

# Individual fit, heterogeneity, and missing data in multi-group SEM

Michael C. Neale

Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Box 980126, Richmond, VA 23298-0126

January 16, 2004

## Abstract

Analysis of raw data allows great flexibility in structural equation modeling. First, data in which some observations are missing at random or missing completely at random are automatically handled within this approach. Second, it is possible to fit finite mixture distribution models to raw data. Third, models for continuous moderator variables, including hierarchical linear models, become easy to specify. Fourth, measures of individual fit that are readily available may be used to detect outliers in the population, or population heterogeneity. The contribution to the likelihood fit function is one such measure, but it depends on the amount of missing data for the case in question. Two Q statistics for measuring individual fit as a z-score, which are independent of the amount of missing data, are compared for the bivariate case.

## Introduction

For many years, structural equation models have been fitted to summary statistics, primarily covariances, but sometimes to the means as well. More recently, programs such as Mx and Amos have shown the advantages of fitting models to raw data, and it is extensions of this method that form the main focus of this chapter. A frequently asked question is “How does one fit structural equation models to the raw data by maximum likelihood?” In some ways, this is a strange question because it is a simpler question to answer than one that asks about the origin of the formula used for fitting models to covariance matrices. Therefore, this chapter begins with an elementary introduction to maximum likelihood (ML), including the concepts of individual fit and the multivariate normal distribution. The second section discusses various alternative measures of individual fit, suitable for use when some of the data are missing. In the third section I consider moderator variables as a potential source of non-normality. If these moderators have been measured, it is possible to

---

The author is grateful for support from PHS grants MH-40828, HL48148, MH45268, MH49492, MH41953, AA09095, RR08123, MH01458

explicitly model their effects. For the case of binary moderators, the model may be specified as a two-group structural equation model, but continuous moderators require an extension to ML analysis of raw data such that there is a different model for every subject in the sample. Finally, although individual fit statistics can be useful for the detection of outliers or mixture distributions, and to judge the value of adding moderating variables, heterogeneity may always not be directly related to an observed moderator variable. Formal methods for detecting ‘latent’ heterogeneity require the application of finite mixture distributions, which are described in the fourth section.

## Fitting Models by Maximum Likelihood

### *Basic principles*

Likelihood is a simple concept based in probability theory. The relationship of likelihood to probability is so close that it can be confusing for the novice. To take a trivial example, suppose a coin is tossed 100 times, and that 53 times it shows heads. The probability of this outcome *given that the coin has a probability  $p = .5$  of heads* is easy to calculate from the binomial distribution, which is:

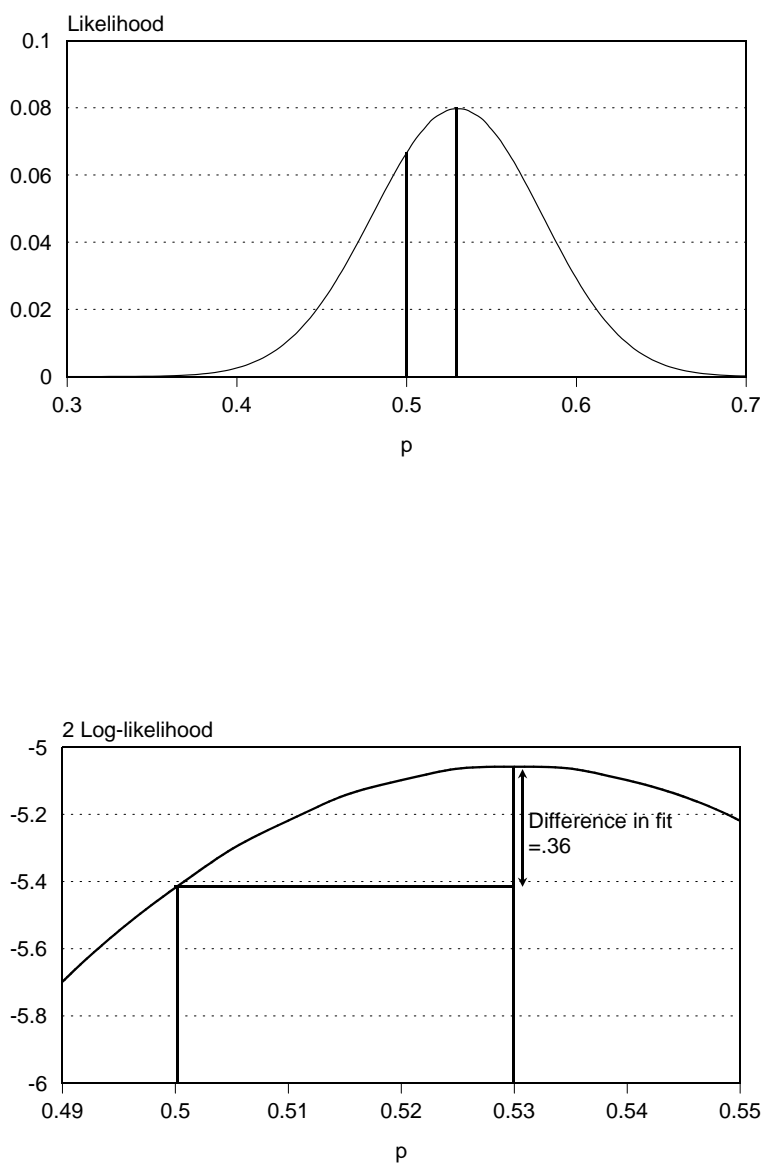
$$\frac{(h+t)!}{(h! \times t!)} \times (p)^h \times (1-p)^t. \quad (1)$$

When this formula is applied to our example, where  $h = 53, t = 47$ , and  $p = .5$  we obtain

$$\frac{100!}{(53! \times 47!)} \times .5^{53} \times (1-.5)^{47} = .07.$$

This probability statement is the usual way of viewing the outcome, or results of an experiment. If we ran a number of such experiments we would obtain a series of probabilities for the particular sequence of number of heads observed in the experiments. In contrast, with likelihood, we are interested in whether the coin is really unbiased, and wish to examine one particular set of results as a function of  $p$ , which would not be  $.5$  for a biased coin. Figure 1 shows the likelihood of obtaining 53 heads plotted for values of  $p$  from zero to one. The original probability we calculate under the assumption that  $p = .5$  appears as  $.07$  in the figure; this is also the likelihood for  $p = .5$ . We notice that the curve has a maximum (the ‘maximum likelihood estimate’ or MLE) at a value somewhat greater than  $.5$ ; in fact this is at  $.53$  corresponding to the  $53/100$ . Elementary calculus can be used to show that the MLE of a proportion is the observed proportion, but this simple relation does not hold for all maximum likelihood estimates. Often, the calculus and algebra involved is very complicated and it may defy analytic solution, so that we resort to using numerical optimization to find MLE’s. This optimization approach is used in all the structural equation modeling programs AMOS, EQS, CALIS, LISREL and Mx.

Likelihood theory concerns itself with comparisons between the height of the curve — the likelihood — at various values of parameters such as  $p$ . In Figure 1 the likelihood at the maximum is  $L_{\hat{p}=.53} = .0797332$  which can be compared with the likelihood under the hypothesis that coin is unbiased,  $L_{p=.5} = 0.0665905$ , via a likelihood ratio test. Twice the difference between the logarithm of these likelihoods is asymptotically distributed as  $\chi^2$



*Figure 1.* (Upper) Likelihood curve based on the binomial distribution for the outcome of 53 heads from 100 tosses. The likelihood varies as a function of the parameter  $p$ , the probability of obtaining heads. (Lower) Twice the log-Likelihood showing the difference ( $\chi^2$ ) between the maximum likelihood estimate of .53 and the population value of point five.

with one degree of freedom (fixing  $p$  at the prior chosen value of .5 instead of allowing it to vary as a free parameter gives one degree of freedom). In this case we have

$$2 \times (\log 0.0665905 - \log .0797332) = 0.360248$$

which is considerably below the value of 3.84 for the .05 level of significance. This difference in log-likelihoods is shown graphically in the lower part of Figure 1.

#### *Individual likelihoods.*

The binomial distribution formula in Equation 1 is probably familiar to most readers. If we think a minute about its origin, it is easy to see that it contains the likelihood for each individual coin toss. Heads occurs with probability  $p$  and there are  $h$  outcomes of this type. Because the coin tosses are independent, their probabilities may be multiplied ( $p(A \text{ and } B) = p(A)p(B)$  for independent events  $A$  and  $B$ ). Therefore, neglecting the constant term  $\frac{(h+t)!}{(h! \times t!)}$ , the contribution to the likelihood of each individual observation is either  $p$  or  $(1-p)$ . It is this most basic, *individual* level of likelihood that will be explored in this chapter.

#### *Normal Theory Maximum Likelihood*

*Univariate case.* The univariate normal distribution is characterized by two parameters, the mean,  $\mu$ , and the variance,  $\sigma^2$ . The likelihood of a particular observed score  $x_i$  is simply the height of the normal curve at that point, as shown in Figure 2. Clearly, the likelihood of the observation  $x_i$  will change as the parameters  $\mu$  and  $\sigma^2$  change. When there is only one observation, the likelihood has a maximum where  $\mu = x_i$  and  $\sigma^2 = 0$ , as the curve has infinite height under these conditions. A sample size of one will lead to problems for numerical estimation because infinity cannot be represented or manipulated by computers with finite precision. Though inconvenient, this result makes some sense because it is difficult to generalize from the particular. Typically, we have more than one observation in a sample, which is useful both scientifically and computationally. These obvious points should be born in mind later when we consider models that are different for each case in the sample.

For the univariate normal distribution the likelihood of the individual observation  $i$  is:

$$L_{U_i} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-.5(x_i - \mu)^2\}$$

and the likelihood for a sample of  $N$  independent observations is simply:

$$L_U = \prod_{i=1}^N \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-.5(x_i - \mu)^2/\sigma^2\} \right).$$

Computationally, this product is difficult to handle because each term is usually less than one in value. As the sample size increases,  $L_U$  gets so small that a computer with limited precision arithmetic cannot distinguish it from zero. Therefore, we take the logarithm of  $L_U$ , and recognizing that  $\log(A \times B) = \log A + \log B$ , we obtain:

$$\log L_U = \sum_{i=1}^N \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-.5(x_i - \mu)^2/\sigma^2\} \right). \quad (2)$$

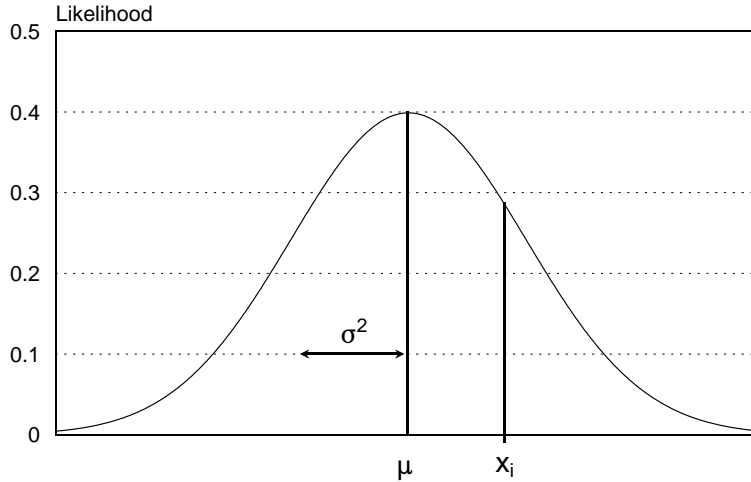


Figure 2. Illustration of the likelihood of an observation  $x_i$  under a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The height of the curve is the likelihood.

This likelihood is made up of two terms:  $1/\sqrt{2\pi\sigma^2}$ , which does not depend on the data, and  $\exp\{-.5(x_i - \mu)^2/\sigma^2\}$ , which does. The latter term is simply the square of the standardized distance, for each observation  $i$ , from the mean. We therefore have two quantities that may be of use in assessing the individual fit of each score. First, the distance measure, known as a *Mahalanobis distance*, which is asymptotically distributed as  $\chi^2$  with one degree of freedom, because it is the square of a normal deviate when the model is correct. Second, we have the contribution of each data vector  $i$  to the  $\chi^2$  fit of the model of the individual data vector, as given by Equation 2. Both these quantities may be used to detect outliers or population heterogeneity; those observations with the largest deviations fit most poorly.

*Multivariate normal distribution.* To generalize to the multivariate normal distribution is very straightforward, although it gets difficult to visualize the Mahalanobis distance in more than three-dimensional space. The algebra is entirely equivalent, however, so there is no need to try to think about hyper-spheres.

For  $k$  observed variables, the multivariate normal probability density function of an observed vector of scores  $\mathbf{x}$  is

$$L_M = |2\pi\mathbf{\Sigma}|^{-n/2} \exp\left\{-.5(\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \quad (3)$$

where  $\mathbf{\Sigma}$  is the covariance matrix and  $\mu_j$  is the (column) vector of means of the variables, and  $|\mathbf{\Sigma}|$  and  $\mathbf{\Sigma}^{-1}$  denote the determinant and inverse of the matrix  $\mathbf{\Sigma}$ , respectively. This likelihood function has become more complex in several ways, but retains the basic form of

Equation 2. There is still the simple product between a constant term  $|2\pi\Sigma|^{-n/2}$  that does not depend on the data, and a Mahalanobis distance. This time the distance measure takes into account not only the variances of the measures (on the diagonal of  $\Sigma$ ) but also the fact that the measures may covary, which affects the distance between points (see Figure 3). It is here that the non-independence of observations (several measures from a subject at one time, or repeated measures) is taken into consideration. Independent observations (different, unrelated subjects sampled from a population) each have their own likelihoods according to Equation 3, and the sum of the log of these likelihoods gives the log-likelihood for the whole sample:

$$\log L_M = N \log |2\pi\Sigma|^{-n/2} + \sum_i^N \left\{ -.5(\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) \right\}$$

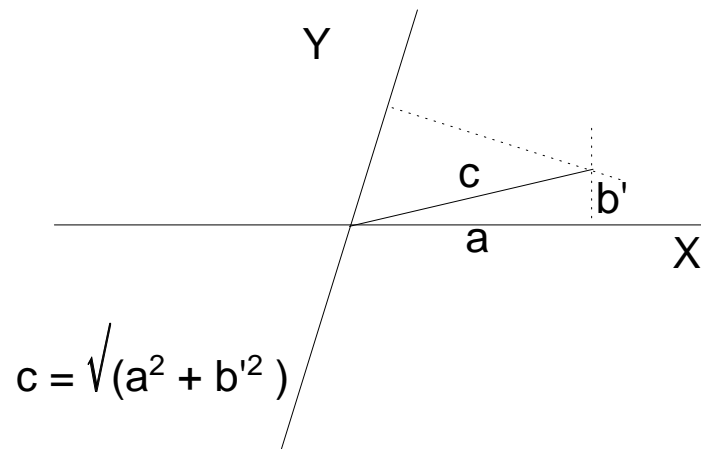
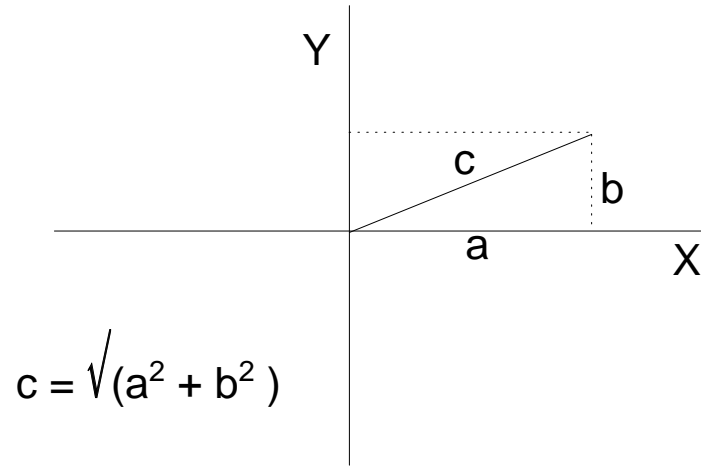
It was Karl Jöreskog's brilliant work (Jöreskog, 1967, 1970) that reduced the above term for the log-likelihood of a set observations to the formulae used to fit models to covariance matrices (see Mardia et al (1979) p. 98 for a succinct account). At that time computers were slow so the great increase in computational speed from using summary statistics was essential to make structural equation modeling practical. Today even palm-top computers exceed the speed of the fastest machines available at that time — machines that filled large rooms. This increase in computational power makes ML analysis of raw data very practical.

### Missing Data

A remarkable property of the individual likelihood method is that it offers a simple but powerful treatment of missing data. The method that is implemented in Mx does not use any imputation, it merely calculates the likelihood of those observations that are present. That is, for any particular vector of observed scores the estimated covariance matrix  $\Sigma$  in Equation 3 is *filtered* to contain only those rows and columns corresponding to the variables that were observed. Likewise,  $\mu$ , the column vector of estimated means, is filtered to contain only the means corresponding to the variables that have been observed. The overall likelihood of a set of data is then made up of the product of individual likelihoods that are based on different numbers of observed variables. The earliest description of this method to my knowledge is that given by Lange et al. (1976), where the application was directed at unbalanced pedigrees; families differ in the number of children, which is a special form of missing data.

When the data are missing due to factors completely independent of any of the observed measures, they are said to be *missing completely at random* (MCAR; Little & Rubin 1987). Correct maximum likelihood estimates will be obtained. It is relatively easy to see, intuitively, why this is so. Suppose we decided to make either one or two measures on a sample, and randomly assigned subjects to being measured on just  $X$  (group 1) or  $X$  and  $Y$  (group 2). One would not expect any difference between the mean and variance of  $X$  in the two groups. Formally, the conditional likelihood given the missingness status is unchanged from the original likelihood because of the independence of the missingness.

A more complicated situation is where random missingness of the MCAR variety is augmented by missingness that is entirely predicted by variables that are not missing. A



*Figure 3.* Illustration of the Mahalanobis distance in two dimensional space. For two uncorrelated dimensions (upper figure), the Mahalanobis distance is the sum of the squared deviations of the two variables from their respective means, or  $c^2 = \sqrt{a^2 + b^2}$  in the diagram. This distance changes to  $\sqrt{a^2 + b'^2}$  when the dimensions correlate, as shown with oblique axes in the lower figure.

practical example of this might be a study in which one measure,  $X$ , is either available for all subjects or missing completely at random. Those subjects with scores above a certain cutoff on  $X$  are selected for a further measurement  $Y$ . In this case, known as *missing at random* (MAR) maximum likelihood estimates will also be unbiased, but it is not very easy to see intuitively why this approach works. Indeed, the sample mean and variance of  $X$  will be quite different for the two groups, and fitting a two-group model with means and variances equated across the groups to the data would fail, given sufficient sample size. However, at the raw likelihood level there is no separation into different groups. The missingness of  $Y$  is completely predicted by a known, measured variable  $X$ , so the conditional distribution of  $Y$  given the value of  $X$  does not contain missingness due to  $Y$  itself. It is when the missingness is associated to a part of the variance not explained by other variables in the model (the “residual” variance) that we say the data are *not missing at random* (NMAR). NMAR data will give biased estimates with the basic raw data ML method, but it may be possible to model the missingness to control for its effects.

The consequence of missing data is that the individual fit statistics have different distributions according to the number of non-missing data points for an individual case. This is because the distances of a set of vectors of  $k$  variables is asymptotically distributed as  $\chi^2$  with  $k$  degrees of freedom. On average, the larger the number of variables for which the likelihood Equation 3 is computed, the larger the value of the Mahalanobis distance, and the larger the contribution to the fit function. Simply identifying the cases with the largest Mahalanobis distances or the largest contributions to the fit function would preferentially select the cases that contained the fewest variables missing. This problem was noted by Hopper & Mathews (1983), who describe two formulae provided by Johnson & Kotz (1970) that may be used to obtain an approximate z-score for the individual fit statistic. These are:

$$Q_i^{(1)} = 2(Q_i)^{.5} - (2n_i - 1)^{.5} \quad (4)$$

$$Q_i^{(2)} = ((Q_i/n_i)1/3 - 1 + 2/(9n_i))(9n_i/2)^{.5} \quad (5)$$

where  $Q_i = (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu)$  is the Mahalanobis distance for subject  $i$ . Early versions of Mx printed out  $Q_i^{(1)}$  because it was simpler to compute. However, following a small simulation study indicated that  $Q_i^{(2)}$  was superior. Admittedly the evidence to date is very flimsy, and a proper simulation study should be done using a variety of distributions of vector lengths. Possibly, a third, superior statistic could be developed that would be superior to both  $Q_i^{(1)}$  and  $Q_i^{(2)}$ . The simulation study involved generating 1000 pairs of scores from a bivariate normal distribution with correlation point five. A saturated model of three free covariance parameters and two free means was fitted to the data, and the two  $Q$  statistics were computed for each observation in the sample. Half normal probability plots for the two statistics are shown in Figure 4. As can be seen, the plot for  $Q_i^{(2)}$  follows the normal distribution much more closely in the tails than does  $Q_i^{(1)}$ , though both are a good approximation in the middle of the range. For now at least, Mx writes  $Q_i^{(2)}$  to a file when `Option Mx\%p=<filename>` is specified and raw data are being analyzed. This file also contains other useful information, namely: (i) the contribution to the likelihood function for that vector of observations; (ii) the unsigned square root of the Mahalanobis distance,  $Q_i$  (iii) the estimated z-score  $Q_i^{(2)}$ ; (iv) the number of the observation in the active (i.e. post

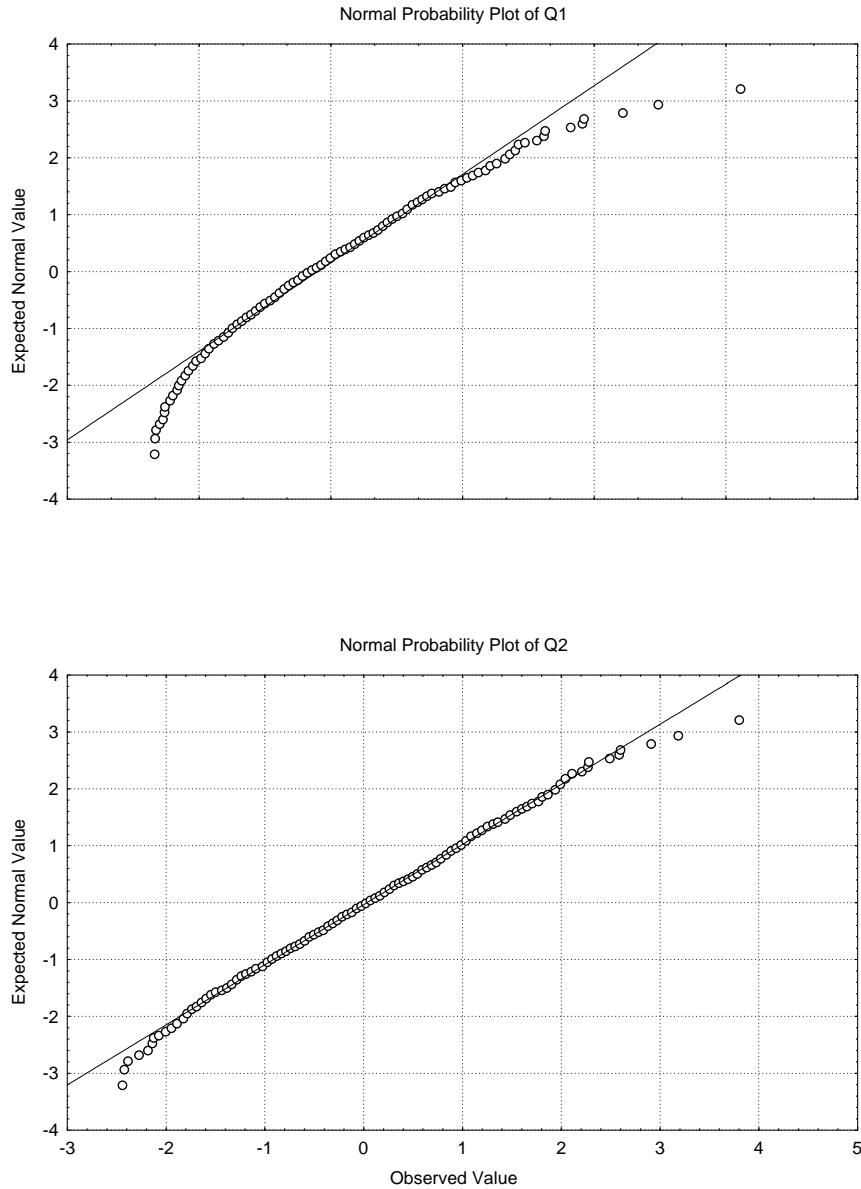


Figure 4. Half normal plots of  $Q_i^{(1)}$  and  $Q_i^{(2)}$  for bivariate standard normal data with correlation .5;  $Q_i^{(2)}$  is closer to the theoretical distribution described by the straight line.

selection) dataset; (v)  $n_i$  the number of data points in the vector; (vi) the number of times that the loglikelihood of this vector was found to be incalculable during optimization; (vii) a flag coded 000 if the likelihood could be evaluated at the solution, or 999 otherwise; and (viii) the model number if there are multiple models requested with the `NModel` argument for mixture distributions (see below). This information may be read into programs such as SAS for sorting and to help the outliers with large  $\chi^2$  or Z-scores.

### *Implementation in Mx*

Structural equation modeling of raw data with Mx is relatively straightforward, but there are some considerations that make it less simple than modeling summary statistics. First, it is necessary to model *both* the means and the covariances. The maximum likelihood function in Equation 3 has parameters for both the means  $\mu$  and the covariances  $\Sigma$ . Usually, we wish to parameterize  $\Sigma$  in terms of the parameters of a structural equation model, and sometimes the mean vector  $\mu$  also will be modeled. In Mx, structural equation models for means and covariances may be modeled using the graphical user interface (GUI) which is available free of charge at the website <http://www.vipbg.vcu.edu/mxgui>. Preparation of the raw data for such a model is relatively straightforward, involving the creation of a ‘.dat’ file which contains all the information concerning the data: the number of variables, their names, and the raw data or the location of the raw data file. For example, the following lines:

```
! Comment - example .dat file for raw data
Data Ninput=4 Nobservations=0
Labels Age Sex Verbal Performance
Rectangular File=verb.rec
```

would identify the source file for the raw data as `verb.rec`. By specifying `rectangular`, the program expects a file in which each case is on a single line, with variables separated by spaces. The default code for missing values is a dot ‘.’ but other codes may be selected with the `Missing` command. Example lines of the file `verb.rec` might therefore read:

```
55 1 103.78 115.40
48 0 112.96 105.33
49 1 98.77 .
. 0 125.86 105.21
```

The first two cases contain complete data, the third has missing data on the fourth (Performance) variable, and the fourth case has missing data on age. Having prepared the .dat file it is a relatively simple matter to draw a diagram to fit the structural equation model for means and covariances with the Mx GUI. Complete instructions are given in the documentation available on the website.

One additional consideration with the modeling of raw data is the use of starting values. Mx, like all model-fitting programs, is sensitive to starting values and may not find the global minimum from bad initial estimates (although it seems better than most in this regard; Hamagami, 1997). This problem is exacerbated when raw data are used, because at bad starting values, especially for the predicted means, the likelihood may evaluate to zero because of the limited numerical precision of computers. Whenever Mx encounters such

a case at the starting values, the user is informed of the problem, and the observed data vector, the predicted mean vector and the standardized deviations of the observed scores from the predicted mean are tabulated. It is up to the user to find a set of starting values that reduces these distances to smallish quantities, perhaps less than two for all variables. Sometimes the deviations of the scores from the means are suitably small but there is still a problem evaluating the likelihood, perhaps because the starting values predict high correlations between the variables, but the observed scores do not follow this correlated pattern of deviations from the means. To avoid this type of problem, it may be best to supply starting values that generate a predicted covariance matrix that is almost diagonal.

### Continuous Moderator Variables

The term moderator is usually applied when a variable changes the relationship between two other variables. For example, a difference in correlation between height and weight for men and women might be termed a moderating effect of sex on the height-weight correlation. These simple effects of binary variables are relatively straightforward to model using regular multiple group SEM. For a comprehensive review of this area, see Schumaker (1998). Here we focus our attention on continuous moderating variables.

As Muthén (1989) pointed out, the problem of moderators becomes more difficult when there are several binary variables to be considered. The group sizes diminish rapidly, and adequate sample sizes for stable parameter estimates become difficult to achieve. Muthén recommends the use of MIMIC models in this case, allowing the group membership variables to be causes of both latent variables and observed variables, thereby moderating the relationship between latent and observed variable as a function of group membership. The MIMIC method is indeed a valuable approach, especially in the context of simple factor models where the relationship between the factor and the observed measures is of interest. Especially useful in the context of analyzing summary statistics is the fact that the sample sizes are not reduced by partitioning into many different groups. Analysis of the raw data offers another way around this problem. In addition it allows us to model moderating relationships directly into any part of a model.

To describe Mx's definition variable approach to continuous moderating variables, we can use the simple example of bivariate regression with interaction, where the dependent variable  $Y$  is a function of two independent variables  $X$  and  $Z$  and their product  $XZ$ . The model is written:

$$Y = b_1X + b_2Z + b_3XZ + e \quad (6)$$

where  $b_i$  are regression coefficients, and  $e$  is a residual error term. This unimaginative example may seem boring, and compared to interactions between latent variables and real-world examples it certainly is, but it serves to illustrate the main statistical points very nicely. First, we should recognize that the interaction term  $XZ$  can be regarded as a form of moderation. We might say that the effect of  $Z$  on  $Y$  depends on the value of  $X$ . A simple rearrangement of Equation 6 yields:

$$Y = b_1X + (b_2 + b_3X)Z + e \quad (7)$$

or equivalently we might describe a moderating effect of the value of  $Z$  on the regression of  $Y$  on  $X$ :

$$Y = (b_1 + b_3Z)X + b_2Z + e. \quad (8)$$

While bivariate regression with interaction is a solved problem, its restatement as a structural equation model is not without difficulties. If we took the strategy of computing  $XZ$  and adding it to the dataset, the first problem would be that if  $X$  and  $Z$  were normally distributed then the product variable  $XZ$  would not. In addition, if  $b_3$  were non-zero, the dependent variable  $Y$  would be non-normal. Goodness-of-fit tests and significance tests on the ML estimates of the parameters would be adversely affected by the non-normal distributions of the variables. Therefore, there is some advantage to recasting the model in the form of Equation 7 or 8 as the non-normal variable  $XZ$  is no longer a part of the data to be analyzed but simply a moderator of the relationship between  $X$  (or  $Z$ ) and  $Y$ . Furthermore, for any particular value of  $Z$ , the distribution of  $Y$  is conditionally normal under this model. Since we will be using the raw maximum likelihood fitting function instead of summary statistics, the normality assumption (conditional on  $X$ ) will be upheld. Better statistical properties of the model should result.

To some extent, interpretation of individual fit becomes more difficult with this approach, because each subject potentially has a different predicted covariance matrix. An outlier might arise either because of extreme values of the observed data, or because of extreme values of the moderators, or because of an unusual combination thereof.

#### *Implementation in Mx*

It is possible to use either the diagram drawing software or the script language to devise models with continuous moderator variables. For the bivariate regression example, an Mx diagram is shown in Figure 5. The diagram, script and data may be downloaded from the Mx website <http://griffin.vcu.edu/mx> and following the links to examples and moderated regression. There are several noteworthy features of the path diagram in Figure 5. First, there is the use of variance arrows — double-headed paths from a variable to itself. This is a graphical tool that makes a 1:1 relationship between the graphical and the algebraic representations of the model (McArdle & Boker, 1990). Second, the use of triangles may be unfamiliar. Triangles allow the modeling of means via simple tracing rules or simple matrix algebra, as described in the documentation for the Mx graphical interface available at <http://vipbg.vcu.edu/mxgui>. Third, the most unusual feature of the diagram is the presence of the variable  $Zd$  inside a diamond on a path. By mapping variables to paths, we are specifying a model for raw data such that the individual values of  $Zd$  change with each subject in the dataset. The diamond graphically flags the presence of this definition variable in the model. In this case,  $Zd$  and  $Z$  are identical, requiring two columns of identical data in the raw data file.  $Z$  is used to model the main effect of  $Z$  on  $Y$  and  $Zd$  is used to model the moderating effect that  $Z$  has on the regression of  $Y$  on  $X$ . A small amount of simulation indicates good recovery of parameter estimates using this method. More comprehensive testing of the method would be relatively simple to automate and should be done to establish whether the modeling of interaction effects this way has superior statistical properties.

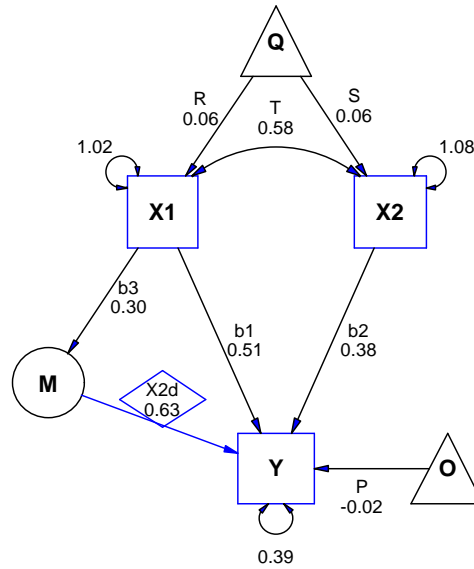


Figure 5. Path diagram of moderating effect of  $Z$  on  $X$  such that bivariate regression of  $Y$  on  $X$ ,  $Z$  and the product  $XZ$  is modeled.

### Finite Mixture Distributions

A frequently challenged assumption of structural equation modeling is that it assumes the same model for all subjects in the sample. This assumption is indeed worth examining, because it seems highly unlikely that the same relationships between variables hold for all subjects unless the population is very homogeneous. As Muthèn (1989) has shown, parameter estimates may be substantially biased when the population contains a mixture of subpopulations, although the goodness-of-fit of the model may be perfectly satisfactory. One solution to this problem, to be considered in this section, is explicit modeling of mixture distributions via maximum likelihood.

When a population consists of a limited number of sub-populations, e.g., two sub-populations with different means, the problem is known as a *finite mixture distribution*. Whole volumes have been written on the advantages, disadvantages and practicalities of modeling mixture distributions (Everitt & Hand, 1981; Jedidi, Jagpal, & DeSarbo, 1997) and would serve little purpose to reiterate these excellent works here. The present aims are simply to describe how Mx may be used to fit models of mixture distributions via maximum likelihood, and to consider briefly some of the problems that may arise.

First, we should consider standard methods for dealing with such mixtures. One extremely powerful method is multiple group structural equation modeling. The usual way of handling heterogeneity in means or covariance structure between two groups is to subdivide the population and to fit models to both groups simultaneously. This approach

has the advantage of being able to statistically test for heterogeneity. For example, Neale & Cardon (1992 Ch.11) tested for sex differences in the parameters of a simple model of genetic and environmental effects on Body Mass Index (BMI). Twin pairs were subdivided into male and female groups and the fit of a model where the parameters were constrained to be equal across sexes was compared with that of a model where the parameters were allowed to differ. A significant improvement in fit was observed when the parameters differed between the sexes. Clearly, splitting the data into groups according to sex is important in the genetic analysis of Body Mass Index (Maes, Neale, & Eaves, 1997).

Wherever possible, tests of heterogeneity should be performed and data from potentially heterogeneous groups should not be combined without first testing for group differences. In the BMI example, such tests would be impossible to perform if the twins' sex had not been assessed. While most studies of humans record the sex of the subjects, some do not. Furthermore, there may be other variables that have not been measured, or that have not been measured accurately, that index population heterogeneity. It is in the treatment of this unmeasured or *latent heterogeneity* that mixture distributions can prove useful.

#### *Normal Likelihood for Mixture Distributions*

It is very simple to write the likelihood for a mixture of multivariate normal distributions. From Equation 3 we can write a weighted sum of likelihoods for  $r$  sub-populations as:

$$L_M = \sum_{d=1}^r w_d |2\pi\Sigma_{\mathbf{d}}|^{-n/2} \exp \left\{ -.5(\mathbf{x} - \mu_{\mathbf{d}})' \Sigma_{\mathbf{d}}^{-1} (\mathbf{x} - \mu_{\mathbf{d}}) \right\}. \quad (9)$$

There are several things to note about this formulation. First, the data vector  $\mathbf{x}$  remains invariant across the different models. This is one place where numerical evaluation of the likelihood of a mixture distribution may run into difficulties, because if the two populations are very different, evaluating the likelihood under the model for the alternate population may yield a value indistinguishable (to a computer) from zero. Taking the logarithm of this likelihood would yield  $-\infty$  and would outweigh all the other data. Attention to starting values is even more important here. In general, the best place to start is where the models slightly differ in their predicted covariances. No difference at all will impede optimization as the derivatives of the mixing proportion parameters cannot be estimated. Too large a difference may lead to incalculable likelihoods for some of the data points.

Second, the  $d = 1 \dots r$  likelihoods of the mixture are weighted by  $w_d$ , often characterized as a mixing proportion, with  $\sum w_d = 1$ . These weights may be measured at the population level, for example, we may fix the weights to be a .25 and .75 in a two component mixture, or they may be estimated at the population level as free parameters. Perhaps more interesting is the possibility that the weights can be measured at the individual level, so that each subject has a different set of group membership probabilities. This formulation encompasses standard multiple group structural equation models as a special case; for the sex differences example the two components would be the parameters to model the covariance matrices  $\Sigma_{\mathbf{d}}$  and the mean vectors  $\mu_{\mathbf{d}}$ . The weights would be the probabilities that each subject is male,  $p(\text{male})$  and  $p(\text{female})$ , and expressed as a vector would be either  $\mathbf{w}' = (1, 0)$  or  $\mathbf{w}' = (0, 1)$ . Subjects of unknown sex would be easy to add to this study,

with weights  $\mathbf{w}' = (.5, .5)$  or some other prior probabilities appropriate for the age and other characteristics of the sample.

Third, there are now several likelihoods being evaluated. If we wish to assess the individual fit then the likelihoods for the component models must all be taken into consideration. Corresponding to the single distribution case, the contributions to the overall fit (the individual likelihood), the Mahalanobis distance, and the z-score  $Q^{(2)}$  may be computed for each component distribution. That is, a particular subject may be an outlier in any or all of the sub-populations. With strong separation between subpopulations (imagine two univariate normal distributions with a mean difference of 6 standard units) almost all points would be outliers in one of the subpopulation distributions. Taken another way, these statistics may inform us of the probability that a given subject belongs to a particular sub-population. One simple function to compute the posterior probability of belonging to sub-population  $k$  is  $L_k / (\sum_{d=1}^r L_d)$ .

Fourth, we note that there are some statistical problems with comparing the likelihood of a mixture distribution with that of a sub-model in which there is no mixture. There are two ways in which the sub-model might be represented: (i) the parameters for the mean and covariance structures of the sub-populations are constrained to be equal and the mixing proportion is fixed at an arbitrary value; and (ii) the mixing proportion parameter is fixed at 1 for one component and at zero for the rest. This latter case is really a *boundary condition* because the mixing proportions are constrained to lie in the interval from 0 to 1. As a result, tests of the difference between a model with a mixture and one without a mixture are not asymptotically distributed as  $\chi^2$ . This problem has been studied quite extensively by Dijkstra (1992), Self & Liang (1987) and Shapiro (1988) among others. All indicate the problem of the fit statistic having a distribution under the null hypothesis that takes the value zero some proportion of the time, and is distributed as  $\chi^2$  otherwise. Some resolution to this problem may be afforded by resort to bootstrapping (Schork, Allison, & Thiel, 1996) although these procedures are not yet implemented in Mx. Another problem is that with small sample sizes or poor separation between the components of the mixture, there is a chance that optimization will converge to a local rather than a global maximum likelihood (Hosmer, 1974). Exactly how sensitive the NPSOL optimization routines (Gill, Murray, Saunders, & Wright, 1986) used by Mx (Neale, 1997) are to starting values, sample sizes and separation of components within mixture distributions is a matter for extensive simulation. Experience to date suggests that they are fairly robust, but this should be investigated thoroughly.

#### *Implementation in Mx*

Built into Mx is an extension of the maximum likelihood fitting function that allows specification of a finite mixture distribution model. In the single distribution case, the user is required to supply a matrix formula for the predicted covariances that yields a matrix of order  $m \times m$  and a formula for the predicted means that yields a vector of order  $1 \times m$ . For a mixture of  $d$  distributions, it is necessary to first tell Mx that there will be more than one model, which is done with the a parameter to the data line, e.g., `NModel=3` for the case of  $d = 3$ . Second, all  $d$  models must be supplied in a single matrix algebra formula for the covariances, stacked on top of each other in a vector of covariance matrices, being of order  $md \times m$ . Similarly, the predicted mean vectors must be stacked to create a matrix of means

of order  $d \times m$ . Finally, it is necessary to supply a matrix formula for the weights that will evaluate to a  $d \times 1$  vector corresponding to the  $d$  models. Because Mx has a general matrix algebra interpreter, it is easy to configure the covariance matrices and mean vectors in this way, using the vertical concatenation operator.

The general case in which the weights vary according to the individual subject may be implemented using the “Definition variable” approach. A definition variable is used to define the model; once a variable is declared to be a definition variable, it is no longer part of the active set of variables to be analyzed. Here I briefly illustrate the use of mixture distribution methods in the context of data collected from studies of genetic linkage. The mixture distribution methods available in Mx were specifically developed to tackle this problem (Eaves, Neale, & Maes, 1996), but it is my hope that they will prove useful in a much broader context.

*Example: Genetic marker data.* Today there is growing interest in the linkage analysis of quantitative traits. This is not the place for a detailed description of the methods in this area, but for a clear exposition on genetic linkage and concepts such as identity by descent and identity by state the reader should consult Sham (1997). Linkage studies use genetic markers to assess directly the genetic similarity of family members. Humans are diploid; they have pairs of chromosomes, one inherited from each parent. At any particular place on the genome or ‘locus’, an individual’s genotype may be expressed as the pair of alleles that they have at that locus. Supposing that we have two parents with entirely different alleles at a particular locus (father=AB, mother=CD), we can describe the genotypes of all the possible offspring that they might create, namely AC, AD, BC and BD. For any particular pair of sibs we can then simply count the number of alleles that they share in common. This type of allele sharing is known as identity by descent (IBD), because it measures whether the alleles originated from the same parental strand of DNA. It is useful because we want to find out if any nearby genes nearby are having an effect on sibling similarity for a trait of interest.

Table 1 shows the number of alleles shared IBD for all the possible configurations of two siblings in a family. For example, if sibling 1 inherits AC and sibling 2 inherits BC, they share one allele identical by descent. Clearly, a population of sibling pairs will consist of a mixture of three types: those sharing 2, 1 or 0 alleles IBD. If a locus has an effect on a trait, then we would expect siblings that share two alleles IBD to be more similar than siblings that share one allele, who in turn would be more similar than siblings that share zero alleles ( $r(IBD2) > r(IBD1) > r(IBD0)$ ). Places on the genome that affect quantitative traits are usually called a quantitative trait loci or QTL.

Usually, identity by descent at the quantitative trait locus itself is not measured, or has a limited number of alleles, so we cannot precisely distinguish sibling pairs into the three types. Instead, we can compute the probability that a particular sibling pair belongs to each type. It is these probabilities, which may be computed by a program such as Mapmaker/sibs (Kruglyak & Lander, 1995), that form the weights for the different models. In practice, these weights may be different for each sibling pair in the sample.

If we knew with certainty whether each pair was IBD2, IBD1 or IBD0, it would be a relatively straightforward case of a three group structural equation model, sometimes known as “Partitioned Twin Analysis” (Nance & Neale, 1989). However, because the genetic data

Table 1: Number of alleles shared identical by descent in all possible pairs of siblings. Parental genotypes are AB and CD, giving siblings of types AC, AD, BC and BD.

Sibling 2	Sibling 1			
	AC	AD	BC	BD
AC	2	1	1	0
AD	1	2	0	1
BC	1	0	2	1
BD	0	1	1	2

are not fully informative, we can only assign probabilities that pairs are of each type, which is where the mixture model comes in.

The model for the resemblance of siblings while allowing for the effects of a QTL is quite simple and is shown in Figure 6. The latent variables are the QTL effect  $Q$ , residual shared (non-QTL genetic and environmental) factors  $C$ , and residual individual-specific factors  $E$  which include measurement error. Key to the identification of this model is the fact that the correlations between the latent variables are fixed. In the model for siblings sharing zero genes IBD at the QTL, the correlation between  $Q1$  and  $Q2$  is fixed at zero; for the models of 1 and 2 genes IBD it is fixed at .5 and 1.0 respectively.

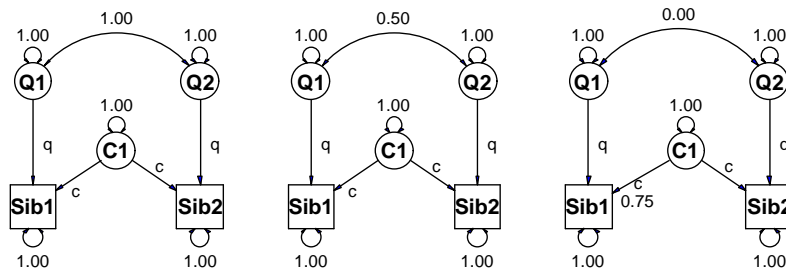


Figure 6. Three models for the covariance of siblings, according to whether they share 2, 1 or 0 alleles identical by descent at a locus. The likelihood of a pair of scores under these three models is weighted and summed by the probability that the pair is that particular type, as computed from genetic marker data.

The test for linkage to a locus is simply a difference in fit (the likelihood ratio test described above) between the model that includes vs. excludes an effect of  $Q$  on the phenotypes. Typically, these tests are repeated many times across a region of interest or even the whole genome, which leads to potential statistical problems from multiple testing. Repeated testing leads to apparently improbable results even when there is no true effect. To account for this, tests of linkage usually involve a high significance level of  $p = .00074$  that corresponds to a  $\chi_1^2$  difference of 13.82 (a lod score of 3.0). This happens to correspond to a value based on a Bayesian argument that takes into account the number of chromosomes

in man. Both these arguments are clearly described in Sham (1997).

Mx scripts to fit this model are available on the website <http://griffin.vcu.edu/mx+> via the examples and QTL links. At this time, it is not possible to use the graphical interface to generate models with mixture distributions. If the model is complex and facility with the Mx script language is not to hand, some part of the work could be done by generating an intermediate script using the graphical interface to build the models and then changing the script to add a data group with the mixture, and using matrix equality constraints to obtain the predicted means and covariances from the previous groups. We hope to offer graphical modeling of mixture distributions in version 2.0 of the Mx GUI.

### References

- Dijkstra, T. (1992). On statistical inference with parameter estimates on the boundary of the parameter space *British Journal of Mathematical and Statistical Psychology*, *45*, 289–309.
- Eaves, L. J., Neale, M. C., & Maes, H. H. (1996). Multivariate multipoint linkage analysis of quantitative trait loci *Behavior Genetics*, *26*, 519–526.
- Everitt, B. S. & Hand, D. J. (1981). *Finite Mixture Distributions*. Chapman and Hall.
- Gill, P. E., Murray, W., Saunders, M. A., & Wright, M. H. (1986). User's guide for npsol (version 4.0): A FORTRAN package for nonlinear programming Tech. Rep. SOL 86-2, Department of Operations Research, Stanford University, Stanford.
- Hamagami, F. (1997). A review of the mx computer program for structural equation modeling *Structural Equation Modeling*, *4*(2), 157–175.
- Hopper, J. L. & Mathews, J. D. (1983). Extensions to multivariate normal models for pedigree analysis. II. Modeling the effect of shared environments in the analysis of variation in blood lead levels, Vol. 117.
- Hosmer, D. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines *Communications in Statistics*, *3*, 995–1006.
- Jedidi, K., Jagpal, H., & DeSarbo, W. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity *Marketing Science*, *16*(1), 39–59.
- Johnson, N. I. & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions*. 1. Boston: Houghton Mifflin.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis *Psychometrika*, *32*, 443–482.
- Jöreskog, K. G. (1970). Estimation and testing of simplex models *British Journal of Mathematical and Statistical Psychology*, *23*, 121–145.
- Kruglyak, L. & Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits *American Journal of Human Genetics*, *57*, 439–454.
- Lange, K., Westlake, J., & Spence, M. A. (1976). Extensions to pedigree analysis: III. Variance components by the scoring method *Annals of Human Genetics*, *39*, 485–491.
- Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley and Son.
- Maes, H., Neale, M., & Eaves, L. (1997). Genetic and environmental factors in relative body weight and human adiposity *Behavior Genetics*, *27*, 325–351.

- Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. New York: Academic Press.
- McArdle, J. J. & Boker, S. M. (1990). *RAMpath* path diagram software. Denver, CO: Data Transforms Inc.
- Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations *Psychometrika*, *54*, 557–585.
- Nance, W. E. & Neale, M. C. (1989). Partitioned twin analysis: A power study. *Behavior Genetics*, *19*, 143–150.
- Neale, M. C. (1997). *Mx: Statistical Modeling* (4th ed.).
- Neale, M. C. & Cardon, L. R. (1992). *Methodology for Genetic Studies of Twins and Families*. Kluwer Academic Publishers.
- Schork, N., Allison, D., & Thiel, B. (1996). Mixture distributions in human genetics research *Statistical Methods in Medical Research*, *5*, 155–178.
- Schumacker, R. & Marcoulides, G. (1998). *Interaction and Nonlinear Effects in Structural Equation Modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Self, S. G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions *Journal of the American Statistical Association*, *82*, 605–610.
- Sham, P. (1997). *Statistics in Human Genetics*. New York: John Wiley and Sons.
- Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis *International Statistical Review*, *56*, 49–62.