lie in any of the bins. This was not done. However, the EM algorithm correctly chose a crossed line solution in all 300 simulations even when the binning procedure started at a straight line starting solution. Thus, we conclude that the EM is quite robust to "reasonable"
starting points. We should reiterate that it is, however, almost always possible for the EM to converge to a spurious maximum if started far enough away from the true parameters. For practical use, we suggest estimating the cross point crudely and choosing three bins which do not contain the cross point. As in the five parameter case, one should also start the EM from various other starting points to compare the likelihoods of the solutions obtained, if different.

The asymptotic variances and covariances were obtained from the information matrix for each configuration. These predicted variances were compared to the sample variances actually obtained from the simulation. In general, the sample variances are all within the right order of magnitude, with the closest agreement from configuration 2. The variances of the slopes and intercepts given that the population membership is known were also calculated and compared to the predicted asymptotic variances. As expected, they are smaller than the predicted asymptotic variances, with the largest discrepancies obtained under the highest overlap. The asymptotic variances can be used to derive asymptotic confidence intervals for each parameter.

The EM in combination with the binning procedure starting point provides an asymptotically efficient procedure for obtaining estimates of the seven mixture of regression
parameters simultaneously. The information matrix can be numerically approximated and used to obtain asymptotic confidence intervals. In each of the configurations we examined the procedure provided estimates with variances approaching the predicted Cramér-Rao lower bounds. We will use this procedure to analyze the data which motivated this dissertation in the following chapter.
Chapter 5

§ 5.1 Introduction

In this chapter, we present an analysis of the data which motivated our research. We review the scientific experiment which generated the data and explain why the mixture of regressions model (4.3.1) may be appropriate. We review the method of obtaining an estimate for the regression mixture parameter \( \theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2) \), by carrying out the estimation procedure in detail on one data set.

We illustrate some common aspects of the data using one of the subjects and provide summaries of parameter estimates, confidence intervals and individual peculiarities for each of 5 subjects. Various aspects of the model are tested including a test of the originally hypothesized model (1.1.3) with \( \alpha_1 = 2.0, \alpha_2 = 0.0, \beta_1 = 0.0 \) and \( \beta_2 = 1.0 \). This corresponds to the scientific hypothesis of interest. A summary of our conclusions is presented in the final section.

§5.2 The experiment

The data which motivated the study of mixtures of regressions resulted from an experiment in musical perception at the Center for Research in Music and Acoustics (CRMA) at Stanford University, carried out by Elizabeth Cohen (Cohen
(1980)). A musical pitch consists of a fundamental tone at a certain frequency plus a series of other frequencies called harmonics. Western musical instruments in general produce harmonics which are at integer multiples of the fundamental. For example, a flute playing an A (at 440 cps) may actually produce the fundamental frequency 440 cps plus a series of harmonics at 880 cps (an octave), 1320 cps, 1760 cps etc. In eastern music, this is often not the case. Thus to the western ear, these notes may be perceived as "out of tune". If the harmonics are at ratios very different from (the normal) integer multiples, one may even experience difficulty in discerning the frequency of the fundamental tone.

Suppose a tone, say an A at 440 cps is generated electronically, and one adds to it a series of harmonics. The ratio of the first harmonic to the fundamental will be called the "tuning ratio" and will be denoted by $x_1$. The rest of the harmonics are related to this ratio logarithmically. For details see Cohen (1980). If $x_1 = 2.0$, (the octave) the harmonic pattern will be the usual integer multiples of the fundamental. If $x_1 < 2.0$, the series is said to be compressed and if $x_1 > 2.0$ it is said to be stretched.

The scientific question of concern to Cohen was the following. Present a subject with a fundamental tone plus a series of harmonics based on a tuning ratio of $x_1$, and ask
him to tune an adjustable tone to the octave above the fundamental tone. We will denote the ratio of the frequency of the adjusted tone to the fundamental tone by $y_1$, and call it the response ratio. If $x_1 \neq 2.0$, will the subject still perceive the tuning at the octave ratio of 2:1, resulting in $y_1 = 2.0$, or will the harmonic series dictate the perceived tuning, resulting in $y_1$ equal to the tuning ratio $x_1$?

For example, suppose we choose A at 440 cps as the fundamental frequency. In addition we add a harmonics series, the first of which is at 660 cps. Thus the tuning ratio $x_1$ is 1.5. We now ask the musician to tune a pure, adjustable tone to the octave above the fundamental frequency. If he is not influenced by these harmonics, he will tune the octave to 880 cps, resulting in a response ratio $y_1$ of $880/440 = 2.0$ which indeed is an octave. He may however perceive the correct tuning based on the harmonics he hears to be at the tuning ratio instead. In this case one expects him to tune the tone to 660 cps, resulting in $y_1 = 1.5 = x_1$. Two theories of musical perception offered these two predictions. The experiment of Cohen was designed to test which (if either) of the theories was the better predictor.

To test this, Cohen subjected 5 trained musicians to a series of trials (which ranged in number from 112 to 360).
For each subject, trial $i$ results in a data pair $(x_i, y_i)$, consisting of the tuning ratio $x_i$ chosen by the experimenter and the response ratio $y_i$. By one theory, $y_i$ should always be 2.0. That is, no matter what the tuning ratio of the harmonics, the musician will tune to the (actual) octave at ratio 2:1. The other theory predicts that $y_i$ should be $x_i$; the perceived octave is actually the tuning ratio. In the experiment, $x_i$ was chosen to range from 1.35 to 3.0.

In Figure 5.1 we display a scatterplot of the pairs $(x_i, y_i)$ from 112 trials for one subject (JMS). One is struck by what appears to be both a flat response at 2.0 and a response of slope 1. Cohen thus theorized that a proportion $\lambda$ of the time the subject perceived the (actual) octave and the rest of the time they chose the tuning ratio for the response ratio. This then resulted in a synthesis of the two theories of musical perception. It was desired to find a method for analyzing data with structure like that exhibited in Figure 5.1. Cohen (1980) presented an exploratory data analysis in which he tacitly assumed the following model:

\[
\begin{align*}
    y_i &= \begin{cases} 
    2.0 + \varepsilon_{1i} & \text{with probability } \lambda; \quad \varepsilon_{1i} \sim N(0, \sigma_1^2) \\
    x_i + \varepsilon_{2i} & \text{with probability } 1-\lambda; \quad \varepsilon_{2i} \sim N(0, \sigma_2^2)
    \end{cases}
\end{align*}
\]
Figure 5.1

Scatterplot of data from JMS
The present research resulted from a desire to analyze more methodically data from the generalization of (5.2.2) as a mixture of regressions:

\[(5.2.3)\]

\[
y_1 = \begin{bmatrix}
\alpha + \beta x_{i1} + \epsilon_{i1} \\
\alpha + \beta x_{i2} + \epsilon_{i2}
\end{bmatrix}
\text{with probability } \lambda
\]

\[(1-\lambda),\]

where the \(\epsilon_{ij}\) are independent, \(N(0, \sigma_j^2)\). In this case, the slopes and intercepts are all assumed unknown. This is our original model (1.1.1).

§5.3 Data Analysis

Before proceeding with the estimation of the seven parameters \(\theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)\) in equation (5.2.3) for each subject, we discuss some aspects of the data. We first discuss elements of the analysis shared by all five subjects. We then discuss the subjects one at a time, discussing the parameter estimates, confidence intervals and model validation procedures. Finally, we discuss the estimates as a whole and draw conclusions about the experiment and directions for further analysis.

The scatterplots of the full data set for each subject are displayed in Figures 5.1, 5.11, 5.12, 5.13 and 5.15. For two subjects, JMS and KIP (Figures 5.1 and 5.15) 112, and 120
trials, respectively, were performed. For the remaining three JWG, GWM and SCH, 360 trials were performed. In all five cases, the tuning ratio, \( x_i \), ranged from 1.35 to 3.00, in intervals of .05. Unfortunately, for our purposes, the experimenter did not test any subject in the range \( x = (1.6 \text{ to } 1.85 \text{ or at the points } 2.15, 2.25, 2.55, 2.65 \text{ and } 2.95) \). Otherwise, all ratios from 1.35 to 3.00 were tested for each subject at intervals of .05. In the range 1.9 to 2.1, observations were also taken at points 1.91, 1.93, ... 2.09.

In three of the scatterplots, Figures 5.11, 5.13, and 5.15, (KIP, JWG and SCH) one observes several data points near \( y = 4.0 \) and \( y = 3.0 \). These points are believed to be due to a phenomenon known as octave doubling. A subject may confuse octaves, and mistakenly tune the tone to 4 times the fundamental instead of 2 times the fundamental frequency. If he confuses the fifth above the tone, he may mistakenly tune the tone to 3 times the fundamental instead of 2. Neither of these mistakes was of particular interest to the experimenter, and thus these points were viewed as outliers and removed from the final analysis. If left in, they seriously change the estimates of all seven parameters.

For SCH, (Figure 5.11), there are many \( y_i \) points between 2.7 and 3.1 for \( x_i = 2 \) to 2.5. Due to the quantity of these points we ran the analysis both with and without them. As a
possible explanation for these numerous points away from the
two apparent lines, the experimenter stated simply that "he
is a percussionist" (Cohen (1985)). The scatterplots after
outliers are removed are shown for JWG and KIP in Figures
5.14 and 5.16.

For each subject, 500 bootstrap replications were
performed on the \((x_i, y_i)\) pairs, \(i = 1, \ldots, n_j\) (\(n_j = 112, 120\) or
360). The bootstrap standard deviations are used to provide
confidence intervals for the parameter estimates. They are
compared to the asymptotic confidence intervals derived from
the information matrix, where the estimated parameters are
used as the true parameters. These confidence intervals are
used to test whether the model (5.2.2) provides a plausible
model for the data.

Subjects

We first discuss subject JMS, whose data are shown in
Figure 5.1. For JMS, we will describe the estimation
procedure in detail. For the other four subjects, we
summarize the estimation procedure results and discuss
differences in their data.

JMS

No points appear to be instances of octave doubling in
Figure 5.1. By Proposition 3.3.6, we will use the binning procedure of §3.3 to find a $\sqrt{n}$-consistent starting value, $\hat{\theta}_n$, for $\theta = (\lambda, \alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$. We will then start the EM algorithm (see §2.2) at $\hat{\theta}_n$, whose limit $\theta^*_n$ will be asymptotically efficient for estimating $\theta$.

The binning procedure is started by choosing three non-overlapping intervals (bins) on the x-axis, which do not contain the cross point (the x-value where the two lines cross). From Figure 5.1, we see that the cross point may be near $x = 2.0$. We therefore choose the following three intervals: $I^{(1)} = [1.35 \text{ to } 1.91]$, $I^{(2)} = [2.60 \text{ to } 3.00]$ and $I^{(3)} = [2.01 \text{ to } 2.30]$. We have labeled the bins this way for consistency with the notation in §3.3. In each bin, we replace the pair $(x_i, y_i)$ by $(\bar{x}, y_i)$ where $\bar{x}$ is the average of the x-values in the bin. This is illustrated for this data set in Figure 5.2.

In each bin, we regard the $y_i$ as a one-dimensional mixture of (approximate) normals and estimate $\theta^{(j)} = (\lambda, -\mu_1^{(j)}, \mu_2^{(j)}, \sigma_1^2(j), \sigma_2^2(j))$, for $j=1,2,3$. The estimates in each bin of the means are shown in Figure 5.2. These are denoted by $\hat{\theta}_i^{(j)}$ for $i=1,2$ and $j=1,2,3$. The estimates of $\lambda$ from bin 1, bin 2 and bin 3 are 0.04, 0.85 and 0.85 respectively. The extremely low estimate from bin 1 is due to the single point which appears below the rest in bin 1. The
Figure 5.2

Bins for JMS data
implications of this will be discussed at the end of this section.

We now want to estimate the intercepts $\alpha_i$ and slopes $\beta_i$ $i = 1,2$ from these binned estimates. From the mean estimates of the two outer bins, $I(1)$ and $I(2)$, we can form two pairs of lines. One pair connects $\hat{\mu}_2 (1)$ with $\hat{\mu}_2 (2)$, the two "higher" mean estimates and $\hat{\mu}_1 (1)$ with $\hat{\mu}_1 (2)$, the two "lower" mean estimates. This pair of lines which does not cross between $\bar{x}(1) = 1.57$ and $\bar{x}(2) = 2.79$, is shown in Figure 5.3. We refer to this pair as the "straight" pair of lines, and denote its intercepts and slopes by $(\hat{\alpha}_{1s}, \hat{\alpha}_{2s}, \hat{\beta}_{1s}, \hat{\beta}_{2s})$. We use (3.3.4) and (3.3.6) to calculate these and find that

$(\hat{\alpha}_{1s}, \hat{\alpha}_{2s}, \hat{\beta}_{1s}, \hat{\beta}_{2s}) = (1.15, 0.97, 0.31, 0.63)$.

Alternatively, we could form a pair of lines by connecting the lower mean in bin 1 with the higher mean in bin 2 and vice versa. This pair, which will cross between 1.57 and 2.79, is shown in Figure 5.4. We refer to this pair as the crossed pair, and denote its intercepts and slopes by $(\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c})$. By (3.3.5) and (3.3.7) we have $(\hat{\alpha}_{1c}, \hat{\alpha}_{2c}, \hat{\beta}_{1c}, \hat{\beta}_{2c}) = (1.90, 0.22, 0.04, 0.90)$.

We must now choose between these two pairs. To do this, we use the mean estimates from bin 3, the middle bin. Start with the straight pair of lines. Refer to Figure 5.3, and note that at $\bar{x}(3) = 2.1$, the upper line overestimates the
Figure 5.3
Straight pair of lines: JMS data
upper mean of the y-values in bin 3, while the lower line underestimates the lower mean of the y-values in bin 3. We calculate absolute distance from the upper mean estimate \( \hat{\mu}_2^{(3)} \) to the upper line and denote it by \( d_{s2} \). (See equation (3.3.8)). We similarly calculate the absolute distance from \( \hat{\mu}_1^{(3)} \) to the lower line and denote it by \( d_{s1} \). These are shown in Figure 5.3. We then define the total error of this pair as \( d_s = d_{s1} + d_{s2} \). For JMS, \( d_s = 0.293 \).

We repeat this procedure for the crossed pair shown in Figure 5.4. Here we find that for JMS, \( d_c = d_{c1} + d_{c2} = 0.204 \). We choose the pair with the smaller total error. We showed in lemma 3.3.3 that this is asymptotically the correct choice. Therefore, the starting intercepts and slopes which we now denote by \( (\hat{\alpha}_1^c, \hat{\alpha}_2^c, \hat{\beta}_1^c, \hat{\beta}_2^c) \) are the crossed line estimates, \( (\hat{\alpha}_1^c, \hat{\alpha}_2^c, \hat{\beta}_1^c, \hat{\beta}_2^c) = (1.90, 0.22, 0.04, 0.90) \). The initial estimate for \( \lambda \) is obtained by averaging \( \lambda \) in all three bins (or theoretically from any one bin). For JMS, \( \lambda = 0.58 \). The initial standard deviation estimates are obtained via equation (3.3.13) and equal 0.012 and 0.010. Thus, the starting value \( \hat{\theta}_n \) for JMS is (0.58, 1.90, 0.22, 0.04, 0.90, -0.010^2, 0.12^2).

This estimate is now used as the starting value for the EM algorithm. The EM algorithm limit obtained was:

127
Figure 5.4
Crossed pair of lines: JMS data
\[ \lambda \quad \alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \sigma_1 \quad \sigma_2 \]

\[ \theta^* = (0.597, \ 1.902, \ -0.00462, \ 0.0402, \ 1.003, \ 0.00716, \ 0.0901), \]

compared with

\[ \hat{\theta}_n = (0.58, \ 1.90, \ 0.22, \ 0.04, \ 0.90, \ 0.010, \ 0.12) \]

The data with the EM lines superimposed are shown in Figure 5.5.

The EM algorithm does not assign the points to each line, but rather for each point \( y_i \), it assigns probabilities \( W_{1i} \) and \( W_{2i} \) \( (W_{1i} + W_{2i} = 1) \) of belonging to the first or second line of (5.2.3). See equation (2.2.11). Because of this, we cannot calculate residuals as in an ordinary linear regression situation. (See also Cox and Snell (1968)). A naive approach to defining a residual in the mixture situation, might be to calculate the distance from each estimated line, divide by the appropriate standard deviation and then take the minimum as the standardized residual. This approach was tried. A plot of the residuals versus \( x_1 \) (Figure 5.6) illustrates the danger of doing this. The points near the cross point seem to be incorrectly assigned, and the straight line pattern near \( x_1 = 2.0 \) is evidence of this.
Figure 5.5

EM limit lines for JMS data
Figure 5.6

Plot of "residuals" vs. $x_1$: JMS data
misassignment.

Rather, we shall define residuals in the following way. We assign points to each line on the basis of whether the final EM probability $W^*_{ji}$ ($j=1,2$) is greater or less than .5. Then the residuals are calculated in the usual way, $r_i = Y_i - (\alpha_j + x_i \beta_j)$ where $j$ is 1 or 2 if $W^*_{i1}$ is > or < .5 respectively. The $r_i$ will be approximately a mixture of normals each with mean 0, since the true errors $e_{ji}$ are a mixture of $N(0,\sigma_1^2)$ and $N(0,\sigma_2^2)$. Figure 5.7 is a stemleaf diagram of the $r_i$. The standardized residuals, $e_i$ are defined to be $e_i = r_i/\hat{\sigma}_j$. Figure 5.8 is a plot of the $e_i$ versus $x_i$. No obvious patterns remain. The standardized residuals should be approximately normally distributed. (The true residuals if population membership known would be normally distributed). Figure 5.9 is a stemleaf diagram of the $e_i$ in which the normality is clearly evident. The only standardized residual greater than 2.33 in absolute value is $e_i = -3.589$.

This point was removed to see the influence on the EM estimates. The removal of the point resulted in a $10\%$ decrease in the standard deviation estimate of the line with slope 0.04 and intercept 1.90. The other six parameters were not affected. Because no scientific justification could be made for deleting this point, it was left in further
Figure 5.7
Frequency histogram of residuals from JMS data
Figure 5.8

Standardized residuals vs. $x_1$: JMS data

134
Confidence intervals are obtained from the standard deviations of bootstrap replications of the data. (See Efron (1977) for an introduction to the bootstrap). The data \((x_1, y_1)\) are resampled with replacement and \(\theta\) is re-estimated by the above procedure at each resampling (replication). The standard deviations of the parameter estimates of these replications then provide estimates of the actual standard deviations of the estimators.

In a multiple regression situation, one is faced with a choice of two methods for resampling the data \((x_1, y_1)\). The first, called the regression model (Freedman (1981)) fixes the \(x_1\) and resamples from the centered residuals \(e_{1}^{*}\) to obtain the bootstrapped pairs \((x_1, y_1)\) where \(y_{1}^{*} = \hat{\alpha} + \hat{\beta} x_1 + e_{1}^{*}\). The second model, the correlation model, treats the \(x_1\) as random also, and resamples with replacement from the pairs \((x_1, y_1)\). Because of the problem with defining residuals discussed above, we have chosen to resample the \((x_1, y_1)\) directly, using the so-called correlation model. Freedman (1981) showed that asymptotically, the bootstrap approximation to the distribution of the least squares estimates is valid under both models. We will not formally extend these results to the mixture of regression case, but will proceed to obtain bootstrap standard errors and compare them to the standard errors obtained from the information matrix.
The data \((x_1, y_1)\) from JMS was resampled with replacement 500 times. The EM estimate was obtained for each replication. The average values of \(\hat{\theta}\) are shown with the original data EM estimates for comparison:

**Original EM estimate:**

\[
\lambda \quad \alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \sigma_1 \quad \sigma_2 \\
\hat{\theta} = (0.403, -0.00462, 1.902, 1.003, 0.0402, 0.00716, 0.0901).
\]

**Average of 500 Bootstrapped EM estimates:**

\[
\hat{\theta} = (0.405, -0.00429, 1.900, 1.003, 0.0416, 0.00679, 0.0880) \\
(0.053) (0.014) (0.060) (0.0071) (0.025) (0.0010) (0.011),
\]

where the bootstrap sample standard deviations of these 500 replications appear in parentheses below each parameter.

For comparison, the information matrix was calculated for these \(x_1\) at \(\hat{\theta} = \hat{\theta}^*\), the EM estimate. The bootstrap standard deviations and the information matrix standard deviations are:

\[
\lambda \quad \alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \sigma_1^2 \quad \sigma_2^2 \\
\text{Boot} \quad 0.053 \quad 0.0144 \quad 0.060 \quad 0.0071 \quad 0.025 \quad 1.32E-5 \quad 1.86E-3
\]
Thus we see that the information matrix standard errors are smaller than the bootstrap with the largest discrepancies occurring for $\alpha_1$ and $\beta_1$. Here the ratios of the bootstrap to the information matrix standard errors are 3.54 and 3.24. This may be due to a few data points which are still high influence points for determining the slope and intercept of the slope 1 line. Because the standard deviation around this line is so small, in those replications where the high influence points are selected or even repeated, the change in the estimates of $\alpha_1$ and $\beta_1$ will be higher than predicted from the information matrix estimates which assume that all of the assumptions of the mixture model are true. We will use the bootstrap standard errors to derive approximate confidence intervals for the parameters by

$$\theta_i \in (\hat{\theta}_i \pm 1.96 \, s_i^*),$$

where $s_i^*$ is the bootstrap sample standard deviation. For JMS, these 95% confidence intervals are:

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\alpha_1$</th>
<th>$\sigma_2$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.300</td>
<td>- .033</td>
<td>1.79</td>
<td>0.99</td>
<td>- .00964</td>
<td>.00516</td>
<td>.0692</td>
</tr>
<tr>
<td>.507</td>
<td>.024</td>
<td>2.02</td>
<td>1.02</td>
<td>.0900</td>
<td>.00917</td>
<td>.111</td>
</tr>
</tbody>
</table>
This now enables us to test the hypothesis that the data are from a mixture of the two perception paradigms postulated in the introduction. That is, we test whether the mixture is $\alpha_1 = 2.0$, $\beta_1 = 0.0$ and $\alpha_2 = 0.0$, $\beta_2 = 1.0$. From the confidence intervals, we see that indeed, all four postulated values are within the confidence limits.

The final assumption to be considered is whether $\lambda$ is constant over the range of $x$. To test this, we have used the weights $W^*_{11}$, the EM probabilities of belonging to the first line as indicators for $\lambda$. If we plot $W^*_{11}$ versus $X$ (Figure 5.9) we see that most of the points with high values of $W^*_{11}$, (high probability of membership to the slope 1 line) have $x$ values between 1.90 and 2.10. Below $x = 1.90$, there are 20 points, all of whom have $W^*_{11}$ near 0, indicating that all of these points are from the "flat line" of intercept 2, slope 1. Similarly, for $x$ above 2.10, 39 of the 49 points have $W^*_{11}$ near 0. To graphically illustrate this, we have plotted (Figure 5.10) a running average of $W^*_{11}$ versus $x_1$. (Thus, the first value is $(x_1, W^*_{11}) = (0, 0.35)$ while the last is $(x_{112}, EW^*_{11}/112) = (\lambda, 3.00)$. It is graphically clear that $\lambda$ is not constant over the $x_1$. This has profound implications for the scientific question at hand. Apparently, for JMS, the ability to hear the octave (represented by the points on the flat line), increases for tuning ratios ($X_1$) which lie
farther away from $x_i = 2.00$ in both directions. One could run separate analyses, and obtain separate confidence

Figure 5.9

$W^*_i$ plotted above JMS data
Figure 5.10

Running average of $\hat{\lambda}$ plotted above JMS data
intervals for $\lambda$ from these three regions to test this, but three things seem clear from our present analysis. First, from our exploratory analysis, the assumption of constant $\lambda$ is clearly violated. Secondly, the binning estimates will suffer from non-constant $\lambda$, since for the first third of the data, nearly all the $y_i$ are from the same line. This estimate of $\lambda$ from the first bin was 0.04 while from the other two it was 0.85 due to varying $\lambda$. Thus the binning procedure may be inefficient at best, and totally misleading at worst. For this data set, however, both the binning starting value and the resulting EM estimate appear to be unaffected by this problem.

The EM seems to be robust to violations of this assumption since there appears to be no pattern in the residuals indicating that this problem caused the EM to allocate points to the wrong line. On the contrary, it seemed able to estimate the overall $\lambda$ probability despite the fact that $\lambda$ was dependent on $x$.

SCH

In looking at the scatterplot of the data from SCH in Figure 5.11, we see that SCH is subject to the octave doubling phenomenon by the abundance of points near $y_i = 4.0, 3.0$ and $1.5$ which do not seem close to either apparent line.
Figure 5.11

Scatterplot of SCH data
These points are circled. We first estimated θ with all 360 
(x₁, y₁) points in and obtained:

\[\begin{array}{ccccccc}
\lambda & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 & \sigma_1 & \sigma_2 \\
\hline
.721 & 1.951 & 0.484 & 0.032 & 0.827 & 0.042 & 0.337
\end{array}\]

We then removed the 30 circled points which the investigator 
felt were indicative of octave doubling and reestimated θ:

\[\begin{array}{ccccccc}
\lambda & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 & \sigma_1 & \sigma_2 \\
\hline
.763 & 1.958 & .330 & .0290 & .874 & .0446 & .113
\end{array}\]

The 500 bootstrap replications gave the following approximate 
95% confidence intervals:

\[\begin{array}{ccccccc}
\lambda & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 & \sigma_1 & \sigma_2 \\
\hline
.702 & 1.923 & .197 & .0140 & .819 & .0387 & .0789 \\
.824 & 1.993 & .462 & .0442 & .930 & .0506 & .147
\end{array}\]

Testing the model (5.2.2), we see that the line which should 
have intercept 2 and slope 0, starts slightly low (\(\hat{\alpha}_1 = 1.958\)) and has slightly positive (.029) slope. This seems to 
be due to the large number of points with large negative 
residuals near \(x_1 = 1.5\). Similarly, the line with slope 1 
and intercept 0, has intercept too high and slope too low.
While the model does seem to fit qualitatively, there are too many points far enough away from both lines which force us to reject the hypothesis that the model (5.2.2), is true.

**GWM**

The most striking feature of the scatter plot of the data from GWM (Figure 5.12) is that nearly all the points seem to fall on the line of slope 1 and intercept 1.00. The other points do not seem to cluster near 2.0 as in the previous subjects. The EM estimates were:

\[
\begin{array}{cccccc}
\lambda & \sigma_1 & \sigma_2 & \beta_1 & \beta_2 & \sigma_1 & \sigma_2 \\
\hline
0.095 & 0.861 & -0.00124 & 0.624 & 1.001 & 0.244 & 0.00411 \\
\end{array}
\]

As expected, over 90% of the points were estimated as coming from the slope 1, intercept 0 line which was estimated to have error standard deviation 0.00411. By contrast, the remaining points were estimated to have come from a line of intercept 0.861 and slope 0.624. Apparently, a few of the points near the slope 1 line were assigned to the other population because of the relatively small error deviation of the slope 1 line. The bootstrap 95% confidence intervals were:

\[
\begin{array}{cccccc}
\lambda & \sigma_1 & \sigma_2 & \beta_1 & \beta_2 & \sigma_1 & \sigma_2 \\
\hline
0.063 & 0.376 & -0.00460 & 0.323 & 0.999 & 0.169 & 0.00398 \\
0.127 & 1.346 & 0.00213 & 0.924 & 1.002 & 0.318 & 0.00482 \\
\end{array}
\]
Figure 5.12

Scatterplot of GWM data
The difference in the structure of the two lines is clear. The slope 1 intercept 0 line has very tight confidence limits due to the small standard error, while the other line ranges from a slope of 0.323 to nearly 1.0. What is clear is that most of the time (87.3 to 93.7%), GWM hears the octave as equivalent to the tuning ratio, (the line $x=y$), while even during the other 6.3 to 12.7% of the time he does not hear the octave, $x=2$, but rather a widely varying set of frequency ratios.

JWG

JWG (Figure 5.13) exhibited one point of octave doubling at $x=3.0, y=4.0$ and three points near $x=2.5, y=3.0$. These were removed, leaving 356 data points. From the scatterplot without the four points (Figure 5.14), one sees the two lines clearly. He seems to have been successful in hearing the octave (choosing $x=2$) most of the time except in the range 1.95 to 2.05 which seems to be a line of slope 1, intercept 0 and very small standard error. In fact, the EM estimates were:

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.623</td>
<td>2.001</td>
<td>0.00951</td>
<td>0.00549</td>
<td>1.000</td>
<td>0.0554</td>
<td>0.00353</td>
</tr>
</tbody>
</table>

146
Figure 5.13

Scatterplot of JWG full data
Figure 5.14

Scatterplot of JWG data, outliers removed
with bootstrap 95% confidence intervals:

\[ \begin{array}{cccccccc}
0.568 & 1.968 & -0.00327 & -0.00987 & 0.989 & 0.0493 & 0.00285 \\
0.679 & 2.034 & 0.0232 & 0.0198 & 1.002 & 0.0623 & 0.00433 \\
\end{array} \]

confirming our observations. The model clearly fits in terms of slope and intercept estimates.

**KIP**

The scatterplot of KIP (Figure 5.15), like that of JWG exhibits several clear outliers due to octave doubling. The scatterplot with outliers removed is shown in Figure 5.16.

Again, one sees the two lines clearly. The EM estimates were:

\[ \lambda \quad \sigma_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \sigma_1 \quad \sigma_2 \]

\[ \begin{array}{cccccccc}
0.622 & 2.026 & 0.0930 & 2.67E-4 & 0.961 & 0.0521 & 0.0291 \\
\end{array} \]

with 95% bootstrap confidence intervals:

\[ \begin{array}{cccccccc}
0.505 & 1.964 & -0.00704 & -0.0271 & 0.920 & 0.0399 & 0.0110 \\
0.738 & 2.088 & 0.193 & 0.0276 & 1.001 & 0.0642 & 0.0472 \\
\end{array} \]

Thus, the estimates for KIP and JWG are quite similar, the main difference being the standard error around the slope 1
line. This standard error is roughly 7 times larger, while

Figure 5.15
Scatterplot of KIP full data
Figure 5.16

Scatterplot of KIP data, outliers removed
the remaining 6 parameter estimates are comparable. The point with the largest residual was the point at (1.35, 1.51) circled in Figure 5.16. Although we had no compelling reason for removing it, we re-estimated the parameters without it to study its influence. These estimates were:

\[
\begin{array}{cccccccc}
\lambda & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 & \sigma_1 & \sigma_2 \\
.615 & 2.026 & .0381 & 8.59E-4 & .983 & .05334 & .0159
\end{array}
\]

The parameters which changed more than one estimated standard error were \( \alpha_2, \beta_2 \) and \( \sigma_2 \), with the largest change in \( \sigma_2 \). Thus the point had some influence in distorting the slope 1 line estimates and in inflating its estimated standard error.

§ 5.4 Conclusions

We have analyzed the data of five subjects of an experiment as a mixture of regressions. We estimated the seven parameters using the EM algorithm with an ad hoc binning estimate starting point. We used bootstrap standard errors to construct approximate confidence intervals for the parameters.

In all five subjects (with outliers removed), one can see the two lines of model (5.2.3) to a varying degree. We analyzed in great detail the data of JMS and concluded that
the parameters of the model (5.2.2) hypothesized by Cohen (1981) were within the 95% confidence limits and thus that the data could be seen as a mixture of two lines, one of slope 1 intercept 0, the other of slope 0 and intercept 2. The assumption of constant $\lambda$ over the range of $x$ was tested graphically and it was concluded that there was enough evidence to cast considerable doubt on this assumption. However, if the data are viewed as having constant $\lambda$, this overall $\lambda$ seems to be estimated accurately by the procedure. (This was true of all subjects, although the details were not given). A further step in the data analysis might be to model $\lambda$ as a function of $x$, or to perform separate analyses for different ranges of $x$.

For three of the subjects (JMS, JWG and KIP), all four of the parameters of the hypothesized model (5.2.2) were within the confidence limits. For GWM only the parameters of the line with slope 1 and intercept 0 were within the limits, while for SCH, all four parameters narrowly missed the limits. GWM stood out as the only one to have $\lambda$ (probability of belonging to the slope 2, intercept 0 line) less than .50. In fact, $\hat{\lambda}$ for GWM was only .095. In the case of SCH, two many points at relatively large distances from either postulated line caused the failure of the postulated parameters to fall within the confidence bounds.
In conclusion, we have shown that the binning procedure combined with the EM algorithm provides asymptotically efficient estimates of the seven parameters of the mixture of regression model. We have demonstrated their small sample characteristics on three simulated data sets and on five actual data sets. In all cases, the EM estimates were within the accuracy predicted by asymptotic efficiency, and seemed to possess good small sample properties of robustness and efficiency.
BIBLIOGRAPHY


