DOCUMENT RESUME

ED 359 262 TM 020 031

AUTHOR Wang, Yuh-Yin Wu; Schafer, William D.
TITLE Maximum Likelihood and Minimum Distance Applied to Univariate Mixture Distributions.
PUB DATE Apr 93
NOTE 45p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computer Simulation; Equations (Mathematics); *Estimation (Mathematics); Graphs; *Mathematical Models; *Maximum Likelihood Statistics; Monte Carlo Methods; Statistical Distributions
IDENTIFIERS EM Algorithm; *Minimum Distance Principle; Mixtures; *Univariate Analysis

ABSTRACT

This Monte-Carlo study compared modified Newton (NW), expectation-maximization algorithm (EM), and minimum Cramer-von Mises distance (MD), used to estimate parameters of univariate mixtures of two components. Data sets were fixed at size 160 and manipulated by mean separation, variance ratio, component proportion, and non-normality. Results indicate that NW is the poorer estimation procedure. EM is less sensitive to different initial inputs and produced the lowest singularity rate. MD is more robust to non-normality and to incorrect model assumption of variance. In practice, MD is recommended. The singularity problem is not severe enough to be a practical concern. (Twelve figures present details of the simulations and analyses.) (Author/SLD)

# Maximum Likelihood and Minimum Distance Applied to Univariate Mixture Distributions



Yuh-Yin Wu Wang

William D. Schafer

Department of Measurement, Statistics, and Evaluation

University of Maryland at College Park



## Abstract

This Monte-Carlo study compared modified Newton (NW), expectation-maximization algorithm (EM), and minimum Cramer-von Mises distance (MD), used to estimate parameters of univariate mixtures of two components. Data sets were fixed at size 160 and manipulated by mean separation, variance ratio, component proportion, and non-normality. Our results indicate that NW is the poorest estimation procedure. EM is less sensitive to different initial inputs and produced the lowest singularity rate. MD is most robust to non-normality and to incorrect model assumption of variance. In practice, MD is recommended. The singularity problem is not severe enough to be a practical concern.

Key Words: Cluster Analysis, EM Algorithm, Mixtures, Modified Newton.

## 1. Introduction

Mixture modeling is one of the non-hierarchical clustering approaches. It is usually expressed as a superposition containing k components with a density function, $f_k$, and proportion, $\pi_k$, i.e.

$$\sum_{k=1}^{K}\pi_k = 1, \text{ where } 0 \le \pi_k \le 1 \text{ , and } f(x;\theta) = \sum_{k=1}^{K}[\pi_k * f_k(x;\theta^{(k)})]$$

where x is an independent observation and $\theta^{(k)}$ is a vector of parameters of subpopulation, k. In this article, we discuss literature findings and issues regarding the application of the model in section 2. The algorithms adopted and their background materials are provided in section 3. Methodology of the simulation study is described in section 4. Results and conclusions are presented in section 5. Finally section 6 presents discussion and suggests implications for future work.

## 2. Research Issues

Various methods have been developed and used to estimate the parameters of finite mixture distributions, such as the method of moments by Pearson (1894), maximum likelihood (ML) by Hasselblad (1966), fussy c-means partition by Davenport, Bezdek, and Hathaway (1988), moment generating function by Quandt and Ramsey (1978), a least square procedure by Fowlkes (1979), minimum distance method by Woodward, Parr, Schucany, and Lindsey (1984), quasi-Baysian approach by Hamilton (1991). The superiority of ML over the other methods for exploratory studies under normal distributions with sufficient sample sizes (say, 300) was confirmed by Day (1969), Tan and Chang (1972), Fryer and Robertson (1972), Kumar, Kicklin and Paulson (1979), Fowlkes (1979), and Woodward et. al. (1984). Within ML, the EM algorithm developed by Dempster, Laird, and Rubin (1977) is preferred (Everitt, 1984) and has been dominantly used as a representative method of ML in both practical and simulation studies.

Redner and Walker (1984) have given a regularity conditions that need to be satisfied in order to have identifiable likelihood estimation for a mixture distribution as follows.

Condition 1: For the parameter matrix $\theta$, the partial derivatives up to 3 orders, $\partial f / \partial \theta_i$, $\partial^2 f / \partial \theta_i \partial \theta_j$, and $\partial^3 f / \partial \theta_i \partial \theta_j \partial \theta_m$ exist and are bounded.

Condition 2: The Fisher information matrix $I(\theta)$ is positive definite at the true parameter values, $\theta^*$.

According to Redner and Walker (1984), if conditions 1 and 2 are satisfied and any sufficient small neighborhood of $\theta^*$ is given, then with probability 1, there is, for all sufficiently large sample size, N, a unique solution $\theta^N$ of the likelihood equations in that neighborhood, and this solution locally maximizes the log-likelihood function. Its distribution is,

$$\sqrt{N}(\theta^N - \theta^*) \sim N(0, I(\theta^*)^{-1}) \quad \text{as} \quad N \to \infty$$

## 2.1 The Problem of Singularity

Although the maximum likelihood (ML) approach to mixture distributions has been preferred since the advent of advanced computing equipment, the estimation procedure contains a problem of singularity which tends to discourage the use of ML. As one of the variances tends to zero and $\mu_k$ is set equal to any observed value of $x_n$, convergence to a singularity occurs. The likelihood function breaks down and becomes infinite at this point. To avoid the problem of singularity, the assumption of homogeneity of variance/covariance matrices between subpopulations can be assumed (Day, 1969). However, in empirical studies, this assumption seems to be very restrictive. Some researchers have found that with the heterogeneity assumption the problem of singularity does not occur under certain conditions, such as large sample size and good initial values (Fryer and Robertson , 1972; Hosmer, 1974), and well separated samples (Everitt, 1984).

On the other hand, Hathaway (1985, 1986) demonstrated that a constraint, such as $\min(\sigma_1 / \sigma_2) \geq c > 0$, imposed on standard deviations will eliminate all the singularities caused by small numbers of observations and poor separation. Hathaway, Huggins, and Bezdek (1984) , and Davenport, Bezdek, and Hathaway (1988) have reported improvement using the constrained EM algorithm by comparing it with unconstrained EM and other algorithms. Hamilton (1991) proposed a quasi-Bayesian approach which covers maximum likelihood estimation as a special case. The common characteristic of both methods is that they need priors.

Leytham (1984) pointed out that although there might be singularities in seeking a maximum likelihood estimate from a mathematical point of view, it has been found that in practice, they do not present serious difficulties. However, it is worth noting that the studies reporting frequencies of singularity (such as Leytham, 1984, and Hamilton, 1991) used true parameter values as initial starting points.

In order to find out whether researchers in empirical studies encountered the problem of singularity and how they dealt with it, a literature search using CD-ROM databases of Agricola, Eric, Life Science, Mathsci, Medline, Psychlic, and Sociofile was conducted. Thirteen empirical studies were found from 1983 to 1990. All of them adopted maximum likelihood approach. One of them did not specify which algorithm was used. Among the other twelve, nine adopted EM algorithm developed by Dempster, Laird and Robin (1977), two adopted Day's Newton method (1969), and one adopted quasi-Newton. It seems that singularity has not appeared as a practical problem, but the fact that only published literature is available could have suppressed the reporting of estimation difficulties. The percentage of singularity in various kinds of data distributions under different algorithms with inaccurate initial values is worth of investigating.

## 2.2 Assumption of Homoscedasticity

Because the assumption of homoscedasticity was first introduced to eliminate the problem of singularity (Day, 1969), it seems unnecessary for researchers to continue to require the assumption. Among the thirteen empirical studies mentioned in 1.4, seven adopted a heteroscedastic model, four adopted a homoscedastic model and the other two did not estimate variances. None of these studies mentioned the association of singularity with model assumption of variance. However, the principal of model parsimony was ignored when homoscedasticity could have represented a more parsimonious model.

Basford and McLachlan (1985) indicated that the adoption of a homoscedastic normal model in the presence of some heteroscedasticity can considerably influence the likelihood

estimates, in particular of the mixing proportions, thus resulting in a higher rate of misallocation. However, there is no study so far that has investigated the adoption of a heteroscedastic model in the presence of homogeneous data. The degree of bias when homoscedastic as well as heteroscedastic models are imposed on the same data set is also of concern.

### 2.3 Non-Normal Mixtures

ML estimation has been demonstrated to be efficient and consistent under normal mixture distributions if regularity conditions are satisfied (Redner and Walker, 1984). When non-normality is presented as difficult distributions, Woodward, Parr, Schucany and Lindsey (1984) found that minimum Cramer-von Mises distance provide better estimates than those of ML under heavy-tailed densities. However, they simulated mixtures of two components with identical shapes. For example, two double exponential components composed a mixture, and two student's t(4) components composed another form of mixture. The situation where a normal component is mixed with a non-normal component or different non-normal components are mixed was not considered. Skewed distributions also were not investigated but they have suggested such simulation would be useful. It is expected that mixtures of various types of distributions will provide more insight about performance of estimation procedures.

### 2.4 Purpose

The purpose of the present study is to explore the effects of variation of mixture distribution parameters and non-normality of two component density functions on the accuracy of parameter estimates with estimation procedures assuming either heteroscedasticity or homoscedasticity. Seven independent variables are manipulated: algorithms, model assumption of variance, initial input, mean separation, variance ratio, component proportion, and shape combination (normality/non-normality).

Three algorithms are evaluated: E(expectation) M(maximization) algorithm developed by Dempster, Laird, and Rubin (1977), a modified Newton approach (Dennis and Schnabel, 1983), and the minimum Cramer-von Mises distance algorithm (Parr and Schucany, 1980). Modified Newton is compared against EM to test how well the simple IMSL subroutine BCOAH performs. Cramer-von Mises distance is used as a contrast to EM and modified Newton to detect how robust these estimation procedures are to non-normality. Four research questions were investigated:

1. The percentage of samples that were failed due to a singularity problem were reported under various data distributions such as mean separation, variance ratio, initial input, and algorithms. The explicit form of a singularity problem is that estimates of component variance and proportion going to boundary. This is designed to evaluate whether the problem of singularity can be expected to occur in practice.

2. Asymptotic variances were calculated and correlated with the absolute distance of corresponding estimates from their true parameter values to investigate the degree to which asymptotic variances can be used to decide how reasonable an estimation event is.

3. The three algorithms mentioned above were compared for sensitivity to mean separation, variance ratio, and initial input. The effect of imposing heterogeneous model for homogeneous data was considered. Criteria for comparison included mean squared error (MSE) and bias of parameter estimates.

4. The robustness of maximum likelihood (ML) estimates to non-normal mixtures was investigated by comparing its performance with that of the minimum distance method. Various degrees of non-normality including positive skewness, negative skewness, leptokurtosis, and platykurtosis were combined to form various shape of a component. Dependent variables used were bias index and MSE.

## 3. Three Estimation Procedures Adopted in the Present Study

### 3.1 EM Algorithm

The procedure of EM algorithm is briefly described given a univariate distribution of two components as an example. Suppose $P(k \mid x_i)$ is the posterior probability that observation $x_n$ belongs to component k, then we have

$$P(k \mid x_n) = \pi_k * f_k(x_n ; \mu_k, \sigma_k,) / f(x_n ; \theta) \qquad (3.1)$$

The likelihood equation solution can be formulated in the context of EM algorithm (Dempster, Laird, and Rubin, 1977) as described in Everitt (1981, p. 37).

$$\pi = \frac{1}{n} \sum_{n=1}^{N} P(k \mid x_n), \quad k = 1, 2 \qquad (3.2)$$

$$\mu_k = \frac{1}{n * \pi_k} \sum_{n=1}^{N} P(k \mid x_n) x_n, \quad k = 1, 2 \qquad (3.3)$$

$$\sigma_k^2 = \frac{1}{n * \pi_k} \sum_{n=1}^{N} P(k \mid x_n) (x_n - \mu_k)^2, \quad k = 1, 2 \qquad (3.4)$$

The EM algorithm proceeds iteratively by two steps, E (expectation) and M (maximization) . In the E step, initial values of $\pi$, $\mu_k$, and $\sigma_k^2$ are used to obtain first estimate of $P(k \mid x_n)$. In the M step, given the posterior probabilities from E step, involves calculation of revised estimates of $\pi$, $\mu_k$, and $\sigma_k^2$ by inserting the posterior probabilities into the right hand side of (3.2) - (3.4). The intent is to maximize likelihood with tentative estimates from E step to give revised parameter estimates. The E step and M step are repeated alternately until some convergence criterion is satisfied. In the present study, an EM source code developed by McLanchlan and Basford (1988) was used as the basic algorithm with some changes to fit the simulation design.

### 3.2 Modified Newton

The modified Newton method is a line-search algorithm varying step size, $\lambda_r$, where $0 < \lambda_r < 1$ at each iteration r. The iteration procedure of the modified Newton is described as follows,

$$\theta^{(r+1)} = \theta^{(r)} - \alpha * \lambda_r * G_r^{'} * H^{-1}{}_{\theta(r)} * G_{\theta(r)}$$

where $\alpha \in (0,0.5)$. Note that the method presented here is a function minimization procedure; therefore, the log likelihood function of the mixture distribution is multiplied by -1 in order to locate a maximum . The strategy is to start with $\lambda_r = 1$ given $\alpha$ bounded by 0 and 0.5. The algorithm according to Dennis and Schnabel (1983, p.126) is given below,

Given $\alpha \in (0,0.5)$, $0 < l < u < 1$

$\lambda_r = 1$;

while $f(t_{r+1}) > f(t_r) + \alpha * \lambda_r * G_r^{'} * (-H_r^{-1}) * G_r$, do

$\lambda_r = z * \lambda_r$ for some $z \in [l,u]$ (z is chosen at each time by the line search);

$t_{r+1} = t_r + \lambda_r * z$; ($t_{r+1}$ is revised with the revised $\lambda_r$).

The search for $\lambda_r$ is named line search. Details about obtaining $\lambda_r$ can be found in Dennis and Schnabel (1983, chapter 6). The IMSL (International Mathematical and Statistical Libraries, 1989) subroutine BCODH was implemented in the present study using the algorithm.

### 3.3 Cramer-von Mises Distance

A minimum-distance measure is a method to estimate an unknown parameter vector $\theta$ by minimizing $\delta(G_n, F_\theta)$, where $G_n$ is the empirical distribution function based on

$x_1, x_2, \ldots x_n$, and $F_\theta$ is the mixture distribution function. The Cramer-von Mises distance is given by

$$\delta(G_n, F_\theta) = \int \{F(x; \theta) - G(x)\}^2 \, \partial F(x; \theta)$$

$$= (12N^2)^{-1} + [\sum_{n=1}^{N} \{F(x_{(n)}; \theta) - (n - \frac{1}{2}) / N\}^2] / N$$

where $x_{(n)}$ denotes the nth order statistic (n = 1,2,...N). The IMSL subroutine UNLSF that adopted Marquardt's (1963) method was used in the present study to minimize the function, as did Woodward et al. (1984). The IMSL special function, NORDF , was used to calculate integral function, $F(x_n)$.

## 4. Methodology

### 4.1 Parameter Values

Since we are using univariate mixtures of two components to investigate the research questions described section 2, there are 5 parameters to be varied, $\pi$, $\mu_1$, $\mu_2$, $\sigma_1$, and $\sigma_2$. Respectively, $\mu_1$ and $\sigma_1$ were fixed at 0 and 1.

The number of modes of a mixture distribution depends on the separation between two component densities. Behboodian (1970) derived a sufficient condition for a mixture to be unimodal,

$$|\mu_1 - \mu_2| \leq 2 \min(\sigma_1, \sigma_2).$$

Eisenberger (1964) has shown that a sufficient condition for a mixture to be bimodal is

$$(\mu_1 - \mu_2)^2 > (8 \sigma_1^2 \sigma_2^2) / \sigma_1^2 + \sigma_2^2$$

We studied the size of $\mu_2$ at (1) $\mu_2$=0.5, (2) $\mu_2$=2.6 to test the relative performance of each algorithm from unimodal and bimodal distributions. The size of $\sigma_2^2$, was studied at by (1) $\sigma_2^2$=1 with a variance ratio of 1 (reflecting homogeneous components), and (2) $\sigma_2^2$ =

4 with a variance ratio equal 4 (reflecting heterogeneous components). We considered three choices of $\pi$, (1) $\pi = 0.5$, (2) $\pi = 0.3$ and (3) $\pi = 0.7$ to represent even and uneven distributions.

### 4.2 Starting Values

Two methods were utilized to get starting values. One of them, cluster analysis, is implemented by IMSL subroutine KMEAN to obtain the estimates by the principle of minimizing the total within-cluster sums of squares. Initial values of $\sigma_k^2$ was set as cluster variances, and the one of component proportion was set 0.5. The other method, called the range method, was suggested by Davenport, Bezdek, and Hathaway (1988) when prior information about a mixture is not available. The range method assumes initial inputs of $\pi$ and $\mu$ to be evenly spread within the range of possible values. In another words, $u_1(0) = x_1 + (x_n - x_1) / 3$, $u_2(0) = x_1 + 2 (x_n - x_1) / 3$, and $p(0) = 0.5$, where $x_1$ is the smallest observation and $x_n$ is the largest observation in a data set. Since Davenport et al. did not report now to set initial values $s_k^2(0)$ for $\sigma_k^2$, we set it as $s_k^2(0) = s^2 - ((x_{median} - x_1) / 3)^2$ for both $\sigma_1^2$ and $\sigma_2^2$. Note that $u_1(0)$ and $u_2(0)$ from the range method were also used as initial inputs of cluster means in the cluster method.

### 4.3 Non-Normality

In the present study, the skewness index was set as 0.0, 0.6, and -0.6 to reflect normal, positively skewed, and negatively skewed distributions. The kurtosis index was set as -0.6, 0.0, and 0.6 to reflect platykurtosis, normality, and leptokurtosis. Exhausting all the combinations of the above two indices, we have 9 kinds of shape regarding a component distribution. Since two components were needed in a mixture, exhausting 9 conditions of shape with pairwise combinations, we have 81 (9*9) different mixtures to provide symmetric variation of the data distributions.

A non-normal distribution generating method developed by Ramberg, Dudewicz, Tadikamalla, and Mykytka (1979) was adopted in the study to provide our needed distributions. Variates (x) in a component were obtained by

$$x = R(z) = \lambda_1 + [z^{\lambda_3} - (1-z)^{\lambda_4}] / \lambda_2$$

where z is a uniform random variable on the interval zero to one, while $\lambda_1$ is a location parameter, $\lambda_2$ is a scale parameter, and $\lambda_3$ and $\lambda_4$ are shape parameters.

Distributions with skewness of -0.6 were obtained by multiplying all variates with skewness of 0.6 by -1 before being relocated and rescaled while mean and kurtosis index remained unchanged. A set of shape ID's from 1 to 9 were given to represent the shape conditions of components. The nine variations of component shape are depicted in Figure 1, and four examples of mixtures of two components are presented in Figure 2.

### 4.4 Simulated Samples

Altogether we have 972-- 2 (choices of $\mu_2$) * 2 (choices of $\sigma_2^2$) * 3 (choices of $\pi$) * 81 (shape combinations) -- different combinations of data distributions. These four variables are generally named here as data distribution variables. Sample size is fixed at 160, which is the median size from the thirteen empirical studies mentioned in section 2. Mean of the sample sizes of the thirteen empirical studies is 1811. The sample sizes range from 28 to 19679 with a single study conducting four mixtures of size larger than 17000 for each mixture.

Within each data distribution condition, there were 12 estimation procedures (3 estimation methods - EM, modified Newton (NW), and Cramer-von Mises distance (MD) * 2 initial inputs - cluster analysis and range method * 2 assumptions of variances - homogeneous model and heterogeneous model). These three variables are called here estimation procedure variables. In all, we have 972*12=11664 cells.

Due to poor performance from certain estimation procedures, the stopping rule for replications in each of the 972 cells was either (1) the minimum successful replications for any of the procedures attaining 50 or (2) 100 replications having been accomplished.

As for the size of both components to reflect $\pi$, a uniform (0,1) distribution of size 160 was first generated. Then the number of observations less than $\pi$ was used as the size of component 1 while 160 minus the size of component 1 was the size of component two.

### 4.5 Secondary Analyses

The generated raw data were subjected to secondary analyses to investigate the four research issues raised in section 2.7. The methods of analyzing raw data regarding each issue are described below:

1. Three loglinear models were adopted to address the problem of failure to locate a maximum: one for iteration exceeding rate, another one for singularity rate, and the other one for failure rate (sum of both rates mentioned above). A dichotomous variable stands for the performance was formed to represent the frequency of success or failure in each design. There were 144 cells (3 methods * 2 initial inputs * 2 assumptions of variance * 2 mean separations * 2 variance ratios * 3 proportions) by two levels of performance ). The purpose of loglinear model was to find the categorical variable or combination which accounts for significant $\chi^2$ loss in terms of degrees of freedom when introduced to the model.

2. A meta-analysis (Hedges and Olkin, 1985) design was conducted to address the issue of correlation between asymptotic variance measures and the absolute distance of corresponding estimates from their true parameter values. A valid replication within a cell was defined as a successful replication (i.e. iterations not exceeding preset values, and none of the parameter estimates going to boundary values), with asymptotic variances being positive through five parameter estimates. One correlation coefficient index was generated under the condition of valid replications larger than five in each cell.for each

estimated parameter. We have six grouping variables. A stepwise regression was conducted first to identify an order to enter them into the meta-analysis. Roughly "averaging" the order across the five parameters, we arrived at this order to enter into the design in an attempt to extract homogeneous groupings: variance ratio, mean separation, model assumption of variance, estimation method, initial input, and component proportion. Partitioning is continued until homogeneous grouping is established or until all available partitioning variables have been exhausted.

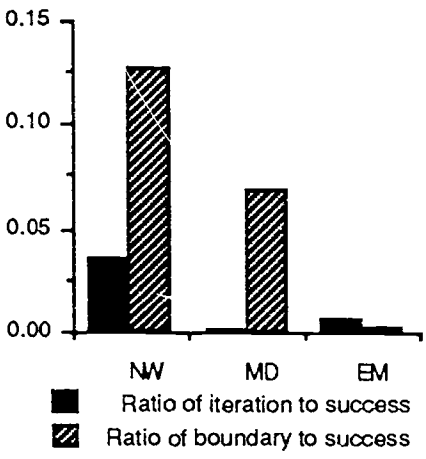
3. To compare the three algorithms with respect to mean separation, variance ratio, and component proportion, a repeated measures analysis was conducted using the three estimation procedure variables repeated across three of the distribution variables. The variable of shape combination was temporarily not included in the design. Dependent criteria were $(\theta^N - \theta)^2$ as a measure of MSE and $(\theta^N - \theta)$ as a measure of bias for each of the five parameters, resulting in 10 ANOVAs.

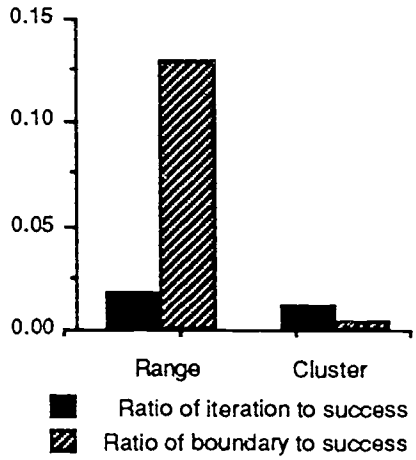
4. To investigate the research question about the robustness of estimates to non-normal mixtures, repeated measures analysis was conducted again. It was intended to adopt the design mentioned above but with shape combination factor added. However, due to memory capacity of the IBM 3180 mainframe (the IBM mainframe has 16 megabyte memory space while 68 megabytes was required by SAS to run the original design.), the design had to be reduced by dropping one of the three data distribution variables. Our strategy was to run two repeated measures independently: one where mean separation was 0.5 and the other for mean separation of 2.6. The mentioned repeated designs were run separately for each of the five parameters for mentioned two measures of MSE and bias, resulting in 20 ANOVAs .

## 5. Results and Conclusion

### 5.1 Failure Rate and Loglinear Model

A saturated loglinear model was imposed on all the mentioned 288 cells. None of the effects predicted cell frequency significantly. However, when we conducted a similar model but with boundary rate replaced by iteration exceeding rate, we obtained a significant effect of method by performance ($\chi^2$=6.26, df=2, p=0.04) which indicated that cell frequencies are depended on method. When a third loglinear model where both boundary rate and iteration exceeding rate were combined as failure rate, we obtained significant effects for method by performance ($\chi^2$=10.44, df=2, p=0.0054) and initial input by performance ($\chi^2$=6.33, df=1, p=0.0119). Figures 3 and 4 depict the ratio of success rate to failure rate by method and by initial guess, respectively.




Note: NW=modified Newton, MD=minimum distance, EM=expectation maximization algorithm

Figure 3. Ratio by Method

Figure 4. Ratio by Initial Input

Figure 3 shows that MD had the smallest iteration exceeding rate (0.16%), while EM had the smallest boundary rate (0.18%). NW ranked poorest for both iteration exceeding rate and boundary rate. As for the performance of initial inputs, Figure 4 indicates that cluster initials was the better strategy for initial values in terms of smaller boundary rate and

iteration exceeding rate than range initials. It is concluded that the use of EM with cluster initials is the preferred method to achieve a successful solution.

## 5.2 Correlations Between the Distance of Estimates to True Parameter Values and the Asymptotic Variances

The size of the correlation coefficient between asymptotic variance and absolute bias shared within each homogeneous group was reported as the result after conducting meta-analysis. There were 11422 correlation coefficients out of 11664 cells for the estimates of $\pi$, $\mu_1$, $\mu_2$, and $\sigma_1^2$, and 5717 coefficients for the estimates of $\sigma_2^2$.

First, we tested whether all coefficients share zero as their common correlation coefficient for each of the 5 parameters. The $\chi^2$ values indicated all the cell's correlations are not simultaneously zero. The test that these cells share the grand mean of their correlation coefficients also failed.

Our next step was to hierarchically group the cells according to the levels of the independent variables until some cells share a common correlation coefficient homogeneously. Finally we exhausted all the six independent variables and obtained 35 groups (rows in Table 1) which were homogeneous at least for one estimate, out of 144 groups. Each group contained observations of coefficients across 81 shape combinations. Table 1 shows the grouping variables, correlation coefficient sizes, $\chi^2$ values, and degrees of freedom.

In Table 1, under the situation where the variance ratio was equal one, there were 33 coefficients for the five parameters, among which 30 entries were in the range from -0.5 to 0.5. Three particularly large negative coefficients occurred when the estimation method was MD (minimum distance) under initial inputs from the cluster method with the assumption of homogeneity (rows 1 to 3 in Table 1) imposed on unimodal mixtures of homogeneous components, implying that in these kinds of mixtures, the larger the

asymptotic variance is, the smaller the MSE for $\sigma_1^2$. If we look at those cells that have variance ratio equal four, there were two groups: one with coefficient values less than 0.2 and the other one with values larger than 0.85. The homogeneous assumption seems to produce large coefficients under all the three methods in bimodal mixtures of heterogeneous components (VR=4, MS=2.6, and ASS=hom in Table 1), implying that in bimodal mixtures, a large asymptotic variance could be an indication of misspecification of variances. Considering the performance of the three estimation methods from the 33 groups, we found seven groups were from EM (expectation-maximization algorithm), four groups were from NW (modified Newton), and 22 groups were from MD. MD tended to create more homogeneous groups than the other two methods, and it seems to produce large values for the estimates of $\sigma_1^2$, and small values for the other four estimates, meaning that the asymptotic variances are not able to predict the stability of estimates except for $\sigma_1^2$.

### 5.3 Comparisons of Three Algorithms in terms of Parameter Estimate Precision

To address research questions 3 and 4, we described the results from the repeated measure design conducted over the raw data set. Because replications with missing values were dropped from the analysis, i.e., only the replications which were successful on all of the 12 estimation procedures (3 methods * 2 initials * 2 models of assumption) were included, 32999 valid observations were analyzed out of the total 79424. The values of the estimates of $\sigma_2^2$ in the mixtures of homogeneous model were copied from the one of the estimates of $\sigma_1^2$. The test for sphericity for MSE and bias from each of the five parameters, respectively, rendered significant $\chi^2$ values; therefore Greenhouse-Geisser adjusted degrees of freedom were used to conduct conservative significance tests. The term, unimodal mixtures, stands for mixtures of two components that differ by 0.5; while bimodal mixtures for two components that differ by 2.6. The term, homogeneous

components, means two components of variance ratio 1; while heterogeneous components for two components of variance ratio 4.

To address the research question 3, the variable, shape, was not included in the design, thus the groupings of the replications always contained the ones from the 81 levels of shape combinations. Regarding the MSE of the three methods under estimation procedure variables (initial selection and model adoption), and data distribution variables (mean separation, variance ratio, and component proportion), we have 214 interaction figures support the following results. Only some of those are presented in this article.

NW and EM were less sensitive to the use of different initials and produced equally acceptable estimates. MD was sensitive to different initials (see Figure 5). With range initials, MD performed better than NW and EM in unimodal mixtures and worse than NW and EM in bimodal mixtures. Nonetheless, cluster initials with MD in most cases outperformed all the other procedures.

There are two types of incorrect model assumption: homogeneous model imposed upon heterogeneous data and heterogeneous model imposed upon homogeneous data. For NW and EM, a homogeneous model imposed on heterogeneous data produced larger MSE for the parameters $\pi$, $\mu_1$, $\sigma_1^2$, and $\sigma_2^2$ compared to a heterogeneous model imposed on homogeneous data (see Figure 6). The estimates of $\mu_2$ did not reflect significant difference by different model assumption. Thus, the better model for NW and EM is heterogeneous in exploratory studies. MD was less affected by imposing an incorrect model in estimating $\pi$, $\mu_1$, $\mu_2$, and $\sigma_1^2$. However, for the estimates of $\sigma_2^2$, MD with cluster initials produced the largest MSE among all the estimation procedures if a homogeneous model was incorrectly imposed on heterogeneous data (see Figure 7). Therefore, the better model for MD is also heterogeneous. Comparing the twelve estimation procedures (3 methods by 2

initials by 2 model assumptions) in terms of MSE, the best was MD with cluster initials when assuming heterogeneous model.

To address the research question 4, we focused on the best estimation procedure, cluster initials combined with a heterogeneous model, for the three methods. There are 67 interaction graphs of method by shape combination to support our results. Some representative ones are depicted in this article. Their performance varied by parameters and shape combinations. MD performed as well as or better than NW and EM in estimating $\pi$ and $\sigma_1^2$ disregard what shape combinations the mixtures had (see Figure 8). As for the other three parameters, MD generally was superior to NW and EM in mixtures of homogeneous components. In mixtures of heterogeneous components, EM was more sensitive than MD to the presence of positive or negative skewness in estimating $\mu_1$, $\mu_2$, and $\sigma_2^2$ (see Figures 9 and 10). However, the MSE difference was small in the cases where MD was poorer than EM for the parameters $\mu_1$ and $\mu_2$. Generally, MD was most robust to non-normality and to uneven component proportion.

Under the use of cluster initials, we concluded that the direction of bias is affected by two factors, the number of modes of mixtures and the component variance ratio. In unimodal mixtures of homogeneous components, we found that the estimates of $\mu_1$ and $\mu_2$ were overestimated, and $\sigma_1^2$ and $\sigma_2^2$ were underestimated. In unimodal mixtures of heterogeneous components, the estimates of $\mu_1$ and $\mu_2$ tended to be biased in opposite directions away from each other and the estimates of $\sigma_1^2$ and $\sigma_2^2$ were biased toward each other. In bimodal mixtures of homogeneous components, bias of the estimates of $\mu_1$, $\mu_2$, $\sigma_1^2$ and $\sigma_2^2$ fluctuated around zero. It is found that positive skewness tends to cause negative bias while negative skewness tends to cause positive bias. However, in heterogeneous components, $\mu_1$ and $\mu_2$ were all overestimated, and $\sigma_1^2$ and $\sigma_2^2$ were nearer

to each other. The above tendency can be applied to the three methods. MD generally produced estimates with sizes smaller than those produced by NW and EM. Therefore, MD estimates were less positively biased if the tendency was positive bias such as for $\pi$, $\mu_1$, and $\mu_2$ (see Figure 11), and more negatively biased if the tendency was negative bias such as for $\sigma_2^2$ (see Figure 12)

## 6. Discussion and Suggestions

### 6.1 Discussion

As mentioned in section 2.1, both Leytham (1984) and Hamilton (1991) reported singularity rate. The common characteristics in both studies were that the largest sample size was 100, that EM was used as the estimation procedure, and that true parameter values were used as initial inputs. However, the two studies led to different conclusions. The two largest singularity rates in Hamilton's study were caused by either the component proportion being as large as 0.9, or the two components being identified. In Leytham's study, all the mixtures were bimodal.

In this study, our main interest is to investigate the relative performances of different methods in an approximation of typical research situations. Therefore, we conducted simulations with a sample size of 160 and used by NW, EM and MD with two types of initial inputs. Our results indicated that EM produced the smallest singularity rate (0.02%) among the three methods. However, the loglinear model did not show significant dependency of singularity rate on method, model assumption, mean separation, variance ratio, and component proportion. Therefore, with a sample size of 160, singularity was not shown to be a concern in either unimodal or bimodal mixtures of homogeneous or heterogeneous components with the smallest component proportion not less than 0.3 by the use of EM method with imprecise initial inputs.

Imposing of a homogeneous model unto heterogeneous data was discussed by Basford and McLachlan (1985). They found that the wrong model particularly affected the estimates of component proportions. Our results show that other than the component proportion, the wrong model also affected the estimates of $\mu_1$, $\sigma_1^2$, and $\sigma_2^2$. For the parameter $\mu_2$, the selection of the model for the three methods was not as crucial because there was no interaction of method by model. For the other four parameters, the 'homogeneous model to heterogeneous components' combination caused the largest MSE compared to the other three types of model-component combination in most mixture forms. An exception was for the estimates of $\sigma_1^2$ in unimodal mixtures where the largest MSE occurred when a heterogeneous model was imposed on heterogeneous data using NW and EM. This phenomenon implies that in unimodal mixtures, a homogeneous model is preferred if component variances are of most interest and either of the estimation procedures of NW and EM is used. Since the estimates of proportion and location parameters are of most interest, we consider that the superiority of the heterogeneous model has been demonstrated and that MD proved to be more robust to mismatch between model assumption of variance and data.

Regarding the sensitivity of MD and EM to non-normality, conclusions from our study may be compared to those of Woodward, Parr, Schucany, and Lindsey (1984). In our study, we used both unimodal and bimodal mixtures. The normal mixture (shape combination 22) is one of the 81 manipulated shape combinations. We demonstrated that MD dominated EM in both unimodal and bimodal mixtures for the estimates of $\pi$. As for the other four parameters, the interaction diagrams of method by shape combination show that MD fluctuated less than EM and was more robust to various shape combinations, but the superiority of one method over the other is not absolute through all the parameters. Our study, does not support the conclusion that EM is better overall than MD under normal

mixtures (see Figure 8), and was neither poorer than MD under non-normal mixtures (see Figure 10).

## 6.3 Suggestions

### 6.3.1 Suggestions for Practitioners

For a practitioner, the most important issue is the selection of the most appropriate estimation procedure. To estimate the parameters of a finite mixture, MD and EM have their own strengths and deficits. The strengths of EM are the lowest singularity rate and insensitivity to the choice of initial inputs. The strengths of MD are its robustness to non-normality and incorrect model assumption of variance. As for availability, EM is easy to program yet there is no handy EM program available for general public. However, MD is installed in popular package IMSL as a general least squares method that only needs a function to define the least squares as introduced in section 3.3.

If only the component proportion parameter is to be emphasized in a practical situation, MD performs equally well with both range and cluster initials and is preferred to EM by both smaller MSE and less bias. If both proportion and location parameters are important, then the availability of initials needs to be considered. With the use of cluster initials which can be obtained by a regular cluster analysis, MD performs better than EM in terms of both MSE and bias, but with simple range initials, MD is inferior to EM. When all the parameters, including scale parameters, are to be estimated, our results suggest the use of cluster initials. Also, their performances differ by non-normality for the estimates of $\sigma_2^2$ in heterogeneous mixtures. Neither MD nor EM was superior over the other through all the shape combinations. EM is preferred if the second component is symmetric or positively skewed while MD is preferred if the second component is negatively skewed. This suggestion should be considered only when the scale parameters are more important than the other parameters; otherwise, MD should be preferred.

Our study found that a homogeneous model imposed unto heterogeneous mixtures produced the largest MSE among the four model-component combinations. Therefore we recommend using a heteroscedastic model.

Our study suggests that asymptotic variances are too unstable to predict the bias in parameter estimates if a heterogeneous model is assumed. If homogeneous model is imposed on a data set and a large value of the asymptotic variance of parameter $\sigma_1^2$ is found, it could be an indication that the wrong model has been used.

### 6.3.2 Suggestions for Further Research

Woodward et al. (1984) has concluded that the results for n=100 were not substantially different from n=50 or n=200 based on their bimodal mixtures. In another words, small sample size did not have a significant effect in bimodal mixtures. In order to complete the comparison between MD and EM in terms of non-normality, a side issue would be how both methods perform with small sample sizes in unimodal mixtures.

Hamilton (1991) has noted the singularity problem occurs when a component proportion is as small as 0.1. How MD compares with EM for small values of $\pi$ is also interesting.

MD has been concluded in the study to be the better estimation procedure. However, its performance in determining the number of components has not been discussed in literature. Wolfe's (1971) modified likelihood ratio can be used to test the null hypothesis of component number equal $C_0$ against the alternative hypothesis of component number equal $C_1$ where $C_1 > C_0$. A comparison of EM with MD on small sample sizes, component proportion close to boundary, and robustness of this procedure for various true number of components would be useful.

Table 1. Correlation Coefficients between the Distance of Estimates from True Parameter Values and Corresponding Asymptotic Variances from Homogeneous Groups for Five Parameters by Data Distribution Variables

| VR | MS | AS | ME | I | P | $\pi$ | $\mu_1$ | $\mu_2$ | $\sigma_1^2$ | $\sigma_2^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.5 | hom | MD | C | 5 | | | | -0.9523<br>4.4 (80) | |
| 1 | 0.5 | hom | MD | C | 3 | | | | -0.8754<br>9.3 (80) | |
| 1 | 0.5 | hom | MD | C | 7 | | | | -0.8575<br>11.6 (80) | |
| 1 | 0.5 | het | EM | R | 5 | -0.2810<br>65.3(80) | | | | |
| 1 | 0.5 | het | EM | R | 3 | -0.1057<br>77.0(80) | | | | |
| 1 | 0.5 | het | EM | C | 5 | -0.2812<br>68.8(80) | | | | |
| 1 | 0.5 | het | EM | C | 3 | -0.1433<br>87.0(80) | | | | |
| 1 | 0.5 | het | EM | C | 7 | -0.1348<br>99.8(80) | | | | |
| 1 | 0.5 | het | NW | C | 3 | -0.2025<br>100.8(80) | | | | |
| 1 | 0.5 | het | MD | R | 5 | -0.1786<br>80.6(67) | -0.1047<br>76.4(67) | | -0.1834<br>61.2(67) | -0.1969<br>84.8(67) |
| 1 | 0.5 | het | MD | R | 3 | -0.1038<br>92.3(73) | | -0.1330<br>87.5(73) | | |
| 1 | 0.5 | het | MD | R | 7 | -0.0308<br>63.8(74) | 0.1114<br>75.2(74) | | | |
| 1 | 0.5 | het | MD | C | 5 | | 0.0260<br>89.6(80) | | | |
| 1 | 2.6 | hom | MD | R | 5 | -0.1010<br>27.2(33) | 0.0026<br>35.8(33) | 0.1939<br>43.4(33) | 0.3428<br>32.4(33) | |
| 1 | 2.6 | hom | MD | R | 7 | | | | 0.4639<br>85.9(69) | |
| 1 | 2.6 | het | MD | R | 5 | -0.1678<br>54.9(42) | | | | |
| 1 | 2.6 | het | MD | R | 3 | | -0.0545<br>54.3(80) | -0.0439<br>67.3(80) | 0.1560<br>94.9(80) | |
| 1 | 2.6 | het | MD | R | 7 | | -0.0040<br>72.2(80) | | | |
| 1 | 2.6 | het | MD | C | 3 | -0.0758<br>78.4(79) | | -0.0711<br>87.4(79) | | |
| 1 | 2.6 | het | MD | C | 7 | | -0.0243<br>76.0(67) | | | |

Table 1. (Continued)

| VR | MS | AS | ME | I | P | $\pi$ | $\mu_1$ | $\mu_2$ | $\sigma_1^2$ | $\sigma_2^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 0.5 | hom | MD | C | 3 | | | | 0.0666<br>64.6(80) | |
| 4 | 0.5 | het | MD | R | 7 | -0.1127<br>82.7(80) | | | | |
| 4 | 2.6 | hom | EM | R | 5 | | | | 0.9271<br>51.6(80) | |
| 4 | 2.6 | hom | EM | C | 5 | | | | 0.9262<br>60.5(80) | |
| 4 | 2.6 | hom | NW | R | 5 | | | | 0.9373<br>4.2(72) | |
| 4 | 2.6 | hom | NW | R | 7 | | | | 0.9441<br>1.0(44) | |
| 4 | 2.6 | hom | NW | C | 5 | | | | 0.9376<br>8.6(80) | |
| 4 | 2.6 | hom | MD | C | 3 | | | | 0.8986<br>7.5(80) | |
| 4 | 2.6 | het | MD | R | 3 | | -0.0438<br>97.9(80) | | | |
| 4 | 2.6 | het | MD | R | 7 | | -0.1014<br>72.7(56) | | | |
| 4 | 2.6 | het | MD | C | 5 | -0.0489<br>87.7(80) | | | | |
| 4 | 2.6 | het | MD | C | 3 | -0.0514<br>97.9(80) | | | | |
| 4 | 2.6 | het | MD | C | 7 | | | -0.0385<br>94.7(78) | | |

Note: $\chi^2$ at the second row. Degrees of freedom in parentheses.
VR=variance ratio
MS=mean separation
AS=assumption of Variance (hom=homogeneity, het=heterogeneity)
ME=method (NW=modified newton, MD=minimum distance)
I=initial inputs (R=range method, C=Cluster method)
P=proportion of the first component (3=0.3, 5=0.5, 7=0.7)

## References


Basford, K.E. and McLachlan, G.J. (1985). Likelihood estimation with normal mixture models. *Applied Statistics, 34*, 282-289.

Behboodian, J. (1970). On the modes of a mixture of two normal distributions. *Technometrics,* 12, 131-139.

Davenport, J.W., Bezdek, J.C. and Hathaway, R.J. (1988). Parameter estimation for finite mixture distributions. *Computers and Mathematics with Applications, 15*, 819-828.

Day, N.E. (1969). Estimating the components of a mixture of two normal distributions. *Biometrika, 56*, 463-474.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B 39*, 1-38.

Dennis, J.E. and Schnabel, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, New Jersey.

Eisenberger, I. (1964). Genesis of bimodal distributions. *Technometrics, 6,* 357-363.

Everitt, B.S. (1984). Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions; a comparison of different algorithms. *The Statistician, 33*, 205 -215.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

Fowlkes, E.B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association. 74*, 561-575.

Fryer, J.G. and Robertson, C.A. (1972) A comparison of some methods for estimating mixed normal distributions. *Biometrika, 59*, 639 -648.

Hamilton, J.D. (1991). A quasi-Bayesian approach to estimating parameters for mixtures of normal distributions. *Journal of Business & Economic Statistics, 9*, 27-39.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics, 8*, 431-444.

Hathaway, R.J. (1986). A constrained EM algorithm for univariate normal mixtures. *Journal of Statistical Computation & Simulation, 23*, 211-230.

Hathaway, R.J. (1985). A constrained formulation of maximum -likelihood estimation for normal mixture distributions. *The Annals of Statistics, 13*, 795-800

Hathaway, R.J., Huggins, V.J. and Bezdek, J.C. (1984). A comparison of methods for computing parameter estimates for a mixture of normal distributions. *Proceedings of Pittsburgh Conference on Modeling and Simulation, 15th A,* 1853-1860.

Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. San Diego: Academic Press.

Hosmer, D.W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics, 29*, 761-770

Kumar, K.D., Nicklin, E.H. and Paulson, A.S. (1979) Comment on 'Estimating mixtures of normal distributions and switching regressions', *Journal of the American Statistical Association, 74*, 52-56.

Leytham, K.M. (1984). Maximum likelihood estimates for the parameters of mixture distributions. *Water Resources Research, 20*, 896-902.

Marquardt, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial Engineers, 11*, 431-441.

McLachlan, G.L. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

Parr, W.C. and Schucany, W.R. (1980). Minimum distance and robust estimation. *Journal of the American Statistical Association, 75*, 616-624.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London, Series A*, 71-110.

Quandt, R.E. and Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions (with discussion). *Journal of the American Statistical Association, 73*, 730-752.

Ramberg, J.S., Dudewicz, E.J., Tadikamalla, P.R. and Mykytka, E.F. (1979). A Probability distribution and its uses in fitting data. *Technometrics, 21*, 201-214.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review, 26*, 195-239.

Tan, W.Y. and Chang, W.C. (1972) Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association, 67*, 702-708.

Woodward, W.A., Parr, W.C., Schucany, W.R. and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *Journal of the American Statistical Association, 79*, 590-598.

## References of the Thirteen Application Studies

Baldetorp, B., Dalberg, M., Holst, U., and Lindgren, G. (1989). Statistical evaluation of cell kinetic data from DNA flow cytometry (FCM) by the EM algorithm. *Cytometry, 10*, 695-705.

Basford, K.E. and McLachlan, G.J. (1985). Likelihood estimation with normal mixture models. *Applied Statistics, 34*, 282-289.

Boothe, P. and Glassman, D. (1987). The statistical distribution of exchange rates - empirical evidence and economic implications. *Journal of International Economics (Netherlands), 22*, 297-319.

Dillon, W. R. and Mulani, N. (1989). LADI: A latent discriminant model for analyzing marketing research data. *Journal of Marketing Research, 26*, 15-29.

Friedlander, Y., Kark, J.D., Cohen, T.m Eisenberg, S., and Stein, Y. (1983). Admixture analysis of high density lipoprotein cholesterol distribution in a Jerusalem population sample. *Clinical Genetics, 24*, 117-127.

Gibbons, R.D. and Davis, J.M. (1986). Consistent evidence for a biological subtype of depression characterized by low CSF monoamine levels. *Acta-Psychiatrica Scandinavica, 74*, 8-12.

Ling, L. and Tolhurst, D.J. (1983). Recovering the parameters of finite mixtures of normal distributions from a noisy record: an empirical comparison of different estimating procedures. *Journal of Neuroscience Methods, 8*, 309-333.

Lwin, T. and Martin P.J. (1989). Probits of mixtures. *Biometrics, 45*, 721-732.

Millar, R.B. (1987). Maximum likelihood estimations of mixed stock fishery composition. Canadian *Journal of Fisheries and Aquatic Science, 44*, 583-590.

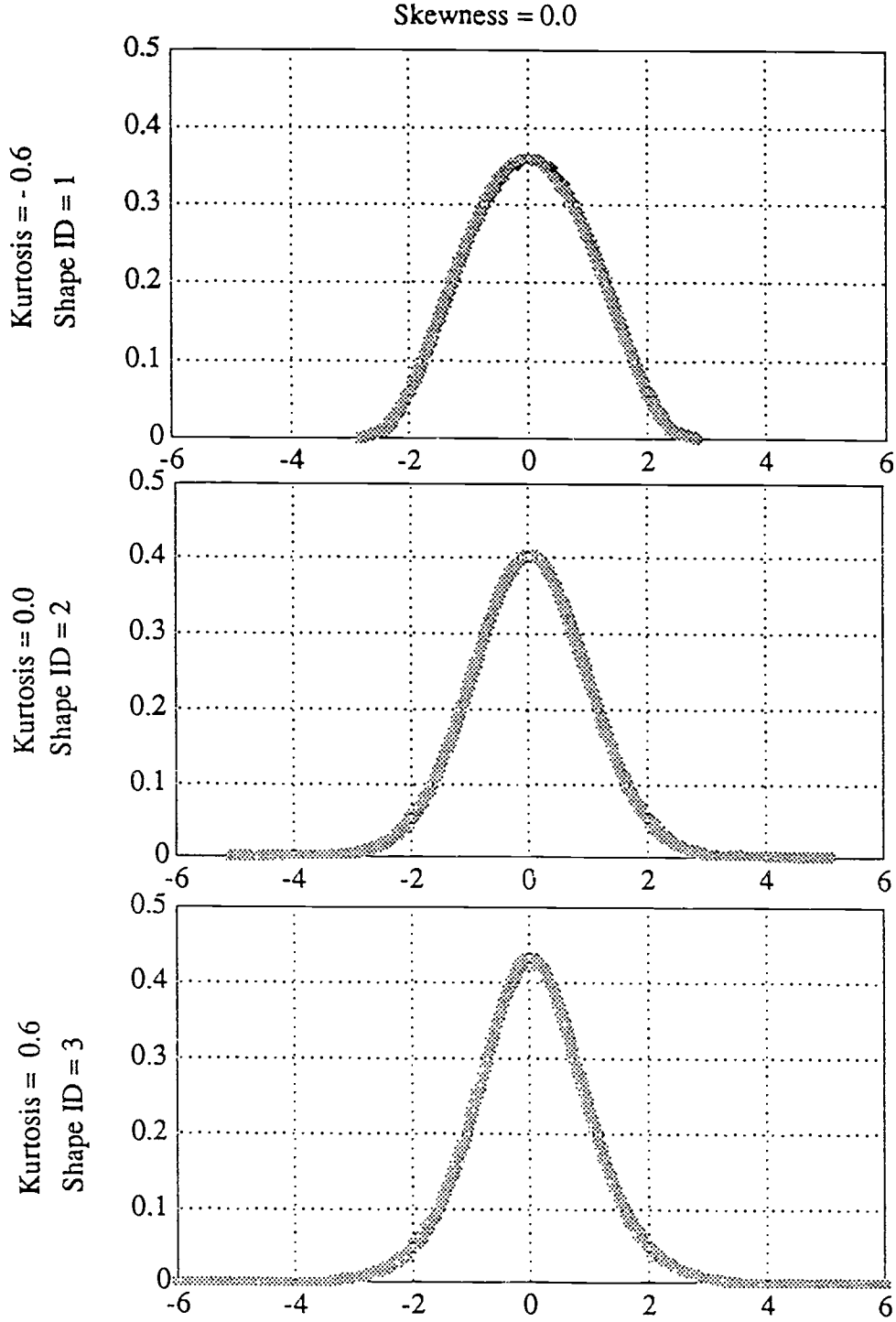
Moreno-Bello, M., Bonilla-Marin, M., and Gonzalez-Beltran, C. (1988). Distribution of pore sizes in black lipid membranes treated with nystatin. *Biochimica et Biophysica Acta, 944*, 97-100.

Summers, R.W., Nicoll, M., Underhill, L.G., and Petersen, A. (1988). Methods for estimating the proportions of Icelandic and British redshanks Tringa totanus in mixed populations wintering on British coasts. *Bird studies, 35*, 169-180.

Thomas, H. (1990). A likelihood-based model for validity generalization. *Journal of Applied Psychology*, 75, 13-20.

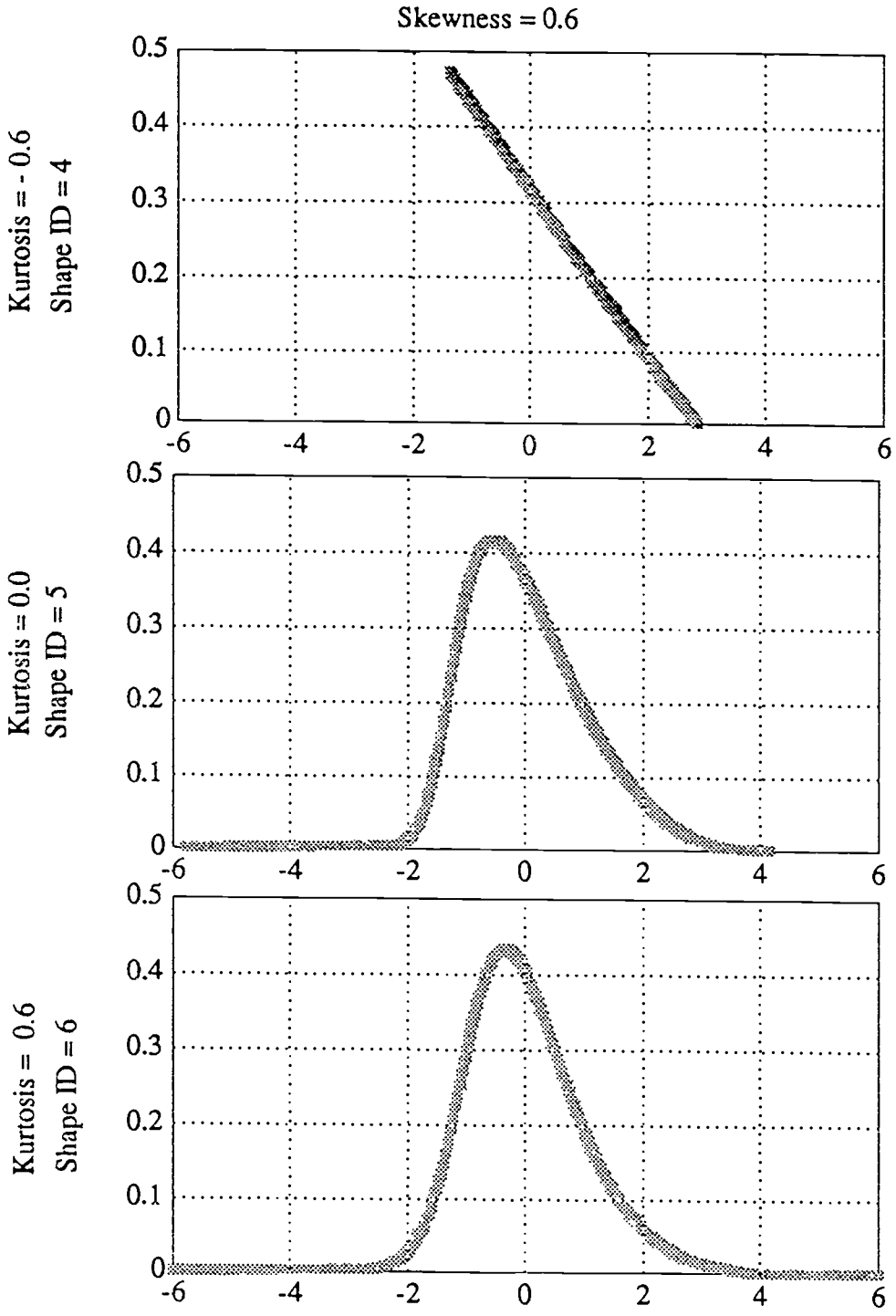
Westenberg, H.G.M. and Verhoeven, W.M.A. (1988). CSF monoamine metabolites in patients and controls: support for a bimodal distribution in major affective disorders. *Acta Psychiatrica Scandinavica, 78*, 541-549.

Figure 1. The Nine Shapes of Components in Mixtures, Represented by Skewness and Kurtosis. ($\mu = 0$ and $\sigma^2 = 1$)



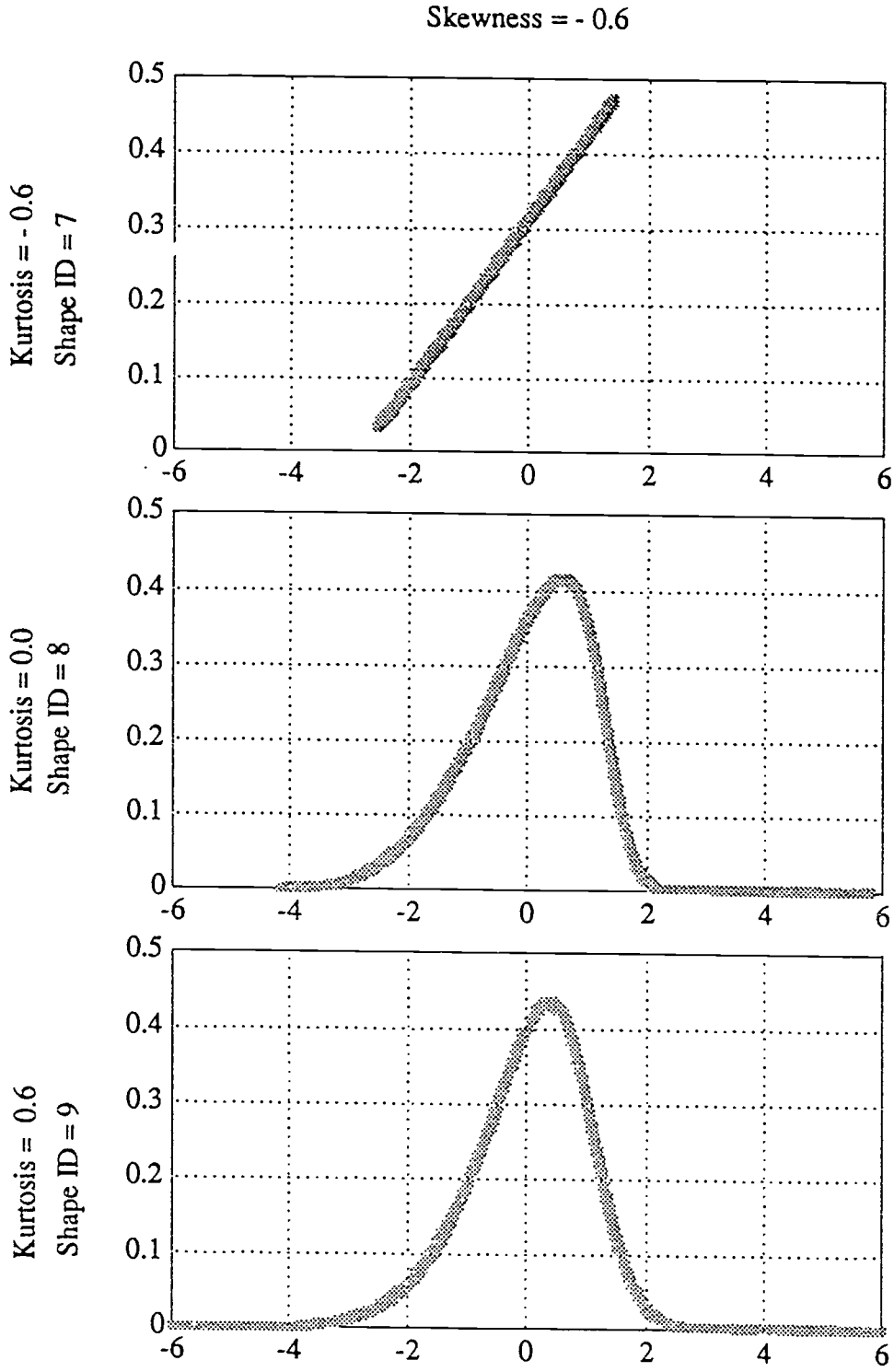

Note: Shape ID is a digit used to symbolize the characteristic of a component in the present study.

Figure 1 (Continued)



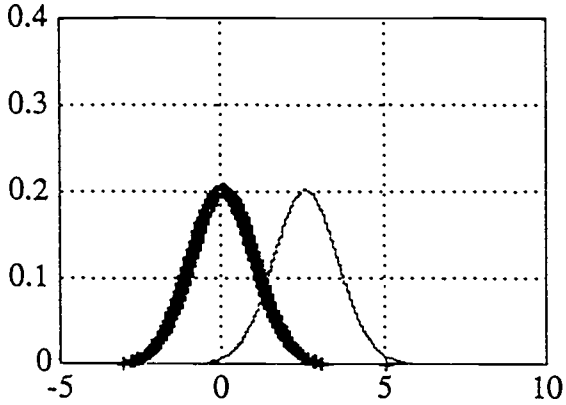

Note: Shape ID is a digit used to symbolize the characteristic of a component in the present study.

Figure 1. (Continued)

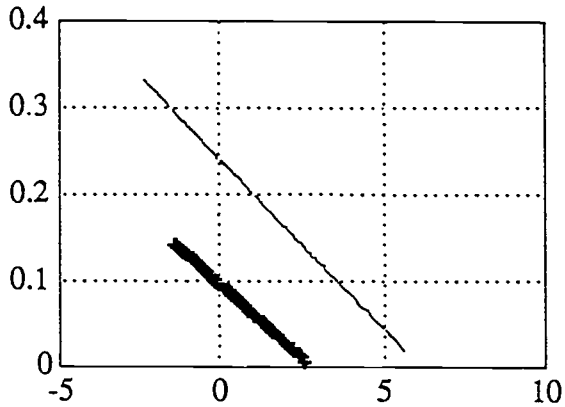



Note: Shape ID is a digit used to symbolize the characteristic of a component in the present study.

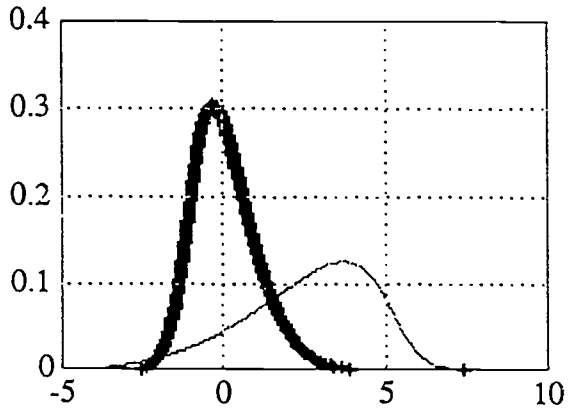
Figure 2. Four Examples of Mixtures



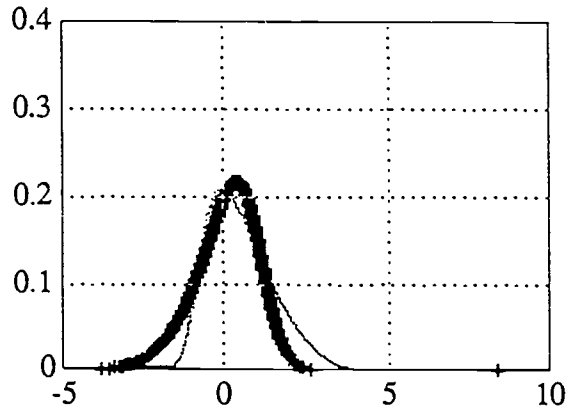

Component 1: S=0, K=0 (Shape ID=2)
Component 2: S=0, K=0 (Shape ID=2)
Mean Separation=2.6
Variance Ratio=1
Proportion of Component 1=0.5




Component 1: S=0.6, K=-0.6 (Shape ID=4)
Component 2: S=0.6, K=-0.6 (Sahpe ID=4)
Mean Separation=0.5
Variance Ratio=4
Proportion of Component 1=0.3




Component 1: S=0.6, K=0.6 (Shape ID=6)
Component 2: S=-0.6, K=0.0 (Shape ID=8)
Mean Separation=2.6
Variance Ratio=4
Proportion of Component 1=0.7




Component 1: S=-0.6, K=0.6 (Shape ID=9)
Component 2: S=0.6, K=0.0 (Shape ID=5)
Mean Separation=0.5
Variance Ratio=1
Proportion of Component 1=0.5

Note: ——— Component 1 ..... Component 2
S = Skewness, K = Kurtosis



32

Figure 5. A Five Way Interaction Plot, Method * Initials * Assumption of Variance * Mean Separation * Variance Ratio, on MSE

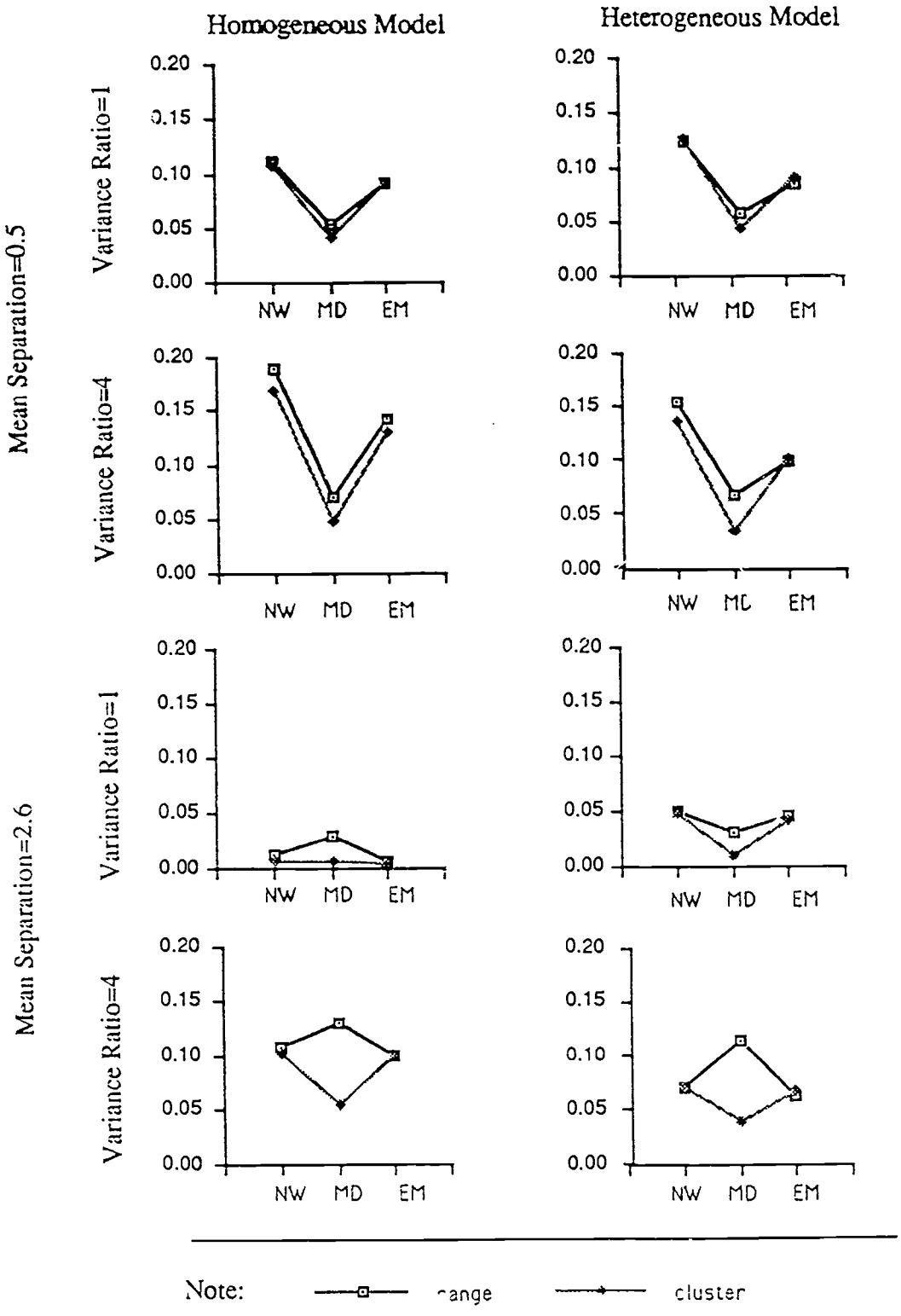


Figure 6. A Six Way Interaction Plot, Method * Initials * Assumption of Variance * Mean Separation * Variance Ratio * Proportion, on MSE index of $\mu_1$ (Mean Separation = 0.5)

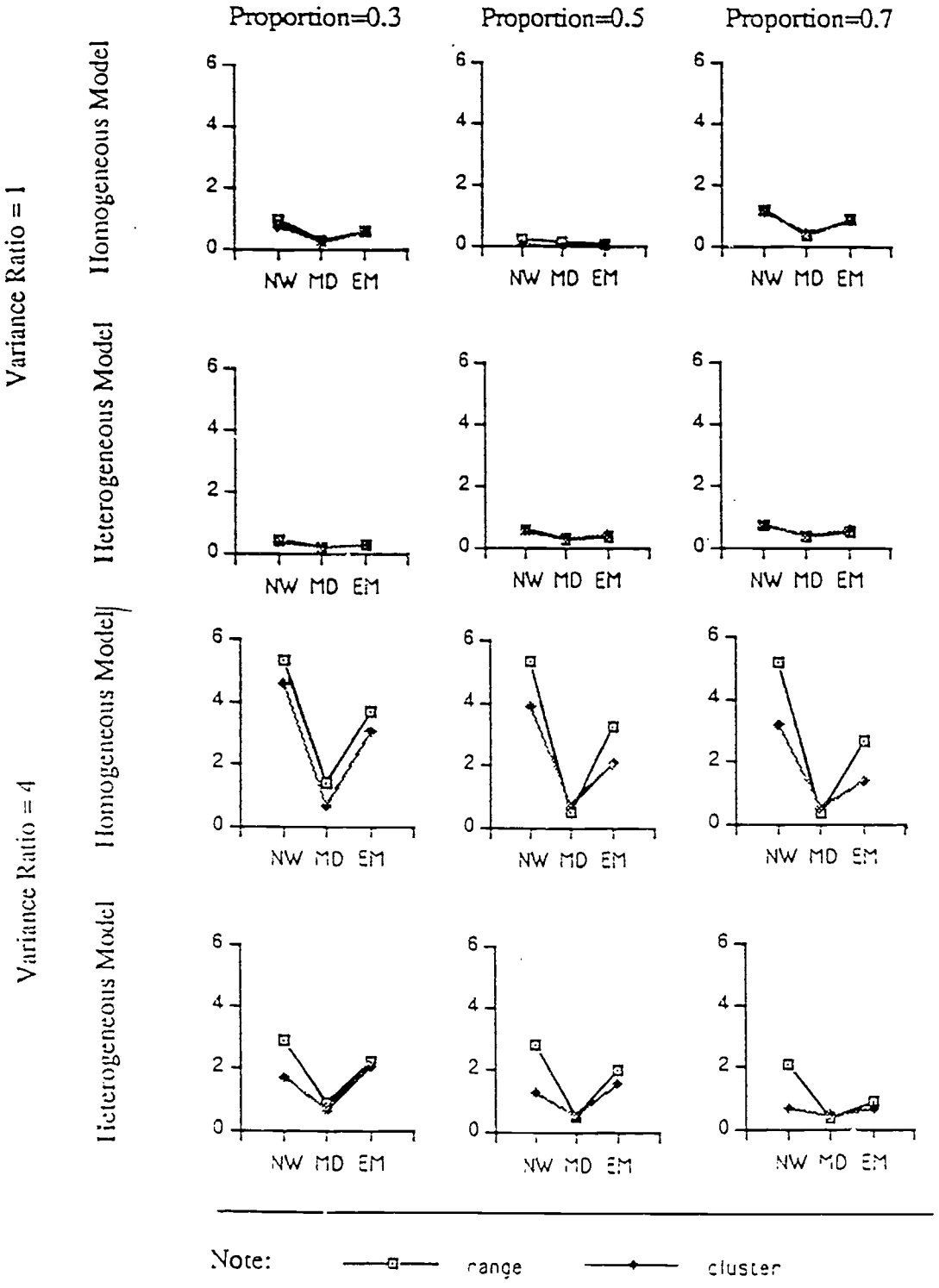
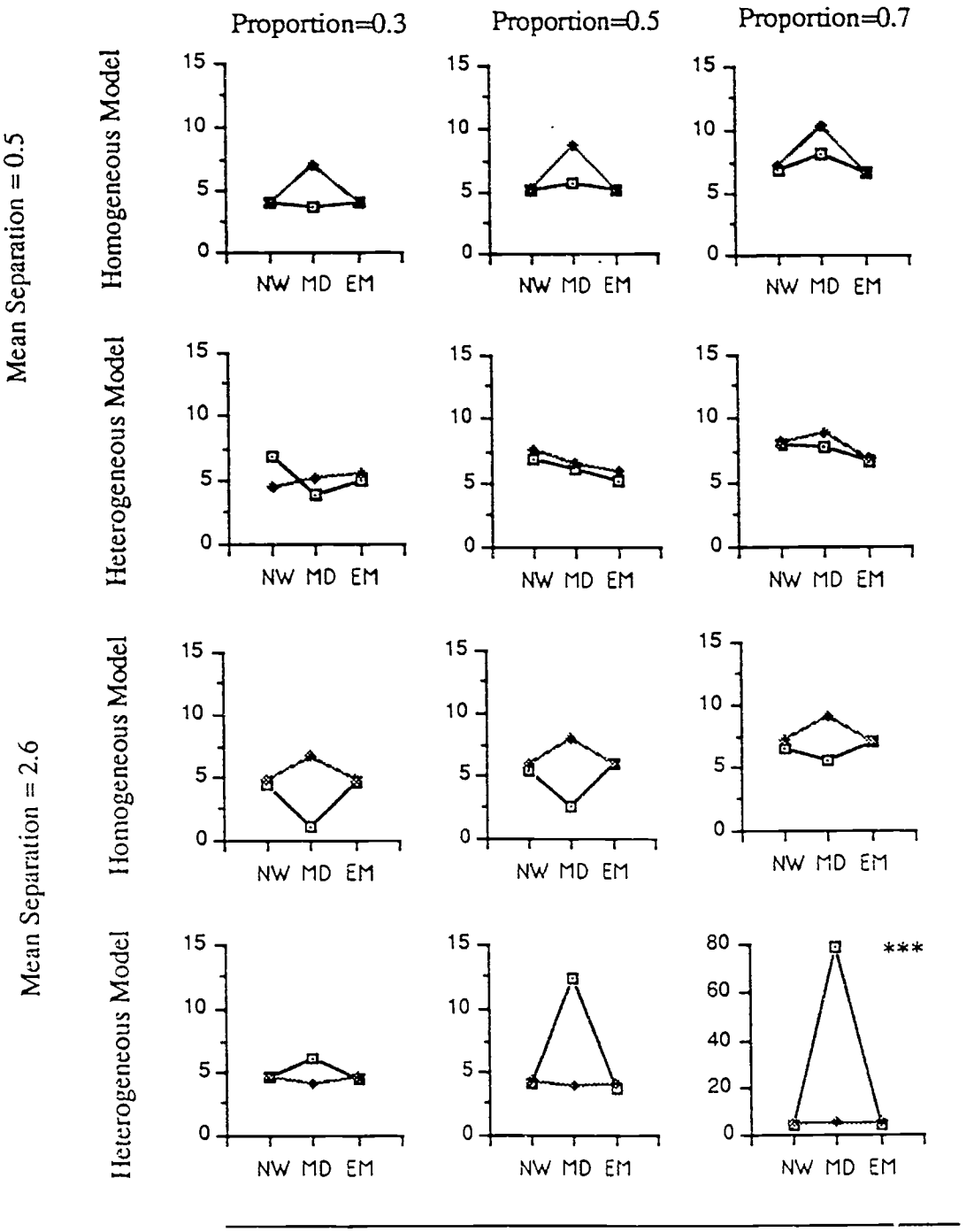


Figure 7. A Six Way Interaction Plot, Method * Initials * Assumption of Variance * Mean Separation * Variance Ratio * Proportion, on MSE index of $\sigma_2^2$ ( Variance Ratio = 4)






35

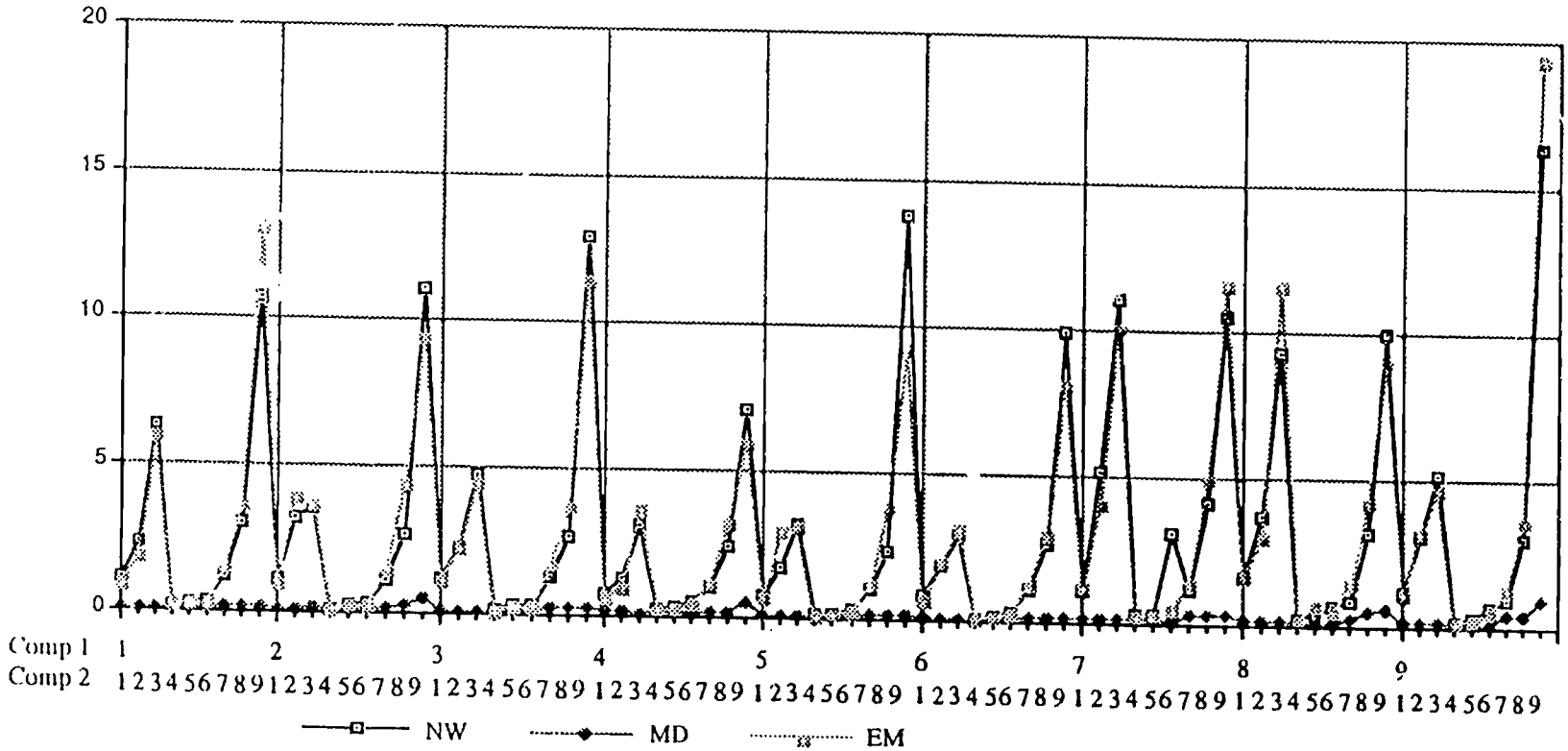



Figure 8. Interaction Plot of Shape * Method on the MSE of $\sigma_1^2$ under the Condition of Mean Separation=0.5, Variance Ratio=4, Initials=Cluster, and Assumption=Heterogeneity Model (Note: X-axis stands for the 81 combinations of the 9 shapes.)

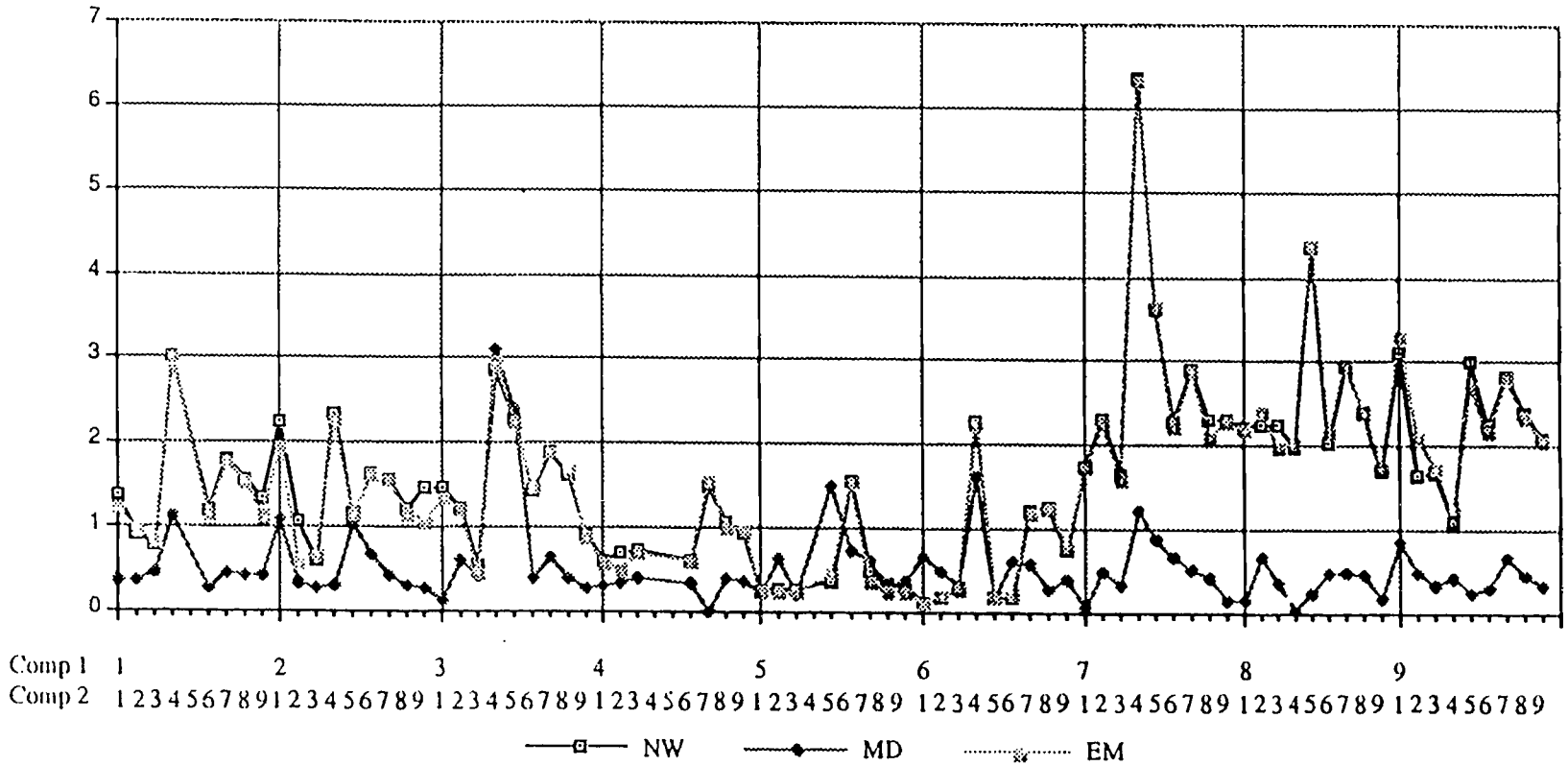



Figure 9. Interaction Plot of Shape * Method on the MSE of μ2 under the Condition of Mean Separation=2.6, Variance Ratio=4, π=0.7, Initials=Cluster, and Assumption=Heterogeneity Model
(Note: X-axis stands for the 81 combinations of the 9 shapes.)

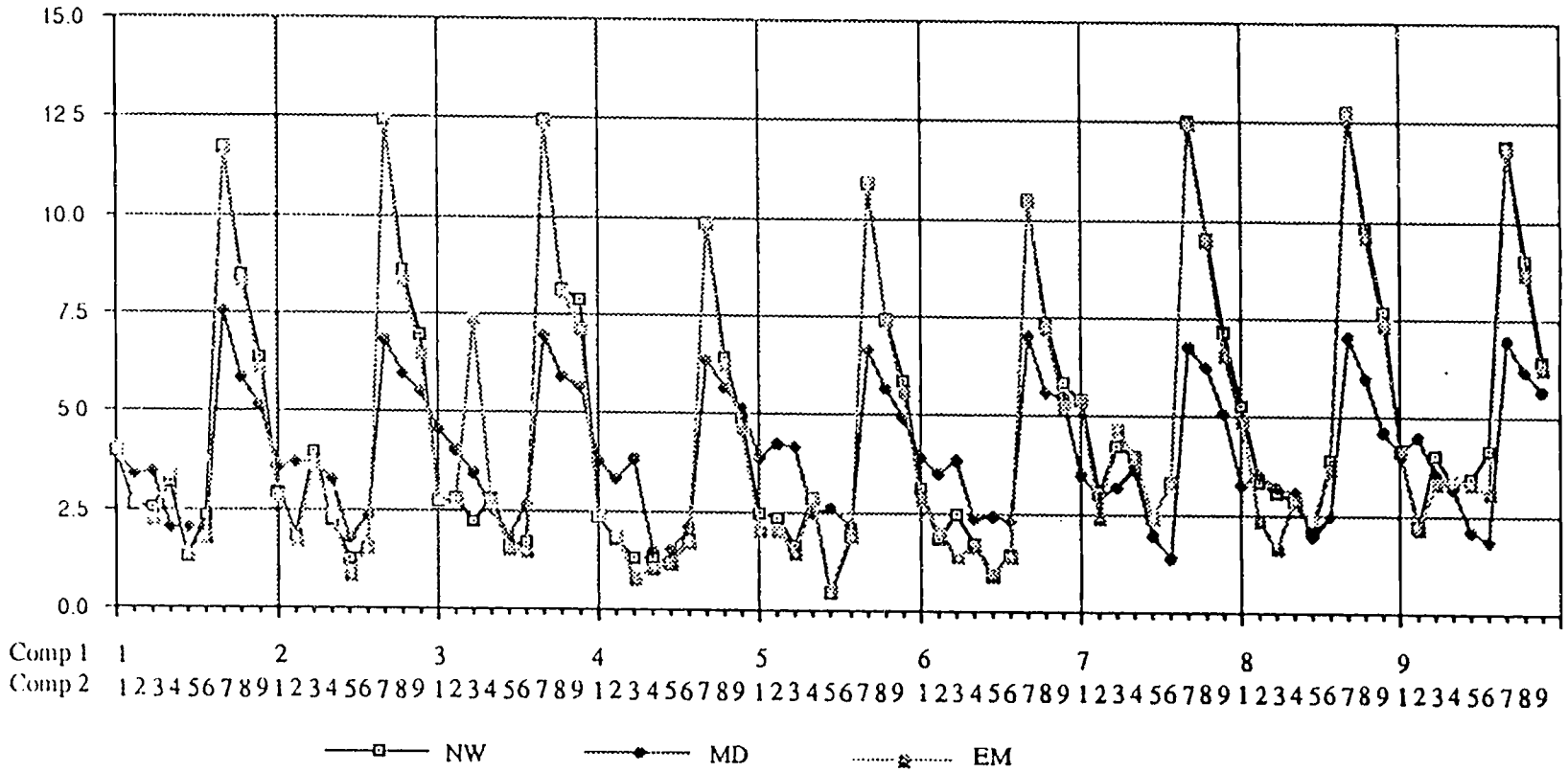



Figure 10. Interaction Plot of Shape * Method on the MSE of $\sigma_2^2$ under the Condition of Mean Separation=2.6, Variance Ratio=4, $\pi$=0.3, Initials=Cluster, and Assumption=Heterogeneity Model
(Note: X-axis stands for the 81 combinations of the 9 shapes.)

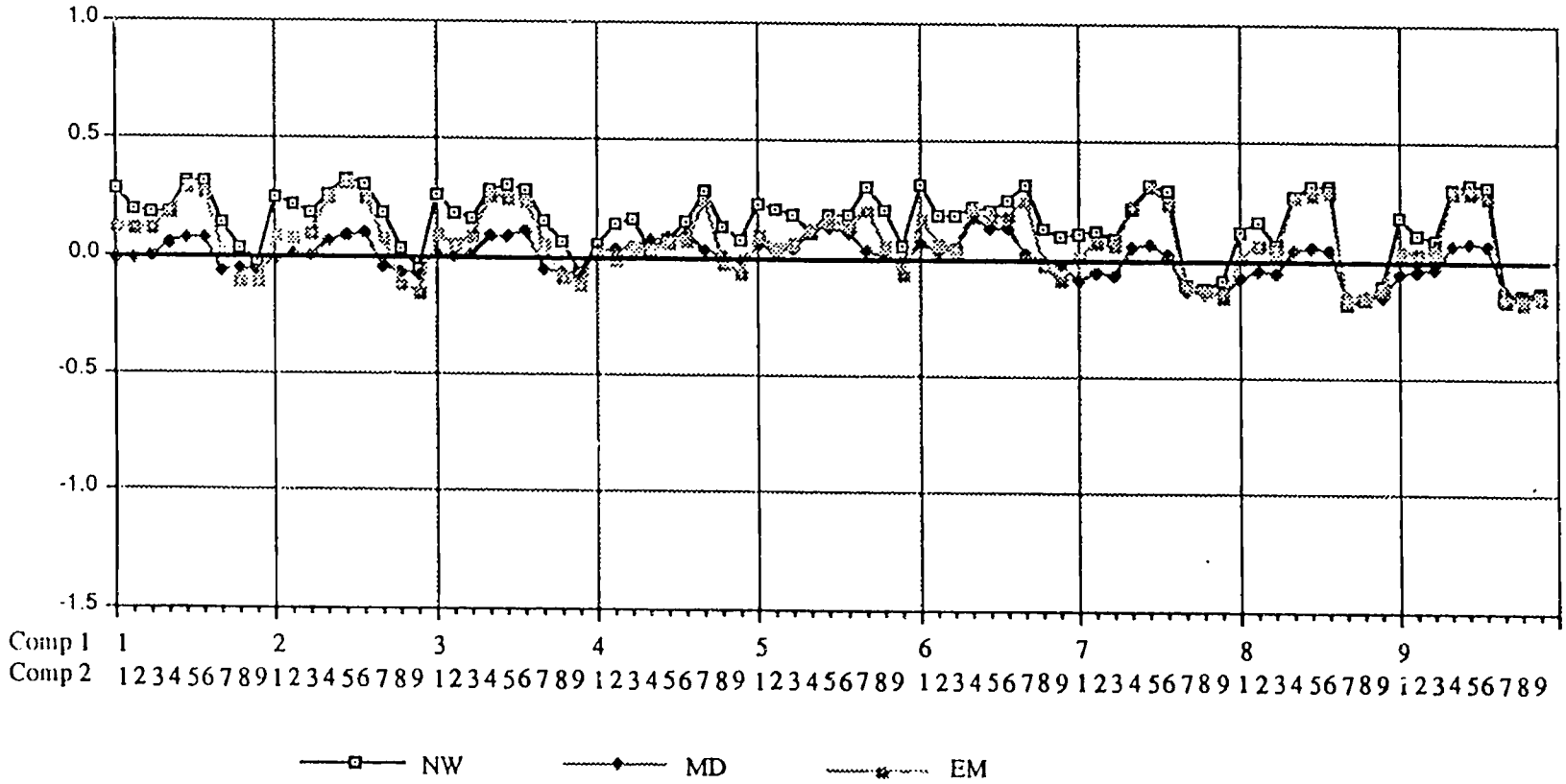



Figure 11. Interaction Plot of Shape * Method on the Bias of $\mu_1$ under the Condition of Mean Separation=0.5, Variance Ratio=4, Initials=Cluster, and Assumption=Heterogeneity Model (Note: X-axis stands for the 81 combinations of the 9 shapes.)

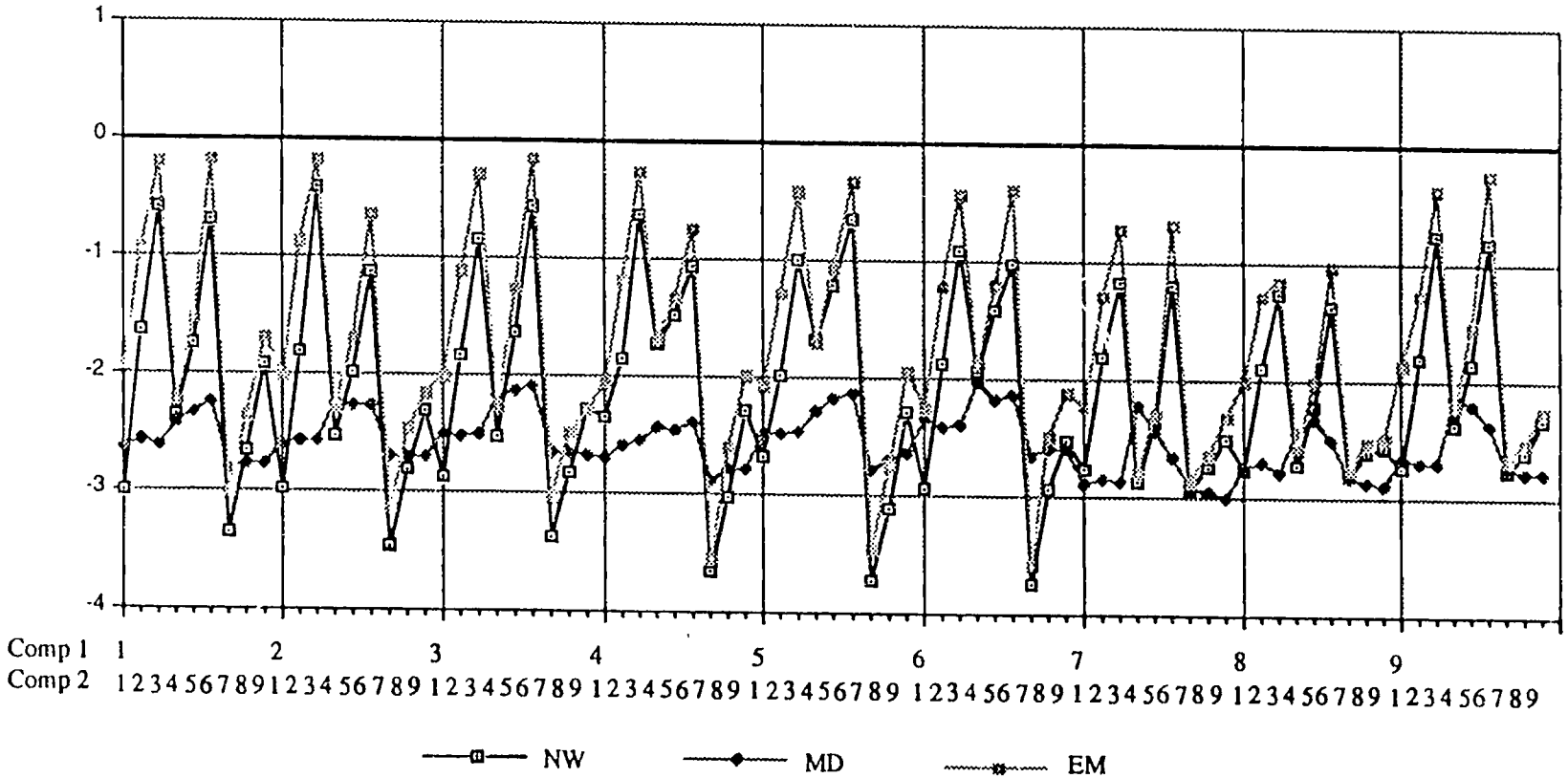



Figure 12. Interaction Plot of Shape * Method on the Bias of $\sigma_2^2$ under the Condition of Mean Separation=0.5, Variance Ratio=4, Initials=Cluster, and Assumption=Heterogeneity Model (Note: X-axis stands for the 81 combinations of the 9 shapes.)