ESTIMATING MIXTURES OF NORMAL DISTRIBUTIONS AND SWITCHING REGRESSIONS

by

Richard E. Quandt

Econometric Research Program
Princeton University

Research Memorandum #190
November 1975

ESTIMATING MIXTURES OF NORMAL DISTRIBUTIONS AND SWITCHING REGRESSIONS*

by

Richard E. Quandt

1. Introduction

A problem which occurs in a wide variety of disciplines is that of separating the components of a probability density function which is the mixture of two or more normal densities. The problem occurs in engineering (Young and Coraluppi, (1970), Yakowitz (1970)), where it is referred to as "unsupervised learning," biology (Bhattacharya (1966)), physiology (Medgyessy (1953)), and economics (Quandt (1972), Ramsey (1975)). For the sake of simplicity we shall restrict the subsequent discussion to the case in which it is known a priori that the number of components is two. The simplest possible such case is when a sample of observations $x_1,\ldots,x_n$ is given on a random variable x where it is known that

$$x \sim N(\mu_1,\sigma_1^2) \quad \text{with probability} \quad \lambda$$

and (1-1)

$$x \sim N(\mu_2,\sigma_2^2) \quad \text{with probability} \quad 1-\lambda \; ,$$

the parameters λ, μ_1, μ_2, σ_1^2, σ_2^2 being unknown. A more complicated case arises in the switching regression context in which observations are given on a random variable y and on a vector of nonstochastic regressors x and where

$$\begin{aligned} y_i &= x_i'\beta_1 + u_{1i} \quad \text{with probability} \quad \lambda \\ y_i &= x_i'\beta_2 + u_{2i} \quad \text{with probability} \quad 1-\lambda \end{aligned} \qquad (1\text{-}2)$$

where $u_{1i} \sim N(0,\sigma_1^2)$, $u_{2i} \sim N(0,\sigma_2^2)$, with λ, the vectors β_1, β_2 and σ_1^2,

[*] I am indebted to James Trussell for useful suggestions. Support from NSF Grant #43747X is gratefully acknowledged.

σ_2^2 being unknown. In an economic context a switching regression systen such (1-2) is equivalent to assuming the presence of structural change. Frequently structural change may be posited to depend deterministically on some observable variables; formulating it specifically as in (1-2) implies that the investigator is ignorant concerning what moves the system from one structural form to another.

The history of attempting to solve the problem is a long one and for the earlier portion of it the reader is referred to Cohen (1967). In attacking the problem the following well-known propositions may be kept in mind: (1) Unlike mixtures of some other densities, the parameters of mixtures of normal densities are identified (Yakowitz (1970)). (2) There exists no sufficient estimator for the parameters of a normal mixture (Dynkin (1961)). (3) If a priori information is available which states that $\sigma_1^2 = k\sigma_2^2$ (in either (1-1) or (1-2)), with k a known number, then maximum likelihood estimates are consistent and may ordinarily be computed by numerical methods without too much difficulty. This is the case treated in detail by Day (1969). (4) If k is not known in $\sigma_1^2 = k\sigma_2^2$, then the likelihood function corresponding to either (1-1) or (1-2) is unbounded and the attempt to determine the location of a global maximum leads to inconsistent estimates.

The purpose of the present paper is to investigate an estimating method for the case in which no prior information is available concerning the σ's. The method makes use of the sample moment generating function and can be applied in reasonably straightforward, "mechanical" manner. It is useable with relatively small samples in contrast to some of the methods relying on graphical techniques (Bhattacharya (1967)) where illustrations with sample sizes of several thousand are given. Finally, it can be applied to the regression case with no more difficulty than to the case of a simple, constant-parameter mixture.

In Section 2 we state the principal notions of (1) the method of moments (Cohen (1967)) which appears to be one of the most easily implemented techniques for estimating constant-parameter mixtures and (2) of the method employing the sample moment generating function. In Section 3 we report the results of some comparative sampling experiments and in Section 4 we extend the results to the regression case. This section also contains a specific economic application. Section 5 contains some brief conclusions.

2. Some Methods for Estimating Normal Mixtures with Constant Parameters

As indicated before, we shall restrict our attention to the case of a mixture of two normal distributions with unequal variances. In the absence of prior information about the variance ratio maximum likelihood estimation is not feasible; however, there are at least two easily computed nongraphical methods that can be employed in this case. These are the method of moments and the method of the sample moment generating function.

Method of Moments. The method of moments has been discussed by Day (1969) and Cohen (1967). In the case of two components the density function the parameters of which are to be estimated is

$$f(x;\lambda,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) = \frac{\lambda}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + \frac{1-\lambda}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2} \qquad (2\text{-}1)$$

Equating the sample mean to the theoretical first moment of (2-1) and the second, third, fourth and fifth sample moments about the mean to the corresponding theoretical central moments we obtain five equations from which it is possible to solve for λ, μ_1, μ_2, σ_1^2, σ_2^2. The solution requires the negative root of the nonic equation

$$a_9z^9 + a_8z^8 + a_7z^7 + a_6z^6 + a_5z^5 + a_4z^4 + a_3z^3 + a_2z^2 + a_1z + a_0 = 0 \quad (2\text{-}2)$$

where $a_9 = 24$, $a_8 = 0$, $a_7 = 84k_4$, $a_6 = 36m_3^2$, $a_5 = 90k_4^2 + 72k_5m_3$, $a_4 = 444k_4m_3^2 - 18k_5^2$, $a_3 = 288m_3^4 - 108m_3k_4k_5 + 27k_4^3$, $a_2 = -(63k_4^2 + 72m_3k_5)m_3^2$, $a_1 = -96m_3^4k_4$, $a_0 = -24m_3^6$ and where m_i denotes the ith central sample moments and k_j is the jth sample cumulant, i.e., $k_4 = m_4 - 3m_2^2$, $k_5 = m_5 - 10m_2m_3$. It is further shown in Cohen (1967) that if we define the differences

$$d_1 = \mu_1 - E(x)$$

$$d_2 = \mu_2 - E(x) \; ,$$

$\hat{z}$ the negative root that solves (2-2)[1] and r as

$$r = \frac{-8m_3\hat{z}^3 + 3k_5\hat{z}^2 + 6m_3k_4\hat{z} + 2m_3^3}{\hat{z}(2\hat{z}^3 + 3k_4\hat{z} + 4m_3^2)}$$

then we obtain as estimates of d_1 and d_2

$$\hat{d}_1 = (r - \sqrt{r^2 - 4\hat{z}})/2$$

$$\hat{d}_2 = (r + \sqrt{r^2 - 4\hat{z}})/2$$

We then have

$$\hat{\mu}_1 = \hat{d}_1 + \bar{x}$$

$$\hat{\mu}_2 = \hat{d}_2 + \bar{x}$$

where $\bar{x}$ is the sample mean. The variances are shown to be estimated by

$$\hat{\sigma}_1^2 = \frac{\hat{d}_1(2r - m_3/\hat{z})}{3} + m_2 - \hat{d}_1^2$$

$$\hat{\sigma}_2^2 = \frac{\hat{d}_2(2r - m_3/\hat{z})}{3} + m_2 - \hat{d}_2^2$$

and the mixture coefficient by

$$\hat{\lambda} = \frac{\hat{d}_2}{\hat{d}_2 - \hat{d}_1}$$

The various formulas can be simplified if prior knowledge exists to the effect that either the means or the variances of the components are identical. No estimates of the sampling variances of the estimates are provided by the technique.. We shall refer to the moment method as MM in the subsequent sections.

[1] See below for possible difficulties in this regard.

<u>Method Using the Moment Generating Function</u>. From (2-1) the moment generating function is

$$E(e^{\theta x}) = \lambda e^{\mu_1\theta+\sigma_1^2\theta^2/2} + (1-\lambda)e^{\mu_2\theta+\sigma_2^2\theta^2/2} \tag{2-3}$$

The sample characteristic function has been used successfully by Arad Wiener (1975) to estimate the parameters of symmetric stable Paretian distributions with characteristic exponent other than 2. In that case the estimation is relatively simple since the logarithm of the characteristic function is linear in the parameters.

In the present case the method involves a nonlinear optimization problem. Let a sample $x_1,\ldots,x_n$ be given and choose k values of θ, $\theta_1,\ldots,\theta_k$, in some small interval (a,b), $a < 0 < b$. For any given value of θ_j the quantity $E(e^{\theta_j x})$ may be estimated by $\sum_{i=1}^n e^{\theta_j x_i}/n$; moreover since the quantities $e^{\theta_j x_i}$ are distributed identically with common mean, $\sum_{i=1}^n e^{\theta_j x_i}/n$ converges to $E(e^{\theta_j x})$ with probability one by the Strong Law of Large Numbers. Since $\sum_{j=1}^n e^{\theta_j x_i}/n$ is, except for sampling error, equal to $\lambda e^{\mu_1\theta_j+\sigma_1^2\theta_j^2/2}$ + $(1-\lambda)e^{\mu_2\sigma_j+\sigma_2^2\theta_j^2/2}$, we shall estimate the parameters by minimizing

$$S_n = \sum_{j=1}^k \left(\frac{\sum_{i=1}^n e^{\theta_j x_i}}{n} - \lambda e^{\mu_1\theta_j+\sigma_1^2\theta_j^2/2} - (1-\lambda)e^{\mu_2\theta_j+\sigma_2^2\theta_j^2/2} \right)^2 \tag{2-4}$$

This method of estimating will be referred to as the MGF method. Let $(\hat\lambda_n,\hat\mu_{1n},\hat\sigma_{1n},\hat\mu_{2n},\hat\sigma_{2n}^2)$ be the parameter values that minimize S_n. Then, by the convergence of $\sum_{i=1}^n e^{\theta_j x_i}/n$ to $E(e^{\theta_j x})$ and by the uniqueness of the moments it follows that $\text{plim}(\min S_n) = 0$ and $\text{plim}(\hat\lambda_n,\hat\mu_{1n},\hat\sigma_{1n},\hat\mu_{2n},\hat\sigma_{2n}^2) = (\lambda,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2)$ and the estimates are consistent.

It is possible that (2-4) may be difficult to minimize since it involves sums of exponentials which occasionally create problems of slow convergence. In such a case it may be worth while to replace the second and third terms in (2-4) by their Taylor series approximations of desired degree. This will clearly introduce some truncation error and the consistency argument above no longer holds; nevertheless such an approximate technique may be worth while under some circumstances. In the sampling experiments described in Section 3 we report some experimentation with 3rd, 4th and 5th degree expansions; minimizing the corresponding expressions (2-4) will be denoted as the MGF3, MGF4 and MGF5 methods. Letting $\lambda_1 = \lambda$, $\lambda_2 = 1-\lambda$, the 5th degree expansion which multiplies $\lambda_\ell (\ell=1,2)$ is

$$1 + \theta_j\mu_\ell + \theta_j^2\left(\frac{\sigma_\ell^2+\mu_\ell^2}{2}\right) + \theta_j^3\left(\frac{\mu_\ell\sigma_\ell^2}{2} + \frac{\mu_\ell^3}{6}\right) + \theta_j^4\left(\frac{\sigma_\ell^4}{8} + \frac{\mu_\ell^2\sigma_\ell^2}{4} + \frac{\mu_\ell^4}{24}\right) +$$
$$\theta_j^5\left(\frac{\mu_\ell\sigma_\ell^4}{8} + \frac{\mu_\ell^3\sigma_\ell^2}{12} + \frac{\mu_\ell^5}{120}\right) + \theta_j^6\left(\frac{\sigma_\ell^6}{48} + \frac{\mu_\ell^2\sigma_\ell^4}{16} + \frac{\mu_\ell^4\sigma_\ell^2}{48}\right) + \theta_j^7\left(\frac{\mu_\ell\sigma_\ell^6}{48} + \frac{\mu_\ell^3\sigma_\ell^4}{48}\right) +$$
$$\theta_j^8\left(\frac{\sigma_\ell^8}{384} + \frac{\mu_\ell^2\sigma_\ell^6}{96}\right) + \theta_j^9\left(\frac{\mu_\ell\sigma_\ell^8}{384}\right) + \theta_j^{10}\left(\frac{\sigma_\ell^{10}}{3840}\right) \qquad (2\text{-}5)$$

3. Sampling Experiments with Normal Mixtures with Constant Parameters

General Discussion of Experiments. Sampling experiments were carried out employing (1) the method of moments, (2) the method employing the sample moment generating function, i.e., consisting of the minimization of (2-4), and (3) approximations to (2) employing third, fourth and fifth degree Taylor series expansions of the right hand side of (2-3), i.e., formulas such as (2-5). These latter approximations were computed in only a subset of all the experiments and are not considered the principal procedures.

Samples of size N of the random variable x were generated according to the pdf (2-1); the true parameters of the pdf in the various experiments and the value of N are given in Table 1. Each experiment was replicated as many times as was necessary to produce 50 successful matched replications of all five estimating methods in Cases 1 through 5 and of the two principal estimating methods in the remaining cases. A replication would be counted as a failure for any of the following reasons: (1) The numerical minimization of (2-4) for the MGF methods may have failed, either because of continual straying of the algorithm into prohibited regions in the parameter space (e.g., $\lambda < 0$ or $\lambda > 1$ or $\sigma_1^2 \leqq 0$ or $\sigma_2^2 \leqq 0$) or because of lack of positive definiteness of the matrix of second partials of (2-4) at the point deemed to be stationary by the minimization algorithm[2]; (2) the computation of the estimates by the method of moments may have failed because (a) (2-2) had more than one negative root and no admissible solution; i.e. one satisfying $0 \leqq \lambda \leqq 1$, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$; (b) it may have had more than one negative root with more than one admissible solution; (c) it may have had no negative root at all.[3] The failure rates of the various methods are exhibited in Table 2. No results are presented for MGF3, MGF4 and MGF5 in Cases 6 through 9 since these were not computed. The non-positive definiteness of the matrix of second partial derivatives for MGF in these cases was disregarded since it occurred so frequently that it would have required a very large number of replications to produce 50 successful ones.[4]

A crucial question for the **MGF-type methods is the choice** of the values of θ_j in (2-4) and in the approximations to it. In the basic set of nine experiments 50 values of θ_j were employed ($k = 50$ in (2-4) with $\theta_j = 2\pi j/25m$, where $j = -25, -24, \ldots, -1, 1, \ldots, 24, 25$ and $m = 10$. Some additional experiments reported at the end of this section consider variations in both

[2] Minimization was performed with the Davidon-Fletcher-Powell algorithm and computations were terminated if either the proportionate value of the step size or the proportionate change in the function value or the length of the gradient fell below .0001.

[3] In the Appendix we exhibit two samples corresponding to (b) and (c) respectively and the corresponding solutions derived from (2-2).

[4] The value of S_n at the minimum was characteristically $\leqq 10^{-6}$; the failure to exhibit positive definiteness was due the difficulty of determining sufficiently accurately the eigenvalues of the matrix of numerically computed second partials.

k and m. The particular choices for the basic experiments represent a compromise between conflicting considerations. Large values of k make for finer resolution but increase the cost of computations. Small values of m exaggerate the effect of atypical sample values but large values of m reduce the width of the interval about zero over which the sample moment generating function is evaluated. As the width of this interval goes to zero, S_n goes to zero and, more importantly, the matrix of second partials of S_n approaches singularity, thus making positive definiteness at the minimum harder to ascertain. The compromise selected was clearly not fully successful in steering an optimal middle course between these dangers.

Basic Results. Table 3 contains the mean estimates and Table 4 their mean square errors. Table 5 displays the fraction of times, for each Case (and aggregated over all coefficients) that a particular method produces the least absolute deviation of the estimate from its true value. Some of the relevant comparisons are as follows. Cases 2 and 1 differ only in sample size and represent a situation in which the component normal densities barely overlap; Cases 4 and 5 also differ only in sample size but the component densities have a substantial overlap. In both sets of cases the mixing parameter is .5. Case 3 is the same as Case 1 but has strongly unequal mixing. Case 6 is similar to Case 1 except one component has a large variance making for more overlap in the densities; Cases 7 and 8 have even larger variance for one component and highly unequal mixing. Case 9 is one in which the two component means are identical.

Comparing the MSE's across Cases 1 through 5 for MGF3, MGF4, MGF5 and MGF we note that MGF3 generally produces inferior results. There are 25 possible comparisons (5 cases x 5 coefficients); MGF produces the lowest MSE's of all four methods in 10 instances. Not considering MGF5, MGF has the lowest MSE's in 19 instances whereas MGF5 beats MGF3 and MGF4 in 17 instances. On the whole therefore MGF5 is quite comparable to MGF. The corresponding comparisons in terms of mean bias are much less sharp although MGF and MGF5 still beat the other two methods on the average. In terms of the fraction of smallest absolute deviations produced by the methods over the replications (Table 5) the conclusion is slightly different: MGF is notable better than MGF3, MGF4 and MGF5, but the latter does not clearly dominate MGF3 and MGF4. Since the results of Table 5 (unlike those in Table 4) are not affected by outliers, this suggests that the relatively poor performance of MGF3 and MGF4 is due to the presence of a few very bad estimates. Henceforth we shall concentrate on examining the behavior of MGF and MM.

Consider first Cases 1 and 2 in which the overlap of the components is small. The MSE's diminish with larger sample size for both MGF and MM, by roughly a factor of 2 for all but one coefficient for MGF but for only two coefficients in the case of MM. The MSE's are smaller for MGF estimates in seven out of the ten possible comparisons. In terms of the fractions reported in Table 5 MM appears to be slightly superior to MGF; however, if we disregard MGF3, MGF4 and MGF5 and compute just the fraction of times that MGF beats MM in terms of mean absolute deviation, the winning fractions for MGF for the five coefficients are .48, .62, .62, .78 and .96 yielding an average of .69.[5] Comparing Cases 4 and 5, the MSE's decline (again roughly by a factor of 2) for all coefficients for MGF; in the case of MM they actually increase in two instances and decline only slightly in the other three. In absolute terms MGF wins the comparison of MSE's in every one of the ten possible comparisons. Table 5 confirms this result. The slight asymmetry introduced in Case 3

[5] Pooling results for different coefficients induces lack of independence among subsets of trials for which methods a re being compared; hence the outcomes are not binomially distributed. For any individual coefficient the reader may test the null hypothesis that the true fraction is .5.

in which $\lambda = .75$ has no major effect on either MSE's or mean biases or on the fraction of best results achieved in Table 5: all of the relevant statistics change in rather minor ways. Case 6 has a substantial amount of overlap, with one of the component densities having a very much larger variance than the other. MGF is again superior to MM by all criteria. The mean square errors for all coefficient estimates except λ increase markedly, however, and tend to increase more for the component with the larger variance. The true variance of one of the components is increased even further in Cases 7 and 8; moreover the true mixing parameter is alternately set at the extreme values of .1 and .9; thus in Case 7 the component with small variance is selected on the average 10 percent of the time and in Case 8 90 percent of the time. It is uniformly true for both methods that the MSE's of estimates of the parameters of the components associated with large values of λ are small and conversely; in general, however, all MSE's as well as mean biases (except for λ) are large. It is still true in Case 7 that MGF is overwhelmingly superior to MM (with MGF having an aggregate winning percentage for absolute mean deviations over the replications of 80 percent); however, this is no longer true in Case 8 where MGF beats MM in terms of MSE's only for the parameters of the seldom selected component. In Case 9, in which the true means of the components are identical, MGF again does substantially better than MM on all counts (although for one coefficient the MSE is slightly smaller for MM than MGF).

Variations in k and m. Some experiments were performed in order to examine the sensitivity of the MGF results to variations in k and m. Case 1 was repeated four times with the combinations ($k = 24$, $m = 10$), ($k = 74$, $m = 10$), ($k = 50$, $m = 5$), ($k = 50$, $m = 20$). The general conclusion that emerges is that the results have a moderate sensitivity to variations in k but a somewhat more significant sensitivity to variations in m. Within the ranges examined the reduction in k from 50 to 24 reduced MSE's by anywhere from 1 to about 70 percent the increase in k to 74 increased all but one MSE. Changes in m had a fairly marked effect: m = 5 made MSE's substantially larger for one coefficient (by a factor of over 10) and about 50 percent larger for three others; m = 20 reduced all MSE's, some by more than a factor of 10. Essentially similar behavior was noted when base 8 was repeated with k = 24 and m = 10 (slightly larger MSE's for two coefficients, noticeably smaller ones for the remaining three) and with k = 50, m = 20, where all MSE's declined markedly.

Concluding Remarks. On the whole the MGF method proved superior to MM. In most cases it had smaller MSE's and tended to beat MM in pairwise comparisons over the replications of the absolute deviation of the estimates from the true values. Increasing the sample size had the expected effect on the MSE of the MGF estimates but did not have a comparable effect on those of MM. Computational failures were also much more numerous for MM than for MGF with the former failing 223 times throughout the basic experiments for one reason or another while the latter failed on 33 times.[6]

There are two cautionary remarks in order. (1) The computation of a positive definite matrix of second partials at the optimum created serious difficulties in a number of cases; these failures were disregarded in the above generalization about failure rates. Even more serious is the fact that no estimates of sampling variability are produced by the MGF technique, and hence, on these grounds there is nothing to choose between the two methods. The reasons for this are as follows: (a) The least squares estimates obtained by minimizing (2-4) are not maximum likelihood estimates, even if we assume that the dependent variable $\sum_{i=1}^{n} e^{\theta_j x_i}/n$ is approximately normal (by the Central Limit Theorem) since the variance of this variable (conditional on x) is

[6] The sequence of computations was that MGF was first computed and MM second. Thus, if MGF failed, the corresponding MM computation was never performed. If MGF succeeded and MM then failed, the successful MGF computation was discarded. Thus the overall MGF failure rate is 33 out of 706 tries (=.047) and the overall MM failure rate is 223 out of 673 (=.331).

clearly a function of θ and hence we have a case of heteroscedasticity for which no provision was made in estimation;[7] (b) Even if an approximation to the appropriate variance-covariance structure were explicitly employed, the negative inverse of the matrix of second partials of the corresponding log-likelihood function (which is a scalar multiple of the inverse of second partials of the weighted sum of least squares) evaluated at the estimates cannot be guaranteed to provide even a consistent estimator of the Cramer-Rao bound since no sufficient statistic exist. (2) The MGF estimates are considerably affected by variations in m and k and we have at present only superficial knowledge of the range of effects which variations in these parameters may produce.

4. Estimating Mixtures of Regressions

Basic Notions. Consider the case when

$$y_i = a_1 + b_1 x_i + u_{1i} \quad \text{with probability } \lambda$$

and

$$y_i = a_2 + b_2 x_i + u_{2i} \quad \text{with probability } 1-\lambda$$

where u_{1i} $(i=1,\ldots,n)$ is i.i.d. as $N(0,\sigma_1^2)$, u_{2i} is i.i.d. as $N(0,\sigma_2^2)$, and where the x_i are nonstochastic and assumed identical in repeated samples. The pdf of the random variable y is

$$h(y) = \frac{\lambda}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(y-a_1-b_1x)^2}{2\sigma_1^2}} + \frac{1-\lambda}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(y-a_2-b_2x)^2}{2\sigma_2^2}} \qquad (4\text{-}1)$$

[7] The variance of $\sum_{i=1}^{n} e^{\theta x_i}/n$ is approximately $\sigma^2\theta^2 \sum_{i=1}^{n} e^{2\theta x_i}/n^2$ where $\sigma^2 = \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2$. A weighted least squares approach would consist of dividing each term of (2-4) by this approximation. This is almost certain to create difficulties in minimization since the denominator will contain weighted sums of σ_1^2 and σ_2^2: it is characteristic of such problems that minimization algorithms attempt to set one or both σ_k's at negative values. As a compromise the σ^2 term might be omitted altogether.

and one could think of developing estimators based on the method of moments as in Section 2; however, the number of parameters is increased here by two (and in the case of k independent variables in the two regressions it would be increased by 2k) and this necessitates employing moments of order even higher than fifth. The results are likely therefore to be fairly unstable.

The method employing the moment generating function can be applied in the present case with one additional approximation. The moment generating function is a function of the regressor x since

$$E(e^{\theta y}) = \lambda e^{(a_1+b_1x)\theta+\theta^2\sigma_1^2/2} + (1-\lambda)e^{(a_2+b_2x)\theta+\theta^2\sigma_2^2/2} \qquad (4\text{-}2)$$

We again select a set of θ_j's and replace $E(e^{\theta_j y})$ by $\frac{1}{n}\sum_{i=1}^{n} e^{\theta_j y_i}$; at the same time we replace $e^{\theta_j b_k x}$ (k=1,2) by $\frac{1}{n}\sum_{i=1}^{n} e^{\theta_j b_k x_i}$.[8] We estimate the parameters by forming the obvious analogue to (2-4) and minimize it with respect to all unknown parameters.[9]

Sampling Experiments. Some very modest sampling experiments were performed to verify the workability of the procedure. The independent variable x was chosen once and for all from the uniform distribution from 0 to 10. Twenty-four values of j were used and m was set to 30 throughout the experiments.

The additional basic characteristics of the experiments are given in Table 6, the mean estimates in Table 7 and the mean square errors in Table 8. In all cases the regression equations are so chosen that the scatters of points generated from the two regressions overlap one another substantially. Cases 1 and 2 differ in that in Case 1 the two pairs of true regression coefficients are more distinctly different from one another than in Case 2. Cases 1 and 3 on the one hand and Cases 2 and 4 on the other differ only in

[8] We assume that the $\sum_{i=1}^{n} e^{\theta_j b_k x_i}/n$ converges to a finite limit as $n \to \infty$.

[9] The resulting optimization problem will in most instances be of a dimensionality that is routinely soluble.

the value of λ; Case 5 is the same as Case 4 except for sample size. Only 30 successful replications were generated for each experiment; the number of failures in computing the minimum was 0 in Case 1, 4 in each of Cases 2 and 3, 13 in Case 4 and 18 in Case 5. The mean biases are quite small although both b_1 and b_2 are slightly biased toward zero in nearly all cases. The MSE's behave at least partly as expected. Specifically, (a) they are uniformly larger in Case 2 than in Case 1; (b) they are uniformly larger in Case 5 than in Case 4; (c) the MSE's of the estimates for the coefficients in the first regression are uniformly smaller in Case 4 than in Case 2; (d) the MSE's of the estimates for the coefficients of the second regression are uniformly larger in Case 3 than in Case 1; (e) The values of the MSE's for the b's are of the general order of magnitude that we would expect for their sampling variance if we had (i) prefect sample separation and (ii) indefinitely large samples with the x's drawn from the uniform distribution (the true sampling variance for b_1 and b_2 under those circumstances being .0072 and .0048 respectively). However the MSE's of the coefficients in the first regression are not all smaller in Case 3 than in Case 1 and the MSE's of the coefficients of the second regression are actually all smaller in Case 4 than in Case 2. These anomalies may be due to the omission of failed replications from the summary statistics and to the relatively small number of replications.

An Economic Example. Hamermesh (1970) examined the determination of wage bargains from a pooled cross-section time-series sample of 180 observations on wage changes $(\dot{w})$, changes in the consumer price index $(\dot{c})$ and unemployment (u). The null hypothesis is that

$$\dot{w} = \beta_0 + \beta_1\dot{c} + \beta_2(1/u) + \varepsilon$$

with β_1 and β_2 both greater than zero and where ε is the error term. The alternative hypothesis entertained by Hamermesh is that the effect of $\dot{c}$ on $\dot{w}$ is of a noticeably positive magnitude only when $\dot{c}$ exceeds some critical figure; this critical figure is selected a priori by Hamermesh to be 2 percent per annum. Accordingly, the alternative is

$$\dot{w} = \beta_{10} + \beta_{11}\dot{c} + \beta_2(1/u) + \varepsilon_1 \quad \text{if} \quad \dot{c} < 2$$

and (4-3)

$$\dot{w} = \beta_{20} + \beta_{21}\dot{c} + \beta_2(1/u) + \varepsilon_2 \quad \text{otherwise}$$

The inclusion of $1/u$ in the regression is the obvious Phillips curve notion that increasing levels of unemployment make it increasingly difficult for unions to drive hard wage bargains; the inclusion of $\dot{c}$ reflects the assumption that the wage bargain will (at least for some large values of $\dot{c}$) reflect increases in the cost of living.[10] Hamermesh estimates his switching model (i.e., Equations (4-3) by introducing suitable dummy variables which alter their values from one level to another at the point at which $c = 2.0$ and restricts β_2 to be identical in the two equations. In terms of the formulation given here he finds

$$\hat{\beta}_{10} = 2.3576 \qquad \hat{\beta}_{20} = 2.4525$$
$$\hat{\beta}_{11} = -0.0424 \qquad \hat{\beta}_{21} = 0.5345$$
$$\hat{\beta}_2 = 7.3539$$

that is to say, for low levels of $\dot{c}$ this variable has a negligible (and negative) effect on wage changes but a more sizeable and positive one for larger values of $\dot{c}$.

We shall assume here that we have no prior information as to the critical value of $\dot{c}$ below and above which different regression regimes are at work. We may continue to assume that nature chooses between the two regressions for any particular data point by comparing $\dot{c}$ to a critical value (known only to nature); thus, if this critical value is $\bar{c}$, and the fraction of observations with $\dot{c} < \bar{c}$ is λ, then nature chooses the first equation of (4-3) with probability λ and the second with probability $1-\lambda$. If we then assume that

[10] For more detail on the theory as well as the data the reader is referred to Hamermesh (1970).

the ε's are normally distributed we are dealing with the case discussed in Section 3, differing from that only in that here we have two regressor variables.

Only point estimates are obtained by the method of the sample moment generating function; we computed the estimates both on the assumption that β_2 is identical in the two regimes and that it may assume different values. In the restricted case we have

$$\hat{\lambda} = .4904$$

$$\hat{\beta}_{10} = 2.1411 \qquad \hat{\beta}_{20} = 2.3947$$

$$\hat{\beta}_{11} = -.1760 \qquad \hat{\beta}_{21} = 1.0797$$

$$\hat{\beta}_2 = 7.3062$$

In the unrestricted case we have

$$\hat{\lambda} = .4905$$

$$\hat{\beta}_{10} = 2.1398 \qquad \hat{\beta}_{20} = 2.3929$$

$$\hat{\beta}_{11} = -.1781 \qquad \hat{\beta}_{21} = 1.0783$$

$$\hat{\beta}_{12} = 7.3097 \qquad \hat{\beta}_{22} = 7.3498$$

Not only are the coefficients nearly identical between the two cases, but the value of the sum of squares at the minimum is 9.095×10^{-8} and 8.955×10^{-8} respectively; differing by a negligible magnitude. What is interesting about the estimates is particularly twofold: (1) the coefficient of $\dot{c}$ in the second equation is much larger than as estimated by Hamermesh; in fact it is slightly greater than unity, suggesting "overcompensation" in wage bargains of changes in the cost of living; (2) if we order the sample values of $\dot{c}$ and compute the value of the cutoff $\bar{c}$ implied by $\lambda = .49$, we obtain by interpolating between the 88th and 89th largest observations a critical value of 1.35, considerably smaller than Hamermesh's a priori value of 2.0. The conclusion thus is that changes in the cost of living begin to matter at an even lower value of $\dot{c}$ than employed by Hamermesh and when it matters, wage bargains overcompensate for it.

5. Conclusion

We have adapted the technique of using the sample moment generating function to estimating mixtures of normal distributions and contrasted it with the well-known method of moments. The method employing the sample moment generating function appears to work well in providing point estimates in the case of mixtures of components with constant means as well as regression mixtures in which the mean value of the random variable depends on the regressors. The method can be successfully applied to realistic data and has in one instance provided a novel interpretation of a sample. Along with some other methods it fails to give estimates of the sampling variance of the estimates; how best to compute such sampling variances is an open question for future research.

Table 1

SUMMARY CHARACTERISTICS OF EXPERIMENTS
WITH CONSTANT PARAMETER MIXTURES

	Case								
	1	2	3	4	5	6	7	8	9
λ	.5	.5	.75	.5	.5	.5	.1	.9	.5
μ_1	-3.0	-3.0	-3.0	-1.0	-1.0	-3.0	-3.0	-3.0	3.0
μ_2	3.0	3.0	3.0	1.0	1.0	3.0	3.0	3.0	3.0
σ_1^2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
σ_2^2	3.0	3.0	3.0	3.0	3.0	16.0	36.0	36.0	9.0
N	50	25	50	50	100	50	50	50	50

Table 2

CAUSES OF FAILURES IN COMPUTING ESTIMATES
FOR CONSTANT PARAMETER MIXTURES

Cause	1	2	3	4	5	6	7	8	9
Convergence									
MGF3	0	2	6	0	1	NA	NA	NA	NA
MGF4	1	0	2	1	1	NA	NA	NA	NA
MGF5	0	0	0	0	0	NA	NA	NA	NA
MGF	0	0	2	3	1	3	13	5	6
Positive Definiteness									
MGF3	13	19	0	0	0	NA	NA	NA	NA
MGF4	13	21	0	0	0	NA	NA	NA	NA
MGF5	6	7	0	0	0	NA	NA	NA	NA
MGF	7	6	0	0	0	NA	NA	NA	NA
Moment Method									
Negative Root(s) without admissible solution	0	0	0	4	4	2	5	4	23
Multiple Negative Roots with multiple admissible solutions	0	0	1	14	46	5	4	8	15
No Negative Root	0	1	0	22	23	15	22	0	5

NA: Not Applicable

Table 3

MEAN ESTIMATES FOR CONSTANT PARAMETER MIXTURES

	Case 1	2	3	4	5	6	7	8	9
MGF3									
λ	.4879	.4712	.8307	.4480	.4535				
μ_1	-2.7353	-2.7476	-2.4824	-.9367	-.9879				
μ_2	2.8056	2.8721	3.8977	.9370	.9709				
σ_1^2	2.6139	2.3924	2.6396	1.0615	1.0692				
σ_2^2	4.7260	4.7928	4.1833	2.9522	2.9831				
MGF4									
λ	.4823	.4718	.7444	.4491	.4579				
μ_1	-3.0182	-2.9485	-2.9459	-.9308	-.9817				
μ_2	2.9796	3.0513	2.9689	.9330	.9801				
σ_1^2	1.1081	1.0812	1.1152	1.0523	1.0473				
σ_2^2	3.0315	3.0980	3.0044	2.9486	2.9851				
MGF5									
λ	.4806	.4704	.7455	.4455	.4561				
μ_1	-3.0244	-2.9575	-2.9423	-.9311	-.9852				
μ_2	2.9646	3.0432	2.9810	.9252	.9737				
σ_1^2	1.0847	1.0708	1.0746	1.0625	1.0509				
σ_2^2	3.0107	3.0504	3.0028	2.9434	2.9812				
MGF									
λ	.4799	.4687	.7420	.4522	.4605	.4380	.1137	.8996	.5439
μ_1	-3.0414	-2.9937	-2.9623	-.9232	-.9886	-3.3406	1.9901	-2.9642	3.0051
μ_2	2.9725	3.0569	2.9591	.9262	.9950	3.0502	2.2822	6.2887	2.9422
σ_1^2	1.0156	.9185	1.0143	1.0694	1.0370	.2259	2.1112	1.0522	.9637
σ_2^2	2.9391	2.9092	2.9729	2.9416	2.9950	13.5460	29.0291	8.8056	9.3692
Method of Moments									
λ	.5057	.5084	.7524	.6969	.6887	.6549	.5002	.9041	.5273
μ_1	-2.9313	-2.8198	-2.9429	-.7385	-.6551	-2.2435	.9187	-2.9128	2.9378
μ_2	3.1671	3.3503	3.1843	2.2197	2.0366	5.2429	4.3744	4.8315	3.2037
σ_1^2	1.0482	1.0358	.9633	1.0450	1.1155	2.0202	10.5851	1.4906	3.5275
σ_2^2	2.4042	2.1839	2.1996	1.0236	1.4197	7.3462	19.1861	23.9106	11.4840

Table 4

MEAN SQUARE ERRORS FOR CONSTANT PARAMETER MIXTURES

	Case								
	1	2	3	4	5	6	7	8	9
MGF3									
λ	.0144	.0217	.0117	.0213	.0124				
μ_1	.1661	.2024	.3077	.0820	.0539				
μ_2	.3359	.3800	1.4638	.0987	.0712				
σ_1^2	3.1099	2.7980	3.1608	.0625	.0469				
σ_2^2	4.3826	5.5903	1.9412	.0592	.0422				
MGF4									
λ	.0069	.0117	.0045	.0207	.0118				
μ_1	.0592	.0535	.0376	.0804	.0513				
μ_2	.1779	.3043	.1894	.0928	.0580				
σ_1^2	.1090	.1660	.0915	.0494	.0340				
σ_2^2	.1551	.2594	.0547	.0524	.0328				
MGF5									
λ	.0068	.0120	.0043	.0218	.0120				
μ_1	.0485	.0534	.0343	.0830	.0514				
μ_2	.1567	.3010	.1723	.0989	.0534				
σ_1^2	.0549	.1156	.0558	.0565	.0322				
σ_2^2	.0905	.1741	.0370	.0540	.0308				
MGF									
λ_1	.0066	.0112	.0043	.0179	.0094	.0094	.0354	.0100	.0334
μ_1	.0505	.0572	.0393	.0900	.0466	.2682	57.6666	.1292	.2647
μ_2	.1599	.3116	.1825	.0912	.0480	.4108	5.2718	10.7377	.6203
σ_1^2	.0738	.1959	.0643	.0596	.0319	1.0649	11.7519	1.8116	.0894
σ_2^2	.1016	.1866	.0450	.0494	.0274	11.5414	110.8230	386.7620	2.0948
Method of Moments									
λ_1	.0046	.0087	.0036	.0533	.0510	.0309	.1890	.0053	.0665
μ_1	.0653	.1238	.0369	.5329	.8024	3.4343	42.2221	.0455	.2075
μ_2	.2558	.3856	.4110	2.0638	1.6615	6.8863	16.0544	34.0330	1.6916
σ_1^2	.1342	.1459	.0602	.1754	.1798	2.0763	177.1570	.2253	97.0687
σ_2^2	1.9253	1.9811	2.1516	4.1866	3.1056	83.9760	405.7640	834.5860	56.6363

Table 5

OVERALL PERCENTAGE WINS FOR CONSTANT PARAMETER MIXTURES

	Case								
	1	2	3	4	5	6	7	8	9
MGF3	.196	.196	.124	.260	.212	–	–	–	–
MGF4	.148	.128	.212	.056	.076	–	–	–	–
MGF5	.136	.120	.148	.200	.228	–	–	–	–
MGF	.244	.276	.248	.324	.344	.748	.800	.468	.728
Method of Moments	.276	.280	.268	.160	.140	.252	.200	.532	.272

Table 6

SUMMARY CHARACTERISTICS OF EXPERIMENTS WITH REGRESSION MIXTURES

			Case		
	1	2	3	4	5
λ	.5	.5	.8	.8	.8
a_1	-1.0	-1.0	-1.0	-1.0	-1.0
b_1	.5	.5	.5	.5	.5
a_2	4.0	-3.083333	4.0	-3.083333	-3.083333
b_2	-.7	1.0	-.7	1.0	1.0
σ_1^2	3.0	3.0	3.0	3.0	3.0
σ_2^2	2.0	2.0	2.0	2.0	2.0
N	50	50	50	50	25

Table 7

MEAN ESTIMATES FOR REGRESSION MIXTURES

			Case		
	1	2	3	4	5
λ	.5030	.5791	.8332	.7704	.7700
a_1	-1.0035	-.9754	-.9874	-1.0120	-1.0483
b_1	.4657	.4885	.4922	.4800	.4711
a_2	4.0027	-3.0627	4.0063	-3.0842	-3.1018
b_2	-.6627	.9222	-.6521	1.0207	.9997
σ_1^2	2.9979	2.9846	2.9901	3.0070	3.0470
σ_2^2	1.9978	1.9861	1.9986	2.0018	2.0084

Table 8

MEAN SQUARE ERRORS FOR
REGRESSION MIXTURES

	Case				
	1	2	3	4	5
λ	.0007	.0473	.0077	.0117	.0283
a_1	.0002	.0096	.0015	.0020	.0419
b_1	.0099	.0312	.0019	.0051	.0079
a_2	.0001	.0089	.0003	.0001	.0061
b_2	.0078	.0262	.0165	.0042	.0062
σ_1^2	.00002	.0024	.0009	.0006	.0457
σ_2^2	.00002	.0018	.00002	.00004	.0012

APPENDIX

Sample 1 was generated from density function (2-1) with $\lambda = .5$, $\mu_1 = -3.0$, $\mu_2 = 3.0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 16$. The sample values are listed in Table A-1. The roots of (2-2) are 4.35492, 13.6424, 37.7525, -8.24714 ± 6.09701i, -27.6368 ± 6.2964i, 8.00902 ± 31.96500i and no negative root exists.[11]

Sample 2 was generated with $\lambda = .5$, $\mu_1 = -1.0$, $\mu_2 = 1.0$, $\sigma_1^2 = 1.0$, $\sigma_2^2 = 3.0$ and the sample values are listed in Table A-2. The roots of (2-2) are -0.82929, -1.68893, 0.85453, 3.35794 ± 1.92864i, -2.70269 ± 0.45069i, 0.17660 ± 5.39189i. Two negative solutions exist and the corresponding estimates for the parameters are

$$\begin{array}{lll} \hat{\lambda} = 0.42698 & \qquad & \hat{\lambda} = 0.80570 \\ \hat{\mu}_1 = -1.71763 & & \hat{\mu}_1 = -1.30087 \\ \hat{\mu}_2 = 1.23416 & & \hat{\mu}_2 = 1.98374 \\ \sigma_1^2 = 0.06128 & & \sigma_1^2 = 3.16491 \\ \sigma_2^2 = 0.907107 & & \sigma_2^2 = 1.28267 \end{array}$$

In a case such as this there is no way in which the method can choose among the several admissible solutions; a difficulty that does not arise if (as happens often) all but one of the negative roots have inadmissible estimates associated with them.

[11] Roots were located by employing subroutine POLRT from the IBM Scientific Subroutine Package. On a variety of test problems this routine based on the Newton-Raphson algorithm, strongly dominated alternatives based on the Bairstow factorization and on the quotient-difference algorithm of Rutishuiser.

Table A-1

VALUES OF x IN SAMPLE 1

1.02	-2.52	4.76	4.92	-4.65	-2.26	0.75	2.62	6.23	-2.69
6.31	6.66	2.92	9.06	-1.80	-1.42	2.54	-3.90	3.65	4.68
2.26	8.69	1.02	-3.55	-2.89	0.15	5.29	-0.95	-2.43	-0.80
-2.49	-2.99	15.93	1.66	10.62	-1.29	-4.08	5.01	-4.94	5.80
-2.20	-3.03	6.38	-4.56	7.84	10.17	-3.14	-5.09	7.20	-3.58

Table A-2

VALUES OF x IN SAMPLE 2

2.00	-2.11	-1.31	-0.39	-0.54	-0.85	-0.68	-0.49	-1.59	3.08
-2.81	-2.81	1.53	0.19	1.05	-2.57	-1.80	0.24	3.71	3.52
-0.67	-2.28	0.10	-1.03	-2.65	2.30	-0.96	-0.91	-2.17	-0.91
-2.46	-1.87	-1.42	-0.91	-1.23	0.02	-0.84	0.32	-2.42	-0.78
1.78	0.02	-3.15	-1.40	-0.27	-0.44	-2.40	-2.37	-0.73	-2.95

REFERENCES

Arad Wiener, Ruth, *The Implications of a Long-Tailed Distribution Structure to Portfolio Selection and Capital Asset Pricing*, Ph.D. dissertation, Princeton University, 1975.

Bhattacharya, C.G., "A Simple Method of Resolution of a Distribution into Gaussian Components," *Biometrics*, 23(1967), 115-135.

Cohen, A. Clifford, "Estimation in Mixtures of Two Normal Distributions," *Technometrics*, 9(1967), 15-28.

Day, N.E., "Estimating the Components of a Misture of Normal Distributions," *Biometrika*, 56(1969), 463-474.

Dynkin, E.B., "Necessary and Sufficient Statistics for a Family of Probability Distributions," *Selected Translations in Mathematical Statistics and Probability*, Vol. I(1961), 17-40.

Hamermesh, Daniel S., "Wage Bargains, Threshold Effects, and the Phillips Curve," *Quarterly Journal of Economics*, 84(1970), 501-517.

Medgyessy, P., "Valószinüség-eloszlásfüggvények Keverékének Felbontása Összetevőire," *Hungarian Academy of Scoence Communications, Institute of Applied Mathematics*, 2(1953), 165-177.

Quandt, R.E., "A New Approach to Estimating Switching Regressions," *Journal of the American Statistical Association*, 67(1972), 306-310.

Ramsey, James B., "Mixtures of Distributions and Maximum Likelihood Estimation of Parameters Contained in Finitely Bounded Compact Spaces," Econometrics Workshop Paper No. 7501, Michigan State University, July 1975.

Yakowitz, Sidney J., "Unsupervised Learning and the Identification of Finite Mixtures," *IEEE Transactions on Information Theory*, (IT-16), (1970), 330-338.

Young, Tzay Y. and G. Coraluppi, "Stochastic Estimation of a Mixture of Normal Density Functions Using an Information Criterion," *IEEE Transactions on Information Theory*, (IT-16), (1970), 258-263.