PSYCHOMETRIKA—VOL. 49, NO. 3, 359–381
SEPTEMBER 1984

ESTIMATING LATENT DISTRIBUTIONS

ROBERT J. MISLEVY

NATIONAL OPINION RESEARCH CENTER

Consider vectors of item responses obtained from a sample of subjects from a population in which ability $\boldsymbol{\theta}$ is distributed with density $g(\boldsymbol{\theta} | \boldsymbol{\alpha})$, where the $\boldsymbol{\alpha}$ are unknown parameters. Assuming the responses depend on $\boldsymbol{\theta}$ through a fully specified item response model, this paper presents maximum likelihood equations for the estimation of the population parameters directly from the observed responses; i.e., without estimating an ability parameter for each subject. Also provided are asymptotic standard errors and tests of fit, computing approximations, and details of four special cases: a non-parametric approximation, a normal solution, a resolution of normal components, and a beta-binomial solution.

Key words: latent distributions, maximum likelihood, EM algorithm, empirical Bayes estimation, Gaussian resolution, beta-binomial model, item response theory

Introduction

The familiar statistical procedures in the armamentarium of the applied educational or psychological researcher, including analysis of variance and regression, assume independent and identically distributed (iid) error terms over observations. Virtually every application of such procedures to mental test scores violates this assumption to some degree. The estimation errors of item response theoretic ability estimates, for example, vary with the numbers and the parameters of the items an individual has been presented, as well as with the resultant response pattern. And while classical true-score test theory posits iid errors for all scores on a given test, this assumption is usually untenable in practice because of "floor" and "ceiling" effects; as Lord and Novick (1968) point out, "... under any model with bounded observed score and unbiased errors (not all zero), the conditional distribution of the observed score cannot be independent of true score."

The consequences of these violations are mitigated as the numbers of items administered to all subjects increase, and the magnitude of estimation errors decreases accordingly. Serious problems are encountered, however, when the number of responses per subject is limited, either by the nature of a test instrument (a checklist of depression symptoms may consist of only ten items) or the design of a study (for a given number of item responses, multiple-matrix sampling designs yield increasingly precise estimates of the population mean as the number of items administered to each subject decreases toward one). The distribution of *estimates* of individual subjects' parameters may then depart radically from the distribution of the parameters themselves, thereby invalidating any analyses that would treat the estimates as if they were the parameters they represent.

An appropriate remedy is to base inferences about population-level parameters directly upon the basic data rather than upon derived data. Sörbom (1974), for example, estimates differences in factor means between groups by maximizing a likelihood equation that contains the differences as parameters, rather than by performing ANOVA on subject-level factor score estimates. Bock and Aitkin (1981) estimate item parameters by

The author would like to thank R. Darrell Bock for his comments, suggestions, and encouragement during the course of this work.

Requests for reprints should be addressed to Dr. Robert Mislevy, Educational Testing Service, Princeton, N.J. 08540

0033-3123/84/0900-5081$00.75/0

the method of marginal maximum likelihood, wherein integration over an ability distribution during item-trait regression analyses replaces summation over ability estimates for each subject. This paper adopts a similar approach to the problem of characterizing the distribution of a latent variable directly from observed response patterns.

The problem of estimating a latent population distribution has been addressed by several workers in recent years. Lord (1969) notes that only the first n moments of the density function $g(\theta)$ can be estimated from scores on an n-item test, and suggests a mathematical procedure for calculating said moments. Andersen and Madsen (1977) and Sanathanan and Blumenthal (1978), working with the one parameter logistic (Rasch) item response model, assume that g belongs to a specified parametric family and provide marginal maximum likelihood equations for estimating its parameters. All of these presentations required all subjects to respond to the same set of items.

Marginal maximum likelihood (MML) estimation is readily extended to other item response models, to general incidence sampling designs, and to multiple dimensions. The following sections develop likelihood equations for the general MML solution and for four special cases: a non-parametric approximation of g, a normal solution, a resolution of mixed normal components, and the beta-binominal model. We begin by reviewing what Dempster, Laird, and Rubin (1977) refer to as the "complete data" problem, in which values of $\boldsymbol{\theta}$ are observed directly.

The "Complete Data" Solution

Suppose that $g(\boldsymbol{\theta})$ is a member of a family of densities characterized by the population parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_s)$. If g is normal, for example, its parameters are the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The probability of a sample of N observations of $\boldsymbol{\theta}$ is given as

$$L = \prod_{i}^{N} g(\boldsymbol{\theta}_i). \tag{1}$$

To estimate $\boldsymbol{\alpha}$ from a sample of observations by the method of maximum likelihood, we consider (1) a function of $\boldsymbol{\alpha}$ for the given data, and find the maximizing values. In practice the log likelihood

$$\log L = \sum_{i} \log g(\boldsymbol{\theta}_i) \tag{2}$$

is maximized. This is done by differentiating (2) with respect to each α_s in turn, and equating the results to zero:

$$0 = \frac{\partial \log L}{\partial \alpha_s} = \sum_{i} g^{-1}(\boldsymbol{\theta}_i) \frac{\partial g(\boldsymbol{\theta}_i)}{\partial \alpha_s}. \tag{3}$$

Solutions of these likelihood equations yield saddle points or relative extrema of the likelihood function. Examination of the likelihood surface in the neighborhood of a solution indicates whether a relative maximum has been found.

For large N, the maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$ follows a multivariate normal distribution with mean $\boldsymbol{\alpha}$ and covariances given by the inverse of the Fisher information matrix

$$E\left[\left(\frac{\partial \log L}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial \log L}{\partial \boldsymbol{\alpha}'}\right)\right].$$

Assuming sampled $\boldsymbol{\theta}$'s to be independent and identically distributed, we may approximate this quantity by

$$N^{-1} \sum_{i} \left(\frac{\partial \log g(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial \log g(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\alpha}'}\right).$$

The "Incomplete Data" Solution

Suppose that observations consist not of values of $\boldsymbol{\theta}$ directly, but of vectors of item responses $\mathbf{x}$ that depend upon $\boldsymbol{\theta}$ stochastically. The response of subject i to item j, namely x_{ij}, is assumed to be a function of $\boldsymbol{\theta}_i$, his ability, and ζ_j, one or more parameters for item j, as follows:

$$P(x_{ij} = k \mid \boldsymbol{\theta}_i, \zeta_j) = f(k, \boldsymbol{\theta}_i, \zeta_j).$$

The response categories and associated response functions may be dichotomous, polytomous, countably infinite, or continuous.

As in Andersen and Madsen (1977) and Sanathan and Blumenthal (1978) we assume both the form of the item response model and the values of the item parameters to be known even though estimates must be used in practice, obtained either from a previous calibration sample or from the same data from which the population parameters will be estimated. When population parameters have been estimated independently from population parameters—either from a separate sample of responses or as conditional estimates under a Rasch item response model as in Andersen and Madsen (1977) and Sanathan and Blumenthal (1978)—adverse effects of this first stage of estimation upon the estimation of population parameters and the calculation of standard errors and tests of fit will diminish as the item parameters are estimated more precisely. Joint estimation of item parameters and population parameters from the same data will be considered in a forthcoming paper (see Rigdon and Tsutakawa, 1983, for results in the univariate normal case.)

Let $\mathbf{x}_i$ be the vector pattern of item responses from subject i on the n_i items he has been administered. Under the usual assumption of conditional independence, the probability of this pattern, given $\boldsymbol{\theta}$, is the product of the probabilities of the individual item responses:

$$L(\mathbf{x}_i \mid \boldsymbol{\theta}) = \prod_j P(x_{ij} \mid \boldsymbol{\theta}, \zeta_j). \tag{4}$$

The marginal probability of response pattern $\mathbf{x}$ from a randomly-selected subject from the population of interest is then given as the integral of the conditional probability (4) over the population density as follows:

$$h(\mathbf{x}) = \int_{\boldsymbol{\theta}} L(\mathbf{x} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{5}$$

From (5), the log marginal likelihood for a sample of N subjects is then given by

$$\log L = \sum_i^N \log h(x_i) = \sum_i^N \log \int_{\boldsymbol{\theta}} L(\mathbf{x} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{6}$$

It will be noted that the response vectors of different subjects need not be based on the same set of test items. The procedures described below are therefore applicable to data solicited through multiple-matrix sampling designs and tailored testing schemes as well as through conventional test administration.

Maximum likelihood estimation of $\boldsymbol{\alpha}$ again begins by differentiating the log likelihood function (6):

$$\begin{aligned} \frac{\partial \log L}{\partial \alpha_s} &= \sum_i h^{-1}(\mathbf{x}_i) \frac{\partial h(\mathbf{x}_i)}{\partial \alpha_s} \\ &= \sum_i h^{-1}(\mathbf{x}_i) \frac{\partial}{\partial \alpha_s} \left[\int L(\mathbf{x}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \right] \qquad s = 1, \ldots, S. \end{aligned} \tag{7}$$

The differentiation and integration in (7) may be interchanged if the resulting integral is convergent for all values of α_s and $\boldsymbol{\theta}$ (Kendall and Stuart, 1979, p. 10.) These regularity conditions are widely satisfied by item response models and density functions in current use. We thus obtain

$$\begin{aligned}
\frac{\partial \log L}{\partial \alpha_s} &= \sum_i h^{-1}(\mathbf{x}_i) \int L(\mathbf{x}_i \mid \boldsymbol{\theta}) \frac{\partial\, g(\boldsymbol{\theta})}{\partial \alpha_s}\, d\boldsymbol{\theta} \\
&= \int \sum_i h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid \boldsymbol{\theta}) \frac{\partial g(\theta)}{\partial \alpha_s}\, d\boldsymbol{\theta} \\
&= \int \sum_i \left[g^{-1}(\boldsymbol{\theta}_i) \frac{\partial g(\boldsymbol{\theta})}{\partial \alpha_s} \right] \left[\frac{g(\boldsymbol{\theta}) L(\mathbf{x}_i \mid \boldsymbol{\theta})}{h(\mathbf{x}_i)} \right] d\boldsymbol{\theta} \\
&= N \int \left[g^{-1}(\boldsymbol{\theta}) \frac{\partial g(\boldsymbol{\theta})}{\partial \alpha_s} \right] \left[N^{-1} \sum_i p(\boldsymbol{\theta} \mid \mathbf{x}_i) \right] d\boldsymbol{\theta} \\
&= N \int \left[g^{-1}(\boldsymbol{\theta}) \frac{\partial g(\boldsymbol{\theta})}{\partial \alpha_s} \right] p(\boldsymbol{\theta} \mid (\mathbf{x}))\, d\boldsymbol{\theta}. \qquad (8)
\end{aligned}$$

The second to last step above will be recognized as an application of Bayes theorem, yielding the derivative of the incomplete data log likelihood with respect to α_s as the expected value of the corresponding complete data derivative, over the posterior density of $\boldsymbol{\theta}$ given the observed data. Such solutions are often referred to as "empirical Bayes," although they are not true Bayesian solutions unless a prior distribution for $\boldsymbol{\alpha}$ is specified (Deely and Lindley, 1981). Computation continues by equating these derivatives to zero to produce the likelihood equations, which may then be solved for $\boldsymbol{\alpha}$ in a number of ways.

Most straightforward is the solution by Newton-Raphson iterations, wherein provisional estimates $\hat{\boldsymbol{\alpha}}^t$ are updated by

$$\hat{\boldsymbol{\alpha}}^{t+1} = \hat{\boldsymbol{\alpha}}^t - \left(\frac{\partial \log L}{\partial \boldsymbol{\alpha}} \bigg|_{\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^t} \right) \left(\frac{\partial^2 \log L}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha}'} \bigg|_{\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^t} \right)^{-1}$$

The dependence of $p(\boldsymbol{\theta} \mid \mathbf{x}_i)$ upon $h(\mathbf{x}_i)$ and $g(\boldsymbol{\theta})$, and therefore upon $\boldsymbol{\alpha}$, must be taken into account in the calculation of second derivatives. The resulting expressions often prove cumbersome.

An alternative solution, based on the method of successive approximations, employs provisional estimates of $\boldsymbol{\alpha}$ to evaluate $p(\boldsymbol{\theta} \mid (\mathbf{x}))$ and then finds zeros of (8) to produce an improved estimate of $\boldsymbol{\alpha}$. This latter solution bears strong resemblance to the cycles of Dempster, Laird and Rubin (1977) EM algorithm for the exponential family. The difference is that rather than computing the conditional expected values of sufficient statistics for g given the data and provisional estimates of $\boldsymbol{\alpha}$, we compute the conditional expectation of the entire density itself through $p(\boldsymbol{\theta} \mid (\mathbf{x}))$. Ramsay (1975) has suggested methods to accelerate the rate of convergence of implicit equations of this type.

Large sample standard errors may be obtained again by first approximating the information matrix by

$$\sum_i \left(\frac{\partial \log h(\mathbf{x}_i)}{\partial \boldsymbol{\alpha}} \right) \left(\frac{\partial \log h(\mathbf{x}_i)}{\partial \boldsymbol{\alpha}'} \right) \qquad (9)$$

where

$$\frac{\partial \log h(\mathbf{x}_i)}{\partial \boldsymbol{\alpha}} = \int g^{-1}(\boldsymbol{\theta}) \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} p(\boldsymbol{\theta} \mid \mathbf{x}_i)\, d\boldsymbol{\theta},$$

then taking the square roots of the diagonal elements of its inverse.

Except in special cases such as the convolution of independent normal densities considered in Dempster, Rubin, and Tsutakawa (1981), the integration required to evaluate (6) and (8) cannot be carried out explicitly and must be approximated numerically. It is instructive to consider the results obtained when integration over g is approximated by quadrature over a finite number of points $\mathbf{X}_q$, $q = 1, \ldots, Q$, with associated weights $A(\mathbf{X}_q)$; that is

$$h(\mathbf{x}_i) \approx \sum_q \underline{L(\mathbf{x}_i} \,|\, \mathbf{X}_q)A(\mathbf{X}_q),$$

$$p(\mathbf{X}_q \,|\, \mathbf{x}_i) \approx \sum_q h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i \,|\, \mathbf{X}_q)A(\mathbf{X}_q),$$

and then

$$\frac{\partial \log L}{\partial \alpha_s} \approx \sum_q \left[g^{-1}(\mathbf{X}_q)\, \frac{\partial g(\mathbf{X}_q)}{\partial \alpha_s} \right] \sum_i p(\mathbf{X}_q \,|\, \mathbf{x}_i) \qquad s = 1, \ldots, S. \tag{10}$$

The computing approximations (10) of the likelihood equations are formally identical to the complete data likelihood equations (3) with respect to weighted pseudo-data points $\mathbf{X}_q$. Solution again proceeds iteratively, using provisional estimates of $\boldsymbol{\alpha}$ to evaluate the posterior weights $p(\mathbf{X}_q \,|\, (\mathbf{x}))$ then re-estimating $\boldsymbol{\alpha}$ by solving fascimilies of the complete data likelihood equations. This solution is equivalent to estimating the parameters of a discrete distribution of ability over the points $\mathbf{X}_q$, a problem treated in Section 4.3 of Dempster, Laird and Rubin (1977). Approximation of the information matrix given in (9) by the same quadrature formula yields

$$\sum_i \left[\sum_q \frac{\partial \log g(\mathbf{X}_q)}{\partial \boldsymbol{\alpha}}\, p(\mathbf{X}_q \,|\, \mathbf{x}_i) \right] \left[\sum_q \frac{\partial \log g(\mathbf{X}_q)}{\partial \boldsymbol{\alpha}'}\, p(\mathbf{X}_q \,|\, \mathbf{x}_i) \right], \tag{11}$$

an expression identical to that given in Louis (1982) for EM solutions in which the missing data consists of indicator variables for a multinomial distribution.

The chi-square approximation to the likelihood ratio test may be used to compare the fit of nested models contemplated for the characterization of a given distribution (e.g., a resolution into two normal components as opposed to three). If Model 1 is nested within Model 2, then

$$\chi^2 = -2 \sum_i \log \left[h_1(\mathbf{x}_i)/h_2(\mathbf{x}_i) \right]$$

where $h_1(\mathbf{x}_i)$ and $h_2(\mathbf{x}_i)$ are the marginal likelihoods of the response vector of subject i under Models 1 and 2 evaluated at their respective maximum likelihood solutions. The corresponding degrees of freedom is the difference in the number of population parameters estimated by the two models.

When the number of potential response patterns is small compared to the number of subjects, a given characterization of g may also be compared against the least restrictive alternative of a general distribution of response pattern counts (as in Bock and Lieberman, 1970). Let the sample of subjects be partitioned into exhaustive and mutually exclusive subsamples in a manner that ensures that the universe of potential response patterns that could have been observed is the same for all subjects in a given subsample. In multiple-matrix test administration, for example, a subsample would be comprised of everyone who had taken the same test form. Let $r(\mathbf{x}_\ell)$ be the count of observations of response pattern $\mathbf{x}_\ell$. Let $N(\mathbf{x}_\ell)$ be the number of subjects in the subsample for which $\mathbf{x}_\ell$ is a potential response pattern. Then a likelihood ratio chi-square against the general multinomial alternative is given by

$$\chi^2 = -2 \sum_\ell r(\mathbf{x}_\ell) \log \left[N(\mathbf{x}_\ell)h(\mathbf{x}_\ell)/r(\mathbf{X}_\ell) \right],$$

where $h(\mathbf{x}_\ell)$ is computed via (2). The degrees of freedom is the number of different response patterns that have been observed, minus the number of parameters estimated in the characterization of g, minus the number of subsamples since the expected proportions of potential response patterns must sum to one in each.

A Non-Parametric Solution

Suppose that $g(\boldsymbol{\theta})$ is a smooth continuous m-variate distribution with finite moments. It is then possible to approximate g without assuming a distributional form. Following the outline of the previous section, we first express the marginal likelihood of a given response vector as the m-fold definite integral

$$h(\mathbf{x}_i) = \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Let $\mathbf{X}_1, \ldots, \mathbf{X}_Q$ be a collection of points spanning the region where the bulk of the density of the distribution lies, sufficient in number and in proximity to insure non-trivial relative likelihoods of the typical response vector in the data at two or more points. The marginal likelihood of $\mathbf{x}_i$ may than be approximated as

$$h(\mathbf{x}_i) \approx \sum_{q=1}^{Q} L(\mathbf{x}_i \mid \mathbf{X}_q) p_q$$

where p_q is the density at point $\mathbf{X}_q$. The p_q are now to be estimated from the data. The continuous density g will be approximated by a discrete distribution on a finite number of points—in effect, an m-dimensional histogram. (In this presentation, the points $\mathbf{X}_q$ are assumed to be specified a priori. See Laird, 1978, on the simultaneous estimation of points and weights in the unidimensional case.)

The log likelihood of the data from the entire sample is now approximated as

$$\log L \approx \sum_{i}^{N} \log \left[\sum_{q} L(\mathbf{x}_i \mid \mathbf{X}_q) p_q\right].$$

A Lagrangian multiplier is introduced to constrain the sum of the p_q's to one:

$$\log L' = \log L + \lambda\left(1 - \sum_q p_q\right).$$

Differentiating with respect to the p_q's and λ yields

$$\frac{\partial \log L'}{\partial p_q} = \sum_i h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid \mathbf{X}_q) - \lambda$$

$$\frac{\partial \log L'}{\partial \lambda} = 1 - \sum_q p_q.$$

Equating to zero yields the following likelihood equations:

$$\lambda = \sum_i h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid \mathbf{X}_q), \qquad q = 1, \ldots, Q \tag{12}$$

$$\sum_q p_q = 1. \tag{13}$$

Multiplying each (12) by its respective p_q, we obtain

$$\lambda p_q = \sum_i h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid \mathbf{X}_q) p_q, \qquad q = 1, \ldots, Q. \tag{14}$$

Summing these expressions over q yields

$$\begin{aligned}\lambda \sum_q p_q &= \sum_q \sum_i h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i \mid \mathbf{X}_q)p_q \\ &= \sum_i h^{-1}(\mathbf{x}_i) \sum_q L(\mathbf{x}_i \mid \mathbf{X}_q)p_q \\ &= N.\end{aligned}$$

But from (13), $\sum p_q = 1$, implying that $\lambda = N$. After substitution into (14) and re-expression, the likelihood equations become

$$\begin{aligned}p_q &= N^{-1} \sum_i h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i \mid \mathbf{X}_q)p_q \\ &= N^{-1} \sum_i P(\mathbf{X}_q \mid \mathbf{x}_i), \qquad q = 1, \ldots, Q. \end{aligned} \tag{15}$$

Because $h(\mathbf{x}_i)$ depends on the values of the p_q, (15) constitutes a set of implicit equations that must be solved iteratively. Improved estimates $\hat{\mathbf{p}}^{t+1}$ of $\mathbf{p}$ are obtained successively by evaluating the right sides of (15) at provisional values $\hat{\mathbf{p}}^t$.

From the characterization of g in terms of the points $\mathbf{X}_q$ and densities p_q, it is possible to estimate attributes of g such as moments and percentile points. For example, the mean and variance of the distribution are approximated as

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \sum_q \mathbf{X}_q \hat{p}_q \\ \hat{\boldsymbol{\Sigma}} &= \sum_q (\mathbf{X}_q - \hat{\boldsymbol{\mu}})(\mathbf{X}_q - \hat{\boldsymbol{\mu}})'\hat{p}_q.\end{aligned}$$

Furthermore, it is possible to obtain approximate standard errors of estimation for these quantities, by first calculating and inverting the observed information matrix of the p's, then applying the delta method. We illustrate with the unidimensional case.

Because the p's are constrained to sum to one, their information matrix would be singular and can not be inverted as required. We therefore to recast the problem with just the first $Q - 1$ p's as parameters to be estimated and p_Q defined as follows:

$$p_Q = 1 - \sum_q^{Q-1} p_q.$$

With this re-expression, the first derivative of log likelihoods takes the following form:

$$\frac{\partial \log h(\mathbf{x}_i)}{\partial p_q} = h^{-1}(\mathbf{x}_i)[L(\mathbf{x}_i \mid \mathbf{X}_q) - L(\mathbf{x}_i \mid X_Q)] \qquad q = 1, \ldots, Q - 1$$

and the approximate information matrix has elements H_{qr} given by

$$\frac{\partial^2 \log L}{\partial p_q \partial p_r} = \sum_i h^{-2}(\mathbf{x}_i)[L(\mathbf{x}_i \mid X_q) - L(\mathbf{x}_i \mid X_Q)][L(\mathbf{x}_i \mid X_r) - L(\mathbf{x}_i \mid X_Q)] \qquad \begin{matrix} q = 1, \ldots, Q - 1 \\ r = 1, \ldots, Q - 1 \end{matrix}$$

A large-sample covariance matrix for the estimation errors of the first $Q - 1$ p's is given by $\mathbf{H}^{-1}$. An application of the delta method gives the large-sample covariance matrix for the estimation errors of $\hat{\mu}$ and $\hat{\sigma}^2$:

$$\underset{2\times 2}{\text{Cov}(\hat{\mu}, \hat{\sigma})} = -\underset{2\times Q}{\mathbf{F}'} \underset{Q\times Q-1}{[\mathbf{I} \vdots \ -1]'} \ \underset{Q-1\times Q-1}{\mathbf{H}^{-1}} \ \underset{Q-1\times Q}{[\mathbf{I} \vdots \ -1]} \ \underset{Q\times 2}{\mathbf{F}},$$

where

$\mathbf{I}$ is the $Q - 1$ by $Q - 1$ identity matrix,

$\mathbf{1}$ is the column vector consisting of $Q - 1$ unit elements, and

$$\mathbf{F} = \begin{bmatrix} X_1 & \cdots & X_Q \\ (X_1 - \hat{\mu})^2 & \cdots & (X_Q - \hat{\mu})^2 \end{bmatrix}.$$

Standard errors for $\hat{\mu}$ and $\hat{\sigma}^2$ are obtained as the square roots of the diagonal elements of the resulting error covariance matrix.

In practice, $\mathbf{H}$ may be poorly conditioned for inversion if the number of points $\mathbf{X}$ is large, especially when points at the extremes have little weight. For this reason, practical application of standard errors for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the non-parametric case will typically require few quadrature points and/or many observations.

A Normal Solution

Suppose that $g(\boldsymbol{\theta})$ is m-variate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The marginal likelihood of $\mathbf{x}_i$ is given by the m-fold definite integral

$$h(\mathbf{x}_i) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{m/2}} \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) \exp\left[-\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right] d\boldsymbol{\theta}. \tag{16}$$

For a sample of N subjects, the log likelihood is again given as

$$\log L = \sum_i^N \log h(\mathbf{x}_i).$$

The derivative of $\log L$ with respect to $\boldsymbol{\mu}$ is then

$$\frac{\partial \log L}{\partial \boldsymbol{\mu}} = \sum_i h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\, d\theta.$$

Equating to zero,

$$\hat{\boldsymbol{\mu}} \sum_i^N h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \sum_i^N h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\boldsymbol{\theta}\, d\boldsymbol{\theta}$$

or

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= N^{-1} \int_{\boldsymbol{\theta}} \boldsymbol{\theta} \sum_i^N h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \boldsymbol{\theta}\left[N^{-1} \sum_i p(\boldsymbol{\theta} \mid \mathbf{x}_i)\right] d\boldsymbol{\theta}. \end{aligned} \tag{17}$$

The similarity to the standard calculation of a mean vector will be recognized, with the density of $\boldsymbol{\theta}$ replaced by the average over subjects of the conditional densities of $\boldsymbol{\theta}$ given each subject's response vector.

Similarly,

$$\begin{aligned} \frac{\partial \log L}{\partial \boldsymbol{\Sigma}} &= \frac{\partial}{\partial \boldsymbol{\Sigma}} N \log |\boldsymbol{\Sigma}|^{-1/2} + \sum_i h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x} \mid \boldsymbol{\theta}) \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}}\, d\boldsymbol{\theta} \\ &= -\frac{N}{2}(2\boldsymbol{\Sigma}^{-1} - \operatorname{diag} \boldsymbol{\Sigma}^{-1}) + \tfrac{1}{2} \sum_i h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta}) \\ &\quad \times \{2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1} - \operatorname{diag}[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}]\}\, d\boldsymbol{\theta} \\ &= \frac{N}{2} \operatorname{diag}[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{S})\boldsymbol{\Sigma}^{-1}] - N\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathbf{S})\boldsymbol{\Sigma}^{-1}, \end{aligned}$$

where

$$\begin{aligned}\mathbf{S} &= N^{-1}\int_{\boldsymbol\theta}(\boldsymbol\theta-\hat{\boldsymbol\mu})(\boldsymbol\theta-\hat{\boldsymbol\mu})'\sum_i^N h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i\mid\boldsymbol\theta)g(\boldsymbol\theta)\,d\boldsymbol\theta\\ &= \int_{\boldsymbol\theta}(\boldsymbol\theta-\hat{\boldsymbol\mu})(\boldsymbol\theta-\hat{\boldsymbol\mu})'\Big[N^{-1}\sum_i p(\boldsymbol\theta\mid\mathbf{x}_i)\Big]\,d\boldsymbol\theta. \qquad (18)\end{aligned}$$

It is clear that the partial derivative of log L with respect to $\boldsymbol\Sigma$ takes the value zero when $\boldsymbol\Sigma = \mathbf{S}$.

Because $h(\mathbf{x}_i)$ depends on $\boldsymbol\mu$ and $\boldsymbol\Sigma$, (17) and (18) again constitute an implicit system requiring iterative solution. In practical calculation, the integration must be approximated by numerical procedures. Three approaches are sketched below.

Gauss-Hermite Quadrature

The preferred method of numerical integration over the normal density function, as is required in (16), (17) and (18), is Gauss-Hermite quadrature. Integration over the normal density is replaced by summation over a finite number of points, each assigned an optimal weight so as to best approximate the integral in question. For the standard normal univariate density, Stroud and Sechrest (1966) provide tables of optimal points and weights, which shall be denoted here as Z_q and $W(Z_q)$ respectively for $q = 1, \ldots, Q$. A grid of quadrature points and weights for the m-variate standard normal density (i.e., with zero means and identity covariance matrix) is obtained as the Cartesian product of univariate sets of quadrature points. A typical point in the grid has the form

$$\mathbf{Z}_q = (Z_{q1}, Z_{q2}, \ldots, Z_{qm}),$$

with weight given by

$$A(\mathbf{Z}_q) = \prod_{t=1}^{m} W(Z_{qt}).$$

Because the integration in (16)–(18) is over a general multivariate normal distribution with mean $\boldsymbol\mu$ and covariance matrix $\boldsymbol\Sigma$, a change of variables of integration is required to apply the Gauss-Hermite formula. We illustrate with (17). Let $\mathbf{z} = (\boldsymbol\theta - \boldsymbol\mu)\mathbf{T}$, where $\mathbf{TT}' = \boldsymbol\Sigma$ is the Cholesky factorization of $\boldsymbol\Sigma$ (implying that $|\mathbf{T}| = |\boldsymbol\Sigma|^{1/2}$). Then

$$\begin{aligned}\hat{\boldsymbol\mu} &= N^{-1}\int_{\boldsymbol\theta}\frac{|\boldsymbol\Sigma|^{-1/2}}{(2\pi)^{m/2}}\,\boldsymbol\theta\sum_i h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i\mid\theta)\exp\,[(\boldsymbol\theta-\mu)\boldsymbol\Sigma^{-1}(\boldsymbol\theta-\boldsymbol\mu)']\,d\theta\\ &= N^{-1}\int_{\mathbf{z}}\frac{|\boldsymbol\Sigma|^{-1/2}}{(2\pi)^{m/2}}\,\boldsymbol\theta(\mathbf{z})\sum_i h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i\mid\boldsymbol\theta(\mathbf{z}))\\ &\qquad\times\exp\,[(\mathbf{zT}+\boldsymbol\mu-\boldsymbol\mu)\boldsymbol\Sigma^{-1}(\mathbf{zT}+\boldsymbol\mu-\boldsymbol\mu))']\,|\mathbf{T}|\,d\mathbf{z}\\ &= \frac{N^{-1}}{(2\pi)^{m/2}}\int_{\mathbf{z}}\boldsymbol\theta(\mathbf{z})\sum_i h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i\mid\boldsymbol\theta(\mathbf{z}))\exp\,(\mathbf{z}\,\mathbf{z}')\,d\mathbf{z}\end{aligned}$$

where

$$\boldsymbol\theta(\mathbf{z}) = \mathbf{zT} + \boldsymbol\mu.$$

Define $\mathbf{X}_q$ as $\boldsymbol\theta(\mathbf{Z}_q)$, where $\mathbf{Z}_q$ is a point in the standard normal quadrature grid and define $A(\mathbf{X}_q)$ as $A(\mathbf{Z}_q)$. Computing approximations of (17) and (18) may now be written as

$$\begin{aligned}\hat{\boldsymbol\mu} &\approx N^{-1}\sum_q \mathbf{X}_q\sum_i h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i\mid\mathbf{X}_q)A(\mathbf{X}_q)\\ &= N^{-1}\sum_q \mathbf{X}_q\sum_i P_{iq} \qquad (19)\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{\Sigma}} &\approx N^{-1}\sum_q (\mathbf{X}_q - \hat{\boldsymbol{\mu}})(\mathbf{X}_q - \hat{\boldsymbol{\mu}})' \sum_i h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i \mid \mathbf{X}_q)A(\mathbf{X}_q)\\ &= N^{-1}\sum_q (\mathbf{X}_q - \hat{\boldsymbol{\mu}})(\mathbf{X}_q - \hat{\boldsymbol{\mu}})'\Big(\sum_i P_{iq}\Big)\end{aligned} \tag{20}$$

where

$$h(\mathbf{x}_i) \approx \sum_q^{Q^m} L(\mathbf{x}_i \mid \mathbf{X}_q)A(\mathbf{X}_q)$$

is the marginal likelihood of response vector $\mathbf{x}_i$ and

$$P_{iq} = h^{-1}(\mathbf{x}_i)L(\mathbf{x}_i \mid \mathbf{X}_q)A(\mathbf{X}_q)$$

is the posterior probability that subject i has ability $\mathbf{X}_q$.

Quadrature Over Fixed Points

Define a grid of Q^m quadrature points given by the Cartesian product

$$(X_1, \ldots, X_Q)x(X_1, \ldots, X_Q)x \ldots x(X_1, \ldots, X_Q)$$

spanning the region where the bulk of the distribution lies. Let a weight be associated with each point $\mathbf{X}_q$ as follows:

$$A(\mathbf{X}_q) = K \exp\left[-\tfrac{1}{2}(\mathbf{X}_q - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{X}_q - \boldsymbol{\mu})\right],$$

where K is the normalizing constant that makes the weights sum to one. These same points will be maintained throughout the solution of the likelihood equations, although the weights are updated at each iteration. The computational advantage accrues that the likelihoods $L(\mathbf{x}_i \mid \mathbf{X}_q)$ required in solution are the same in each cycle and need therefore be calculated only once. The computing approximations remain otherwise unchanged.

Monte Carlo Integration

The quadrature described above becomes unwieldy if the number of dimensions m is large and inefficient if the dimensions are highly correlated in the data. Under these circumstances, Monte Carlo numerical integration is preferable (see Bock, 1983). First, a number of multivariate random variables are generated in each iterative step in accordance with provisional estimates of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. Next, the likelihood of each subject's response vector is evaluated at each of these points; the sum of these values is taken as the marginal probability of his data. Then, the sum of normalized likelihoods over subjects at a given point is taken to be proportional to the density of the distribution at that point. The likelihood expressions are re-evaluated to produce improved estimates of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$.

Approximating the Information Matrix

The information matrix is approximated by the sum of gramian products over subject of the terms

$$\partial \log h(\mathbf{x}_i)/\partial(\boldsymbol{\mu}, \mathbf{\Sigma}),$$

a column vector of dimension $m + m(m + 1)/2$ consisting of the m elements of

$$\sum_q P_{iq}(\mathbf{X}_q - \hat{\boldsymbol{\mu}})'\,\hat{\mathbf{\Sigma}}^{-1}$$

and the $m(m + 1)/2$ elements of the lower half of the symmetric matrix

$$1/2\ \hat{\mathbf{\Sigma}}^{-2}\Big[\sum_q P_{iq}(\mathbf{X}_q - \hat{\boldsymbol{\mu}})'(\mathbf{X}_q - \hat{\boldsymbol{\mu}}) - \hat{\mathbf{\Sigma}}\Big].$$

For example, the elements of the information matrix for univariate normal case are approximated as follows:

$$\mu,\ \mu\colon\quad \sum_i \left[\sum_q h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid X_q) A(X_q) \frac{X_q - \hat{\mu}}{\hat{\sigma}^2} \right]^2$$

$$\mu,\ \sigma^2\colon\quad \sum_i \left[\sum_q h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid X_q) A(X_q) \frac{X_q - \hat{\mu}}{\hat{\sigma}^2} \right] \times \left[\sum_q h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid X_q) A(X_q) \frac{(X_q - \hat{\mu})^2 - \hat{\sigma}^2}{2\hat{\sigma}^4} \right]$$

$$\sigma^2,\ \sigma^2\colon\quad \sum_i \left[\sum_q h^{-1}(\mathbf{x}_i) L(\mathbf{x}_i \mid X_q) A(X_q) \frac{(X_q - \hat{\mu})^2 - \hat{\sigma}^2}{2\hat{\sigma}^4} \right]^2.$$

Resolution of Normal Components

Suppose that $g(\boldsymbol{\theta})$ is a mixture of K normal components, with mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, and common covariance matrix $\boldsymbol{\Sigma}$. Let $p_1, \ldots, p_K$ be the unknown proportions of the mixture. Day's (1969) maximum likelihood equations could be applied to estimate these population parameters if values of $\boldsymbol{\theta}_i$ were observed directly rather than implied by $\mathbf{x}_i$; the following solution represents in this sense a generalization of Day's procedure.

To begin, the marginal likelihood of a given esponse vector $\mathbf{x}_i$ is given by the m-fold definite integral

$$h(\mathbf{x}_i) = \sum_k p_k \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

where

$$g_k(\boldsymbol{\theta}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{m/2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_k)\right].$$

The log likelihood of a sample of size N, augmented by a Lagrangian multiplier to ensure that the p_k's sum to one, is then

$$\log L' = \sum_i^N \log h(\mathbf{x}_i) + \lambda\left(1 - \sum_k p_k\right). \tag{21}$$

We obtain likelihood equations by differentiating with respect to the p_k's, the $\boldsymbol{\mu}_k$'s, $\boldsymbol{\Sigma}$, and λ in turn and equating these expressions to zero.

Differentiating (21) with respect to p_k, we obtain

$$\frac{\partial \log L'}{\partial p_k} = \sum_i h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta} - \lambda.$$

Equating to zero,

$$\lambda = \sum_i h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \qquad k = 1, \ldots, K.$$

Multiplying both sides by p_k,

$$\lambda p_k = \sum_i p_k h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \qquad k = 1, \ldots, K. \tag{22}$$

Summing these equations over k yields

$$\begin{aligned}\lambda \sum_k p_k &= \sum_k \sum_i p_k h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \sum_i h^{-1}(\mathbf{x}_i) \sum_k p_k \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= N. \qquad (23)\end{aligned}$$

But differentiating (21) with respect to λ yields

$$\sum_k p_k = 1,$$

which, together with (23), implies that $\lambda = N$. Substituting this value back into (22) and dividing both sides of the result by N produces the following likelihood equations for the p_k:

$$\begin{aligned}\hat{p}_k &= N^{-1} \sum_i h^{-1}(\mathbf{x}_i)\hat{p}_k \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\mathbf{u})\, d\boldsymbol{\theta} \\ &= N^{-1} \int_{\boldsymbol{\theta}} \sum_i h^{-1}(\mathbf{x}_i)\hat{p}_k L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= N^{-1} \int_{\boldsymbol{\theta}} \sum_i p(\boldsymbol{\theta} \mid \mathbf{x}_i, I_i = k)\, d\boldsymbol{\theta}, \qquad k = 1, \ldots, K \qquad (24)\end{aligned}$$

where I_i is an indicator variable that takes the value k if and only if subject i belongs to component k of the mixture. The summand for each subject in (24) may be interpreted as the portion of the conditional density of $\boldsymbol{\theta}$ given $\mathbf{x}_i$ that is associated with component k.

Differentiating (21) with respect to $\boldsymbol{\mu}_k$ yields

$$\frac{\partial \log L'}{\partial \boldsymbol{\mu}_k} = \sum_i h^{-1}(\mathbf{x}_i) p_k \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1} g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Equating to zero,

$$\sum_i \hat{p}_k h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_k \mid \boldsymbol{\theta}) \hat{\boldsymbol{\mu}}_k g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \sum_i \hat{p}_k h^{-1}(\mathbf{x}_i) \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) \boldsymbol{\theta} g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

or, after simplification,

$$\hat{\boldsymbol{\mu}}_k = \frac{\int_{\boldsymbol{\theta}} \boldsymbol{\theta} \sum_i p(\boldsymbol{\theta} \mid \mathbf{x}_i, I_i = k)\, d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} \sum_i p(\boldsymbol{\theta} \mid \mathbf{x}_i, I_i = k)\, d\boldsymbol{\theta}}, \qquad k = 1, \ldots, K. \qquad (25)$$

Differentiating (21) with respect to $\boldsymbol{\Sigma}$ yields

$$\begin{aligned}\frac{\partial \log L'}{\partial \boldsymbol{\Sigma}} &= -\frac{N}{2}(2\boldsymbol{\Sigma}^{-1} - \operatorname{diag} \boldsymbol{\Sigma}^{-1}) \\ &\quad + \tfrac{1}{2} \sum_i h^{-1}(\mathbf{x}_i) \sum_k p_k \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta}) g_k(\boldsymbol{\theta}) [2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1} \\ &\quad - \operatorname{diag} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}]\, d\boldsymbol{\theta}.\end{aligned}$$

Equating to zero and proceeding as in the normal solution, we obtain

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} &= N^{-1} \sum_i h^{-1}(\mathbf{x}_i) \sum_k \hat{p}_k \int_{\boldsymbol{\theta}} L(\mathbf{x}_i \mid \boldsymbol{\theta})(\boldsymbol{\theta} - \hat{\boldsymbol{\mu}})'(\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}) g_k(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} (\boldsymbol{\theta} - \hat{\boldsymbol{\mu}})'(\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}) \Big[N^{-1} \sum_i \sum_k p(\boldsymbol{\theta} \mid \mathbf{x}_i, I_i = k)\Big]\, d\boldsymbol{\theta}. \qquad (26)\end{aligned}$$

Computing Approximations

All three techniques of numerical integration described in the section on the normal solution can be applied to solve the likelihood equations (24), (25), and (26).

For solution by quadrature over fixed points, define again the grid of Q^m points by the Cartesian product

$$(X_1, \ldots, X_Q) \times (X_1, \ldots, X_Q) \times \cdots \times (X_1, \ldots, X_Q).$$

Associated with each individual point $\mathbf{X}_q$ are K weights, one for each component in the mixture:

$$A_k(\mathbf{X}_q) = C_k \exp\left[-\tfrac{1}{2}(\mathbf{X}_q - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}(\mathbf{X}_q - \boldsymbol{\mu}_k)\right],$$

where C_k is the normalizing constant that makes the weights sum to one over all points for component k.

Replacing integration with summation, the marginal likelihood of the response vector $\mathbf{x}_i$ is approximated as

$$h(\mathbf{x}_i) \approx \sum_k \sum_q p_k L(\mathbf{x}_i \mid \mathbf{X}_q) A_k(\mathbf{X}_q). \tag{27}$$

Denote the conditional probability that subject i has ability $\mathbf{X}_q$ and belongs to component k as

$$\begin{aligned} P_{ikq} &= h^{-1}(\mathbf{x}_i) p_k L(\mathbf{x}_i \mid \mathbf{X}_q) A_k(\mathbf{X}_q) \\ &= P(\mathbf{X}_q, I_i = k \mid \mathbf{x}_i). \end{aligned} \tag{28}$$

The likelihood equations are now approximated as

$$\hat{p}_k \approx N^{-1} \sum_q \sum_i P_{ikq} \qquad k = 1, \ldots, K \tag{29}$$

$$\hat{\boldsymbol{\mu}}_k \approx \Big(\sum_q \mathbf{X}_q \sum_i P_{ikq}\Big) \Big/ \Big(\sum_q \sum_i P_{ikq}\Big) \qquad k = 1, \ldots, K \tag{30}$$

and

$$\hat{\boldsymbol{\Sigma}} \approx N^{-1} \sum_q \sum_k (\mathbf{X}_q - \hat{\boldsymbol{\mu}})(\mathbf{X}_q - \hat{\boldsymbol{\mu}}_k)' \sum_i P_{ikq}. \tag{31}$$

These equations are identical in form to Day's (Day, 1969, p. 473), with the quadrature points interpreted as weighted data points. Solution is again iterative, since the quantities on the right sides of the likelihood equations depend on the population parameters to be estimated.

For solution by Gauss-Hermite quadrature, it is necessary to produce a grid of points for each component of the mixture in each iteration. As with the normal solution, first define a grid of points of quadrature points for the univariate normal; a typical point will be

$$\mathbf{Z}_q = (Z_{q1}, Z_{q2}, \ldots, Z_{qm})$$

and its associated weight will be

$$A(\mathbf{Z}_q) = \prod_{t=1}^{m} W(Z_{qt}),$$

where Z_{qt} and $W(Z_{qt})$ are taken from the tables of Stroud and Sechrest (1966). Again let $\mathbf{TT}' = \boldsymbol{\Sigma}$ be the Cholesky factorization of $\boldsymbol{\Sigma}$. Finally, define grids of quadrature points and associated weights by

$$\mathbf{X}_{kq} = \mathbf{Z}_q \mathbf{T} + \boldsymbol{\mu}_k$$

and

$$A(\mathbf{X}_{kq}) = A(\mathbf{Z}_q).$$

These terms replace $\mathbf{X}_k$ and $A_k(\mathbf{X}_q)$ respectively in equations (27) and (28). The computing formulas remained otherwise unchanged.

For solution by Monte Carlo integration, random normal deviates are generated in accordance with provisional values of the $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$, in proportions given by the provisional values of the p_k. The equations have the same appearance as those (29)–(31), except that the $A_k(\mathbf{X}_q)$ are all identical due to the self-weighting nature of the points.

Approximating the Information Matrix

As in the normal solution, we may apply (9) to approximate the information matrix. As in the non-parametric solution, however, it is first necessary to recast the linear-dependent set of p_k's in terms of just the first $K - 1$ as parameters to be estimated, in order to avoid a singular information matrix. We illustrate with the resolution of two univariate normal components.

With two univariate normal components to resolve, there are four parameters to estimate: μ_1, μ_2, σ^2, and p, the proportion of the population represented by the first component (implying that the proportion represented by the second component is $1 - p$). The elements of $\partial \log h(\mathbf{x}_i)/\partial(p, \mu_1, \mu_2, \sigma^2)$ are then given by

$$p\colon \quad \frac{1}{2\sigma^2} \sum_q \left[P_{i1q}(X_q - \hat{\mu}_1)^2/p - P_{i2q}(X_q - \hat{\mu}_2)^2/(1 - p)\right]$$

$$\mu_1\colon \quad \frac{1}{\sigma^2} \sum_q P_{i1q}(X_q - \hat{\mu}_1)$$

$$\mu_2\colon \quad \frac{1}{\sigma^2} \sum_q P_{i2q}(X_q - \hat{\mu}_2)$$

$$\sigma^2\colon \quad \frac{1}{2\sigma^4} \sum_k \sum_q \{P_{ikq}[(X_q - \hat{\mu}_k)^2 - \sigma^2)]\}.$$

The information matrix is obtained as the accumulated sum over subjects of these vectors squared.

A Beta-Binomial Solution

Lord and Novick (1968) describe a model for number-correct scores on an n-item test under which correct scores have a binomial distribution conditional on expected percent-correct scores ("true scores") and true scores in a population of interest follow a beta distribution. That is, the probability of test score r given true score θ is

$$L(r \mid \theta) = \frac{n!}{(n - r)!\,r!}\, \theta^r(1 - \theta)^{n-r} \qquad r = 0, \ldots, n \tag{32}$$

while the density of the true score distribution is

$$g(\theta) = B^{-1}(\alpha, \beta)\theta^{\alpha-1}(1 - \theta)^{\beta-1} \tag{33}$$

where

$$B(\alpha, \beta) = \int_0^1 \phi^{\alpha-1}(1 - \phi)^{\beta-1} d\phi$$

and α and β are the parameters of the population distribution. It will be noted that the integration in the definition of the beta function $B(\alpha, \beta)$ runs from zero to one, the range of potential true scores. This range will be assumed for all remaining integrals in this section.

As in the cases considered above, the objective is to estimate the population parameters directly from a sample of observed response vectors—or, observed test scores in this case, as items are considered interchangeable. Excellent starting values for maximum likelihood estimation are available from the analysis of Lord and Novick, who note that observed scores will have a negative hypergeometric distribution with parameters α and $\beta + n - 1$. Replacing sample statistics for population parameters in the formulas for the mean and varaince of the negative hypergeometric distribution yields the following approximations:

$$\tilde{\alpha} = (-1 + 1/\mathrm{KR}_{21})\bar{x}$$

$$\tilde{\beta} = (-1 + 1/\mathrm{KR}_{21})n - \tilde{\alpha}$$

where

$$\mathrm{KR}_{21} = \frac{n}{n-1}\left[1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right]$$

is the Kuder-Richardson 21 reliability coefficient, and $\bar{x}$ and s^2 are the sample mean and variance of the test scores.

Solving the "Complete Data" Problem

It is instructive to consider at this point the maximum likelihood estimation of α and β when the data consist of values of θ rather than of test scores. Suppose that values $\theta_1, \ldots, \theta_k$ have been observed with respective frequencies $N_1, \ldots, N_k$. Let N be the sum of the N_k's, or the total sample size. The likelihood function for this data set is the product over subjects of terms like (33)

$$L = B^{-N}(\alpha, \beta) \prod_k [\theta_k^{\alpha-1}(1-\theta_k)^{\beta-1}]^{N_k}$$

and the log likelihood function is

$$\log L = (\alpha - 1) \sum_k N_k \log \theta_k + (\beta - 1) \sum_k \log (1 - \theta_k) - N \log B(\alpha, \beta).$$

The first derivatives of log L with respect to α and β are given as

$$\frac{\partial \log L}{\partial \alpha} = \sum_k N_k \log \theta_k - NB^{-1}(\alpha, \beta) \int_\phi \phi^{\alpha-1}(1-\phi)^{\beta-1} \log \phi d\phi \tag{34}$$

$$\frac{\partial \log L}{\partial b} = \sum_k N_k \log (1 - \theta_k) - NB^{-1}(\alpha, \beta) \int_\phi \phi^{\alpha-1}(1-\phi)^{\beta-1} \log (1-\phi)\, d\phi. \tag{35}$$

Equating to zero yields the likelihood equations, which after re-expression are written as

$$\alpha\colon\ B^{-1}(\alpha, \beta) \int_\phi \phi^{\alpha-1}(1-\phi)^{\beta-1} \log \phi\, d\phi = N^{-1} \sum_k N_k \log \theta_k \tag{36}$$

$$\beta\colon\ B^{-1}(\alpha, \beta) \int_\phi \phi^{\alpha-1}(1-\phi)^{\beta-1} \log (1-\phi)\, d\phi = N^{-1} \sum_k N_k \log (1-\theta_k). \tag{37}$$

Solution of the likelihood equations may be accomplished by Newton-Raphson iterations. Improved estimates are obtained from provisional estimates at iteration t by

$$\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix}^{t+1} = \begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix}^{t} - \mathbf{H}^{-1}\left[\frac{\partial \log L}{\partial \alpha}\ \frac{\partial \log L}{\partial \beta}\right]$$

where $\mathbf{H}$ is the matrix of second derivatives of log L with respect to α and β, and both the first and second derivatives are evaluated at the provisional estimates. The elements of $\mathbf{H}$ are given by

$$\frac{\partial \log L}{\partial \alpha^2} = -NB^{-2}(\alpha, \beta)\left\{\left[B(\alpha, \beta)\int \phi^{\alpha-1}(1-\phi)^{\beta-1}\log^2 \phi d\phi\right] - \left[\int \phi^{\alpha-1}(1-\phi)^{\beta-1}\log \phi d\phi\right]^2\right\}$$

$$\frac{\partial^2 \log L}{\partial \alpha \partial \beta} = -NB^{-2}(\alpha, \beta)\left\{\left[B(\alpha, \beta)\int \phi^{\alpha-1}(1-\phi)^{\beta-1}\log \phi \log(1-\phi) d\phi\right] - \left[\int \phi^{\alpha-1}(1-\phi)^{\beta-1}\log \phi d\phi\right]\left[\int \phi^{\alpha-1}(1-\phi)^{\beta-1}\log(1-\phi) d\phi\right]\right\}$$

$$\frac{\partial^2 \log L}{\partial \beta^2} = -NB^{-2}(\alpha, \beta)\left\{\left[B(\alpha, \beta)\int \phi^{\alpha-1}(1-\phi)^{\beta-1}\log^2(1-\phi) d\phi\right] - \left[\int \phi^{\alpha-1}(1-\phi)^{\beta-1}\log(1-\phi) d\phi\right]^2\right\}.$$

Solving the "Incomplete Data" Problem

Suppose that rather than values of θ, test scores depending on θ through (32) are observed. Let x_i represent the score of subject i and let $N_0, N_1, \ldots, N_n$ be the counts of test scores of 0, 1, ..., n, summing to the total sample size N. As before, the marginal probability of test score r is given by

$$h(r) = \int_\theta L(r \mid \theta)\, \theta^{\alpha-1}(1-\theta)^{\beta-1}\, B^{-1}(\alpha, \beta)\, d\theta.$$

and the log likelihood of the entire sample is given by

$$\begin{aligned}\log L &= \sum_i h(r_i)\\ &= \sum_r N_r h(r). \qquad (38)\end{aligned}$$

The maximum likelihood estimates of α and β are the values that maximize (38) for the observed test score distribution. Differentiating with respect to α, we obtain

$$\begin{aligned}\frac{\partial \log L}{\partial \alpha} &= \sum_r N_r h^{-1}(r)\int_\theta L(r \mid \theta)\frac{\partial g(\theta)}{\partial \alpha}\, d\theta\\ &= \sum_r N_r h^{-1}(r)\int_\theta L(r \mid \theta)[B^{-1}(\alpha, \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &\quad - B^{-1}(\alpha, \beta)\textstyle\int \phi^{\alpha-1}(1-\phi)^{\beta-1}\log \phi d\phi]\, d\theta.\end{aligned}$$

Equating to zero and simplifying yields the likelihood equation for α:

$$B^{-1}(\alpha, \beta) \int \phi^{\alpha-1}(1-\phi)^{\beta-1} \log \phi d\phi = N^{-1} \sum_r N_r h^{-1}(r) \int \log \theta \, L(r \mid \theta) g(\theta) \, d\theta$$

$$= N^{-1} \int \log \theta \left[\sum_r N_r p(\theta \mid r)\right] d\theta. \qquad (39)$$

By similar calculation, the likelihood equation corresponding to β is found to be

$$B^{-1}(\alpha, \beta) \int \phi^{\alpha-1}(1-\phi)^{\beta-1} \log (1-\phi) \, d\phi$$

$$= N^{-1} \int \log (1-\theta)\left[\sum_r N_r p(\theta \mid r)\right] d\theta. \qquad (40)$$

The resemblance of (39) and (40) to (36) and (37), the likelihood equations for the complete data problem, is clear. The difference is that the sum of logarithms of observed true-score values has been replaced integration over the range of potential true-score values, each point weighted by its expectation given the observed score distribution. The required integrals may be evaluated by means of standard procedures such as Simpson's rule or Monte Carlo methods. One simple solution follows.

Let $X_1, \ldots, X_q$ be set of evenly-spaced points covering the unit interval. Define weights $A(X_q)$ as follows:

$$A(X_q) = \frac{X_q^{\alpha-1}(1-X_q)^{\beta-1}}{\sum_S X_S^{\alpha-1}(1-X_S)^{\beta-1}}.$$

These points and weights define a discrete distribution that may be used to approximate the continuous beta distribution, with the approximation as good as desired by sufficiently large Q. For example, $h(r)$ may be approximated as follows:

$$h(r) \approx \sum_q L(r \mid X_q) A(X_q).$$

The likelihood equations (39) and (40) may also be approximated, as

$$\frac{\sum_q X_q^{\alpha-1}(1-X_q)^{\beta-1} \log X_q}{\sum_q X_q^{\alpha-1}(1-X_q)^{\beta-1}} = N^{-1} \sum_q \log X_q \sum_r N_r P(X_q \mid r)$$

and

$$\frac{\sum_q X_q^{\alpha-1}(1-X_q)^{\beta-1} \log (1-X_q)}{\sum_r X_q^{\alpha-1}(1-X_q)^{\beta-1}} = N^{-1} \sum_q \log (1-X_q) \sum_r N_r P(X_q \mid r)$$

where

$$P(X_q \mid r) = \frac{L(r \mid X_q) A(X_q)}{\sum_S L(r \mid X_S) A(X_S)}.$$

Approximating the Information Matrix

Again (9) is the basis for approximating the information matrix, employing expressions for the complete data derivatives given in (34) and (35). Terms are identical for all subjects with the same score, so the required summation may be run over scores rather than over subjects. The elements of the resulting approximation are as follows:

$$\alpha, \alpha: \quad \sum_r N_r[\sum_q P(X_q | r) \log X_q - C]^2$$

$$\alpha, \beta: \quad \sum_r N_r[\sum_q P(X_q | r) \log X_q - C][\sum_q P(X_q | r) \log (1 - X_q) - D]$$

$$\beta, \beta: \quad \sum_r N_r[\sum_q P(X_q | r) \log (1 - X_q) - D]^2$$

where

$$C = [\sum_q X_q^{\alpha-1}(1 - X_q)^{\beta-1} \log X_q]/[\sum_q X_q^{\alpha-1}(1 - X_q)^{\beta-1}]$$

$$D = [\sum_q X_q^{\alpha-1}(1 - X_q)^{\beta-1} \log (1 - X_q)]/[\sum_q X_q^{\alpha-1}(1 - X_q)^{\beta-1}].$$

Example 1: The Law School Admissions Test

Table 1 presents observed counts of response patterns from samples of 1,000 subjects to two five-item subsets of the Law School Admissions Test. Several analyses of these data sets, referred to hereafter as LSAT-6 and LSAT-7, have appeared in the psycho-

TABLE 1

FREQUENCY COUNTS OF RESPONSE PATTERNS FOR LSAT EXAMPLE

	ITEM SCORE					LSAT 6		LSAT 7	
INDEX	1	2	3	4	5	OBSERVED	EXPECTED	OBSERVED	EXPECTED
1	0	0	0	0	0	3	2.363	12	10.191
2	0	0	0	0	1	6	5.458	19	17.343
3	0	0	0	1	0	2	2.475	1	4.331
4	0	0	0	1	1	11	8.241	7	8.896
5	0	0	1	0	0	1	0.851	3	5.969
6	0	0	1	0	1	1	2.832	19	16.259
7	0	0	1	1	0	3	1.284	3	4.284
8	0	0	1	1	1	4	6.212	17	13.948
9	0	1	0	0	0	1	1.819	10	4.733
10	0	1	0	0	1	8	6.057	5	9.822
11	0	1	0	1	0	0	2.746	3	2.509
12	0	1	0	1	1	16	13.283	7	6.162
13	0	1	1	0	0	0	0.944	7	7.647
14	0	1	1	0	1	3	4.565	23	25.642
15	0	1	1	1	0	2	2.070	8	6.822
16	0	1	1	1	1	15	14.732	28	27.852
17	1	0	0	0	0	10	10.280	7	11.225
18	1	0	0	0	1	29	34.232	39	34.965
19	1	0	0	1	0	14	15.520	11	8.756
20	1	0	0	1	1	81	75.075	34	33.729
21	1	0	1	0	0	3	5.334	14	10.434
22	1	0	1	0	1	28	25.803	51	54.230
23	1	0	1	1	0	15	11.699	15	13.750
24	1	0	1	1	1	80	83.262	90	92.203
25	1	1	0	0	0	16	11.406	6	7.354
26	1	1	0	0	1	56	55.174	25	28.104
27	1	1	0	1	0	21	25.015	7	7.070
28	1	1	0	1	1	173	178.037	35	33.450
29	1	1	1	0	0	11	8.598	18	19.062
30	1	1	1	0	1	61	61.190	136	130.759
31	1	1	1	1	0	28	27.743	32	32.766
32	1	1	1	1	1	298	295.702	308	309.731

metric literature. This section presents estimates of underlying ability distributions through the methods described above, using previously-published item parameter estimates.

The data from LSAT-6 have been found to be in accord with assumptions of a one-parameter logistic (Rasch, 1980) item response model and an underlying normal distribution of ability (see Bock and Lieberman, 1970; Andersen, 1972; Andersen and Madsen, 1977; Bock and Aitkin, 1981; Andersen's 1972) conditional maximum likelihood estimates of the item parameters are −1.256, .476, 1.236, .168, and −.623. Taking these values as known, marginal maximum likelihood estimates of the mean and variance of an underlying normal distribution of ability have been computed from the data from all five items, and from items 1 through 4, 1 through 3, and 1 and 2, as well. In each solution, 20 fixed quadrature points between −4.75 and +4.75 in steps of .50 were employed. The results are presented in Table 2.

Table 2 contains the maximum likelihood estimates for the parameters in question and their associated standard errors. It may be seen that the solution for items 1 through 5—a mean of 1.475 and a variance of .551—agrees with Andersen and Madsen (1977), in which the same likelihood equation was maximized. The expected response pattern counts in this solution are also shown in Table 1. The resulting chi-square test against the general multinomial alternative yields a chi-square of 21.799, which agrees with that obtained by Bock and Aitkin (1981). Had the item parameters been specified a priori rather than estimated from the same data, the degrees of freedom would be 27, or the number of non-zero response patterns (30) minus the number of population parameters estimated (2) minus one.

The last three columns of Table 2 may be used to study the effects on the estimation of the mean and variance that result from the second stage sampling, i.e., basing estimates not on observed values of θ but on samples of responses given values of θ. The column labeled "observed information" is the approximation based on (9). It will be seen that these values are not the squared reciprocals of the standard errors, because the estimates of the mean and variance in the "incomplete data" problem are not uncorrelated as they are in the "complete data" problem. The column labeled "complete data information" contains the values given by $\sigma^2/1000$ and $2\sigma^4/1000$ for means and variances respectively, indicating the precision of estimation that would be obtained if 1,000 values of θ had been observed from a normal distribution with variance equal to that estimated from the corresponding item subset. The final column, "effective sample size," is 1,000 times the ratio of the preceeding two columns. This value is an estimate of the information about the parameter in question contained in the item responses, in terms of the number of observed

TABLE 2

PARAMETER ESTIMATES FOR LSAT-6 DATA

	ITEMS	MLE	S.E.	OBSERVED INFORMATION	COMPLETE DATA INFORMATION	EFFECTIVE SAMPLE SIZE
MEAN:	1-5	1.475	.047	516.323	1814.882	284
	1-4	1.478	.050	458.887	1760.563	261
	1-3	1.492	.058	350.512	1497.006	234
	1-2	1.490	.098	213.949	1594.896	134
VARIANCE:	1-5	.551	.099	117.920	1646.898	72
	1-4	.568	.117	84.516	1549.792	55
	1-3	.668	.161	45.509	1120.513	41
	1-2	.667	.306	21.737	1271.847	17

TABLE 3

OBSERVED AND EXPECTED NUMBER-CORRECT DISTRIBUTION FOR LSAT-6

	SCORE							
	0	1	2	3	4	5	CHI-SQUARE	DF
OBSERVED	3.0	20.0	85.0	237.0	357.0	298.0		
EXPECTATION UNDER BETA-BINOMIAL	2.4	20.9	88.5	228.1	364.2	296.0	.83	3
EXPECTATION UNDER NORMAL-RASCH*	2.2	20.3	88.1	228.7	367.1	293.8	1.03	3

*From Andersen and Madsen (1977)

values of θ that would be required to match the precision of estimation. It is clear from these values that the loss of precision in estimation is less serious for the mean than for the variance.

The fit of the beta-binomial model to the total score distribution of LSAT-6 is shown in Table 3. The maximum likelihood estimates of α and β are 15.55 and 4.81, with standard errors of 4.05 and 1.25 respectively. Andersen and Madsen's (1977) fit of the normal distribution with the Rasch item response model is also shown in Table 3. That both models fit extremely well serves as a reminder that a meager data set provides little power to reject incorrect models or to distinguish among competing models.

The data from LSAT-7 have been found to be in accord with a two-factor model of the following form:

$$P(x_{ij} = 1) = \Phi(Z_{ij})$$

where

$$Z_{ij} = c_j + a_{j1}\theta_{i1} + \alpha_{j2}\,\theta_{i2}$$

and (θ_1, θ_2) has a bivariate normal distribution (see Christofferson, 1975; Muthén, 1978: and Bock and Aitken, 1981). Bock and Aitkin's (1981) varimax-rotated two factor solution, consisting of item intercepts and slopes (i.e., factor loadings), is shown in Table 4.

TABLE 4

ESTIMATED FACTOR MEANS AND COVARIANCE MATRIX FOR LSAT-7 DATA

		FACTOR LOADINGS			
ITEM	INTERCEPT	DIMENSION 1		DIMENSION 2	
1	1.6177	1.354		.265	
2	.4722	.231		.538	
3	1.4303	.388		1.551	
4	.2938	.345		.284	
5	1.0902	.330		.250	
ESTIMATED MEANS:		-0.006	(.116)	.013	(.088)
ESTIMATED SIGMA:		1.009	(.291)	-.009	(.203)
		-.009	(.203)	1.036	(.222)

Taking these values as known, it is possible to estimate the mean vector and covariance matrix of the assumed underlying bivariate normal distribution.

This solution, also shown in Table 4, was obtained through the use of a 10 × 10 grid of Gauss-Hermite quadrature points. As expected, the solution consists of essentially zero means and an identity covariance matrix; that is, the distribution assumed in the estimation of the item parameters. The expected counts of response patterns are shown in Table 1. The chi-square text against the general multinomial alternative yields a value of 21.245 (again matching Bock and Aitkin). Had the term parameters been specified a priori, degrees of freedom would be 26, or the number of response patterns with non-zero observations (32) minus the number of population parameters estimated (5) minus one.

Example 2: The Vandenberg-Shepard Mental Rotations Test

As early as 1943, Johnson O'Connor had noticed a sex effect in spatial ability tasks, such that only about one-fourth of the women in a given population would score above the median of the men on a spatial abilities test (O'Connor, 1943). He speculated that this effect might be due to a X-linked recessive major gene for spatial visualization, in which case the proportion of females exhibiting the trait would be the square of the proportion of males exhibiting it. Proportions of .25 and .50 for females and males, in particular, would be consistent with O'Connor's conjecture. Since O'Connor's work, some other studies have lent support to this hypothesis (e.g., Bock and Kolakowski, 1973), but other have not (e.g., Balker et al., 1981).

In this section, data gathered by Zimowski (in progress) in dissertation research on sex difference in spatial ability tasks are analyzed. Responses from 158 male and 175 female seniors from York High School in Elmhurst, Illinois were obtained to the items of the Vandenberg-Shepard Mental Rotations Test. Each item in this 20-item test depicts a primary object, then asks the subject to identify which two of four other objects could be the same as the primary object, as seen from a different point of view. Table 5 presents the distribution of observed scores.

Using marginal maximum likelihood methods, the underlying distributions of ability in the male and female populations may be resolved into one, two or three normal components, and the fit of the solutions may be examined via the likelihood ratio test against the general multinomial alternative. The items will be assumed homogeneous, so that a binomial item response model may be employed. The logit scale will be assumed for θ in order to avoid the floor and ceiling effects of the bounded percent-correct variable. The probability of obtaining r correct responses, then, for a subject with latent ability θ is given as

$$P(r \mid \theta) = \frac{20!}{r!(20-r)!} p(\theta)^r [1 - p(\theta)]^{20-r}$$

where

$$p(\theta) = \exp(\theta)/[1 + \exp(\theta)].$$

The resultant model for the observed score distribution thus corresponds to a three-stage experiment. First, a component is selected in accordance with the proportion of the population that that component represents. Second, a value of θ is drawn at random from the normal distribution comprising the selected component. Finally, random item responses are obtained in accordance with the binomial model with a probability determined by the selected θ.

Table 5 presents, along with the observed score counts, their expectations under the assumptions of 1, 2, and 3 components—all estimated separately within sexes. The esti-

TABLE 5

OBSERVED AND EXPECTED FREQUENCIES FOR VANDENBERG DATA

	MALES				FEMALES			
		EXPECTATION				EXPECTATION		
SCORE	OBSERVED	K=1	K=2	K=3	OBSERVED	K=1	K=2	K=3
1	0	1.156	0.83	0.350	1	2.105	0.879	0.854
2	2	3.239	2.755	1.981	6	7.166	4.819	4.750
3	6	5.567	5.297	5.384	12	13.717	12.843	12.772
4	11	7.637	7.777	9.368	24	19.458	22.141	22.117
5	10	9.215	9.730	11.783	22	22.783	27.751	22.776
6	10	10.276	10.996	11.661	29	23.285	27.051	27.111
7	11	10.932	11.632	10.096	23	21.486	21.673	21.753
8	11	11.289	11.786	9.050	16	18.301	15.335	15.392
9	8	11.402	11.609	9.616	11	14.608	10.718	10.712
10	12	11.335	11.221	11.330	8	11.034	8.288	8.233
11	11	11.140	10.706	12.740	3	7.925	7.077	7.034
12	13	10.780	10.119	12.587	11	5.409	5.999	6.006
13	13	10.187	9.498	10.727	2	3.489	4.597	4.637
14	7	9.391	8.868	8.189	5	2.103	3.030	3.064
15	6	8.490	8.227	6.379	2	1.166	1.673	1.684
16	6	7.521	7.529	6.016	0	0.582	0.757	0.752
17	7	6.446	6.677	6.528	0	0.254	0.274	0.266
18	6	5.207	5.545	6.457	0	0.093	0.076	0.072
19	5	3.773	4.059	4.845	0	0.027	0.015	0.014
20	3	2.212	2.338	2.358	0	0.005	0.002	0.002
21	0	0.806	0.800	0.553	0	0.001	0.000	0.000

mated parameters of these solutions are shown in Table 6. It may be seen that the data for the females are in good accord with the major gene hypotheses. A two-component model fits the data well—significantly better than a one-component model—while a three-component model offers little additional improvement. Moreover, the estimated proportions in the two components are 79 percent for the lower and 21 percent for the higher, very close to the predicted proportions of 75 percent and 25 percent. The interpretation of results for males is less straightforward. To begin with, a one-component model offers a very good fit to the data, with a chi-square of 10.8 on 17 degrees of freedom. A two-component model offers no significant improvement in fit, but a three-component model decidely does. It is likely that the three-component model represents overfitting, but it is interesting to see that the lower two components in the three-component solution

TABLE 6

PARAMETER ESTIMATES FOR VANDENBERG DATA

	P_1	P_2	P_3	μ_1	μ_2	μ_3	σ	χ^2	DF
MALES:	1.000	-	-	-0.136	-	-	1.002	10.794	17
	.670	.330	-	-0.611	.850	-	.713	10.030	15
	.357	.466	.177	-1.220	0.095	1.514	.105	4.626	13
FEMALES:	1.000	-	-	-0.964	-	-	.578	20.251	13
	.794	.206	-	-1.200	.030	-	.157	13.624	11
	.679	.128	.193	-1.244	-0.874	.064	.112	13.543	9

have means virtually identical to the means in the females' two-component solution and account in roughly equal proportions for most of the males' distribution.

REFERENCES

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34*, 42–54.

Andersen, E. B. and Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika, 42*, 357–374.

Balker, J. T., Krasnoff, A. G. and Peaco, D. (1981). Visuo-spatial perception in adolescents and their parents: the X-linked recessive hypothesis. *Behavior Genetics, 11*, 403–413.

Bock, R. D. (1983). The Discrete Bayesian. In H. Wainer and S. Messick (Eds.), *Principals of Modern Psychological Measurement*. Hillsdale, New Jersey: Erlbaum.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D. and Kolakowski, D. F. (1973). Further evidence of sex-linked major gene influence on human spatial visualizing ability. *American Journal of Human Genetics, 25*, 1–14.

Bock, R. D. and Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*, 179–197.

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5–32.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika, 56*, 463–473.

Deely, J. J. and Lindley, D. V. (1981). Bayes empirical Bayes. *Journal of the American Statistical Association, 76*, 833–841.

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

Dempster, A. M., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in Covariance Components Models. *Journal of the American Statistical Association, 76*, 341–353.

Kendall, M. and Stuart, A. (1979). *The Advanced Theory of Statistics, Volume II*. New York: Macmillan.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73*, 805–811.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (An empirical Bayes problem). *Psychometrika, 34*, 259–299.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44*, 226–233.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551–560.

O'Connor, J. (1948). *Structural Visualization*. Boston: Human Engineering Laboratory.

Ramsay, J. O. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika, 40*, 361–372.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research, 1960. Chicago: University of Chicago Press, (reprint).

Rigdon, S. E. and Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48*, 567–574.

Sanathanan, L. and Blumenthal, N. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association, 73*, 794–798.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 37*, 222–239.

Stroud, A. H. and Sechrest, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, New Jersey: Prentice-Hall.

Zimowski, M. (in progress). *Implications of item difficulties for visuo-spatial information processing*. Doctoral dissertation, Department of Behavioral Sciences, University of Chicago.

Manuscript received 3/28/83
Final version received 5/2/84