Bayes Modal Estimates of a Discrete Latent Variable Distribution in Item Response Models Using the EM Algorithm

Bradley A. Hanson
ACT, Inc.

Abstract

There are typically two sets of parameters used in models of the observed distribution of item responses in a population: 1) item parameters, and 2) parameters of the distribution of the latent variable measured by the items. Several authors have presented Bayesian estimates of the item parameters. This paper presents Bayes modal estimates of a discrete latent variable distribution for item response models. The EM algorithm for computing maximum likelihood estimates of a discrete latent variable distribution is first reviewed, followed by a presentation of the EM algorithm for computing Bayes modal estimates using a Dirichlet prior for the discrete probabilities. Examples are presented comparing the maximum likelihood and Bayes modal estimates. A small simulation study is performed to examinee the performance of the Bayes modal estimates of a discrete latent variable distribution as applied to estimating average domain scores. The results of the simulation illustrate that whether Bayes modal estimates will have less error than maximum likelihood estimates depends on the tradeoff between the higher bias and lower variance of the Bayes modal estimates relative to the higher variance and lower bias of the maximum likelihood estimates. Bayes modal estimates are more likely to perform well relative to maximum likelihood estimates when sample sizes are small and the variance of the maximum likelihood estimates is large.

Models of observed item responses in a population of examinees are typically written to contain two types of parameters (Holland, 1981; Holland 1990): 1) item parameters that determine the conditional probabilities of each possible response to an item conditioned on the latent variable measured by the set of items on the test; and 2) the parameters of the distribution of the latent variable in the population. Several authors have presented Bayesian estimates of item parameters in item response models (Mislevy, 1986; Tsutakawa & Lin, 1986; Harwell & Baker, 1991; Tsutakawa, 1992; Zeng, 1997). This paper presents Bayes modal estimates of a discrete latent variable distribution for item response models.

The general item response model considered in this paper is presented first, followed by a review of using the EM algorithm to compute maximum likelihood (ML) estimates of a discrete latent variable distribution. Bayes modal estimates of a discrete latent variable distribution using the EM algorithm are then presented. Examples are given comparing maximum likelihood estimates to Bayes modal estimates using several prior distributions. A small simulation study is conducted to examine the performance of Bayes modal estimates of a latent variable distribution as applied to estimating average domain scores.

Item Response Model

The data to be modeled are the responses of $i = 1, \ldots, n$ examinees, randomly sampled from a population of examinees, to a fixed non-random set of $j = 1, \ldots, J$ items. The responses of the n examinees to the J items are contained in a $n \times J$ matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_n)^t$, where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})$ is a $1 \times J$ row vector that contains the responses of the ith randomly sampled examinee to the J fixed items. It is assumed that the set of responses to each item is finite (the item responses can be dichotomous or polytomous).

In addition to the observed item responses, there is a realization of a latent ability random variable Θ for each randomly sampled examinee. Unlike the realization of the item responses, the realization of Θ for the ith randomly sampled examinee (denoted θ_i) is not observed. This paper considers the case of a real-valued Θ.

It would be most natural to consider the distribution of Θ to be continuous, but in this paper the latent variable is taken to be discrete, and estimation procedures are derived based on the discrete latent variable. It is assumed the latent random variable Θ takes on K known discrete values $q_k, k = 1, \ldots, K$, with associated unknown probabilities $\pi_k, k = 1, \ldots, K$. This is in contrast to deriving estimation procedures based on a continuous latent variable and then implementing approximations of those procedures with a discrete version of the continuous latent variable (e.g., Bock and Aitken, 1981). The discrete distribution is taken to be a general multinomial distribution, so it can assume any shape. A continuous latent variable distribution would need to be assumed to belong to a parametric family of distributions since a nonparametric continuous latent variable distribution could not be identified from the observed discrete data.

The distribution of the observed item responses is modeled as

$$
\begin{aligned}
f(\mathbf{y} \mid \boldsymbol{\Delta}, \boldsymbol{\pi}) &= \sum_{k=1}^{K} f(\mathbf{y}, q_k \mid \boldsymbol{\Delta}, \pi_k) \\
&= \sum_{k=1}^{K} f(\mathbf{y} \mid q_k, \boldsymbol{\Delta}) \Pr(\Theta = q_k \mid \pi_k) \\
&= \sum_{k=1}^{K} f(\mathbf{y} \mid q_k, \boldsymbol{\Delta}) \pi_k \,,
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\Delta}$ represents the item parameters which determine the probability of a particular set of item responses occurring given a fixed value of $\boldsymbol{\Theta}$, and $f(\mathbf{y} \mid q_k, \boldsymbol{\Delta})$ is the conditional probability distribution of the item responses for examinees with a value of the latent variable equal to q_k. An assumption made in going from the first to the second line in Equation 1 is that the distribution of the item responses when conditioned on the item parameters and latent variable does not depend on the parameters of the latent variable distribution, and that the distribution of the latent variable conditioned on the parameters of the latent variable distribution does not depend on the item parameters. This assumption is made throughout the paper. The notational convention used in this paper is that $q_k, k = 1, \ldots K$ are the K possible values of the latent variable, whereas θ is a specific undetermined value of the latent variable (θ can be equal to any of the q_k).

The M step of the EM algorithm for computing maximum likelihood or Bayes modal estimates can be computed separately for $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ and $\boldsymbol{\Delta}$ (Woodruff and Hanson, 1996). Bayes modal estimates of $\boldsymbol{\Delta}$ for dichotomous item response models using the EM algorithm have been presented by several authors (Mislevy, 1986; Tsutakawa & Lin, 1986; Harwell & Baker, 1991; Tsutakawa, 1992; Zeng, 1997). This paper presents Bayes modal estimates of $\boldsymbol{\pi}$.

Maximum Likelihood Estimates

This section reviews computation of maximum likelihood estimates of $\boldsymbol{\pi}$ using the EM algorithm. The maximum likelihood estimates are the values of $\boldsymbol{\pi}$ that maximize the likelihood of the observed data ($\mathbf{Y}$) given the parameters $\boldsymbol{\pi}$ (the observed data likelihood). Several authors have presented maximum likelihood estimates of $\boldsymbol{\pi}$ using the EM algorithm (Mislevy, 1984; Titterington, Smith, and Makov, 1985; Woodruff and Hanson, 1996). The EM algorithm uses the complete data likelihood to find values of the parameters $\boldsymbol{\pi}$ which maximize the observed data likelihood (Dempster, Laird, Rubin, 1977). The observed data are $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})$ is the vector of realized item responses to the J items for examinee i. The missing data are $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$, where θ_i is the realized value of the latent variable for examinee i. The complete data are $[(\mathbf{y}_1, \theta_1), (\mathbf{y}_2, \theta_2), \ldots, (\mathbf{y}_n, \theta_n)]$. The complete data log-likelihood can be written as

$$
\log \left[\prod_{i=1}^{n} f(\mathbf{y}_i, \theta_i \mid \boldsymbol{\Delta}, \boldsymbol{\pi}) \right] = \sum_{i=1}^{n} \log[f(\mathbf{y}_i, \theta_i \mid \boldsymbol{\Delta}, \boldsymbol{\pi})]
$$

$$
= \sum_{i=1}^{n} \log[f(\mathbf{y}_i \mid \theta_i, \boldsymbol{\Delta}, \boldsymbol{\pi})] + \sum_{i=1}^{n} \log[f(\theta_i \mid \boldsymbol{\Delta}, \boldsymbol{\pi})]
$$

$$
= \sum_{i=1}^{n} \log[f(\mathbf{y}_i \mid \theta_i, \boldsymbol{\Delta})] + \sum_{i=1}^{n} \log[f(\theta_i \mid \boldsymbol{\pi})]. \tag{2}
$$

In the last line of Equation 2 the parameters $\boldsymbol{\pi}$ only appear in the $\log[f(\theta_i \mid \boldsymbol{\pi})]$ terms. Consequently, for the purpose of estimating the parameters $\boldsymbol{\pi}$ the only portion of the log-likelihood that is relevant is

$$
\sum_{i=1}^{n} \log[f(\theta_i \mid \boldsymbol{\pi})]. \tag{3}
$$

Since the latent variable is assumed to be discrete, the distribution $f(\theta \mid \boldsymbol{\pi})$ is multinomial with parameters $\boldsymbol{\pi}$ and n. The log-likelihood in Equation 3 is for a sample from a multinomial distribution. If the missing data $(\theta_1, \theta_2, \ldots, \theta_n)$ were known then the maximum likelihood estimate of π_k would be n_k/n, where n_k is the number of the n examinees for whom the value of the latent variable is equal to q_k (the n_k are complete-data sufficient statistics for the parameters of the multinomial distribution). The EM algorithm is easy to apply in this case since the maximum likelihood estimates of π_k are easy to compute using the complete data.

Since the multinomial distribution is a member of the regular exponential family of distributions the EM algorithm consists of an E step in which the expected values of the complete data sufficient statistics (the n_k) are computed over the distribution of the missing data conditioned on the observed data and provisional values of the parameters, and an M step in which the expected values of the sufficient statistics computed in the E step are used in place of the unobserved n_k to compute complete data maximum likelihood estimates of the parameters. For computing estimates of π_k the E step at iteration $s = 0, 1, \ldots$ consists of computing

$$
n_k^{(s)} = E(n_k \mid \mathbf{Y}, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}) \tag{4}
$$

where $\boldsymbol{\pi}^{(s)} = (\pi_1^{(s)}, \pi_2^{(s)}, \ldots, \pi_K^{(s)})$ are parameters computed at iteration $s - 1$ (or starting values for $s = 0$), and the expectation is over the distribution of the missing data given $\mathbf{Y}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\pi}^{(s)}$. The quantity in Equation 4 can be expressed as (Woodruff and Hanson, 1996)

$$
n_k^{(s)} = E(n_k \mid \mathbf{Y}, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}) = \sum_{i=1}^{n} p(q_k \mid \mathbf{y}_i, \boldsymbol{\Delta}, \boldsymbol{\pi}^{(s)}) = \sum_{i=1}^{n} \frac{f(\mathbf{y}_i \mid q_k, \boldsymbol{\Delta})\pi_k^{(s)}}{\sum_{k'=1}^{K} f(\mathbf{y}_i \mid q_k, \boldsymbol{\Delta})\pi_{k'}^{(s)}}, \tag{5}
$$

where $p(q_k \mid \mathbf{y}_i, \mathbf{\Delta}, \boldsymbol{\pi}^{(s)})$ is the conditional probability that $\Theta_i = q_k$ given fixed known values $\mathbf{y}_i$, $\mathbf{\Delta}$, and $\boldsymbol{\pi}^{(s)}$ (this is the posterior distribution of the latent variable given $\mathbf{y}_i$, for fixed values of the parameters $\mathbf{\Delta}$, and $\boldsymbol{\pi}^{(s)}$).

The M step at iteration s consists of using the $n_k^{(s)}$ computed in the E step to compute

$$\pi_k^{(s+1)} = \frac{n_k^{(s)}}{n} \, . \tag{6}$$

The values of $\pi_k^{(s+1)}$ computed in the M step at iteration s are used in Equation 5 in the E step at iteration $s + 1$. The E steps and M steps are repeated until convergence occurs. One convergence criterion is the relative difference in values of the observed data likelihood evaluated at the parameter estimates in consecutive iterations (the observed data likelihood evaluated at the parameter estimates will always increase on consecutive iterations). Another convergence criterion is the relative difference in parameter estimates in consecutive iterations.

These maximum likelihood estimates are computed by the BILOG computer program (Mislevy & Bock, 1990), although they are labeled "posterior weights" in the BILOG output. Note that even though posterior distributions are used in the E step (Equation 5), the estimates produced are maximum likelihood estimates not Bayesian estimates.

Bayes Modal Estimates

The EM algorithm can be used to compute Bayes modal estimates (Dempster, Laird, & Rubin, 1977, pp. 6; Tanner, 1996). The Bayes modal estimate of $\boldsymbol{\pi}$ is the value that maximizes the observed posterior distribution (the posterior distribution of the parameters given the observed data). The EM algorithm for computing Bayes modal estimates uses the complete data posterior (or the augmented posterior) to compute estimates of the mode of the observed data posterior (Tanner, 1996). The complete data posterior is proportional to the product of the complete data likelihood and the prior distributions of $\boldsymbol{\pi}$ and $\mathbf{\Delta}$. The constant of proportionality can be ignored for the purposes of computing Bayes modal estimates of $\mathbf{\Delta}$ and $\boldsymbol{\pi}$. The Bayes modal estimates of $\mathbf{\Delta}$ and $\boldsymbol{\pi}$ using the complete data posterior are the values that maximize

$$\log \left[g(\boldsymbol{\pi}) h(\mathbf{\Delta}) \prod_{i=1}^{n} f(\mathbf{y}_i, \theta_i \mid \mathbf{\Delta}, \boldsymbol{\pi}) \right] = \log[g(\boldsymbol{\pi})] + \log[h(\mathbf{\Delta})] + \sum_{i=1}^{n} \log[f(\mathbf{y}_i, \theta_i \mid \mathbf{\Delta}, \boldsymbol{\pi})]$$

$$= \sum_{i=1}^{n} \log[f(\mathbf{y}_i \mid \theta_i, \mathbf{\Delta}, \boldsymbol{\pi})] + \log[h(\mathbf{\Delta})] + \sum_{i=1}^{n} \log[f(\theta_i \mid \mathbf{\Delta}, \boldsymbol{\pi})] + \log[g(\boldsymbol{\pi})]$$

$$= \sum_{i=1}^{n} \log[f(\mathbf{y}_i \mid \theta_i, \mathbf{\Delta})] + \log[h(\mathbf{\Delta})] + \sum_{i=1}^{n} \log[f(\theta_i \mid \boldsymbol{\pi})] + \log[g(\boldsymbol{\pi})], \quad (7)$$

where $f(\mathbf{y}_i, \theta_i \mid \mathbf{\Delta}, \boldsymbol{\pi})$ is the complete data likelihood for examinee i, $h(\mathbf{\Delta})$ and $g(\boldsymbol{\pi})$ are the prior distributions of $\mathbf{\Delta}$ and $\boldsymbol{\pi}$, and the prior distributions of $\mathbf{\Delta}$ and $\boldsymbol{\pi}$ are assumed to be independent. In the last line of Equation 7 the parameters $\boldsymbol{\pi}$ only appear in the terms $\log[f(\theta_i \mid \boldsymbol{\pi})]$ and $g(\boldsymbol{\pi})$, and the parameters $\mathbf{\Delta}$ only appear in the terms $\log[f(\mathbf{y}_i \mid \theta_i, \mathbf{\Delta})]$ and $\log[h(\mathbf{\Delta})]$. Thus, Bayes modal estimates of $\mathbf{\Delta}$ and $\boldsymbol{\pi}$ can be found independently. For the purpose of estimating the parameters $\boldsymbol{\pi}$ only the portion of Equation 7 that depends on $\boldsymbol{\pi}$ needs to be considered:

$$\sum_{i=1}^{n} \log[f(\theta_i \mid \boldsymbol{\pi})] + \log[g(\boldsymbol{\pi})] \, . \tag{8}$$

Since the θ_i take on a discrete set of values, the distribution $f(\theta_i \mid \boldsymbol{\pi})$ is a multinomial distribution (with parameters n and $\boldsymbol{\pi}$). The values of $\boldsymbol{\pi}$ that maximize Equation 8 are the Bayes modal estimates of $\boldsymbol{\pi}$.

The conjugate family of prior distributions for a multinomial distribution is the Dirichlet (Bishop, Fienberg, and Holland, 1975). If the prior distribution of $\boldsymbol{\pi}$ is Dirichlet, then the posterior distribution of $\boldsymbol{\pi}$ will also be Dirichlet. The probability distribution of the Dirichlet is

$$f(\boldsymbol{\pi} \mid \boldsymbol{\beta}) = \Gamma \left(\sum_{k=1}^{K} \beta_k \right) \prod_{k=1}^{K} \frac{\pi_k^{\beta_k - 1}}{\Gamma(\beta_k)} \, , \tag{9}$$

3

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$, $\beta_k > 0$ for all k, and $\Gamma(x)$ is the gamma function. If the prior distribution of the parameters of a multinomial distribution is Dirichlet with parameters $\boldsymbol{\beta}$ then the posterior distribution is Dirichlet with parameters $\beta_k + n_k$, $k = 1, 2, \ldots, K$ (Bishop, Fienberg, and Holland, 1975), where n_k is the number of the n examinees for whom the value of the latent variable is equal to q_k (the n_k are complete data sufficient statistics for the parameters of the multinomial distribution). The prior mean of π_k is β_k / n_β, where

$$n_\beta = \sum_{k=1}^{K} \beta_k \,. \tag{10}$$

The posterior mean of π_k is

$$\frac{n}{n + n_\beta} \frac{n_k}{n} + \frac{n_\beta}{n + n_\beta} \frac{\beta_k}{n_\beta} \,. \tag{11}$$

Hence, the posterior mean of π_k is a weighted average of the prior mean of π_k and the maximum likelihood estimate of π_k (n_k/n) with weights proportional to the observed sample size (n) and the "prior sample size" (n_β).

It is convenient to choose the prior distribution by choosing values of n_β (denoted the prior sample size) and β_k / n_β (denoted the prior means). The prior means would be chosen to give the multinomial distribution thought to be most likely, and the prior sample size would be chosen to represent the strength of belief in the accuracy of the prior means in representing the true multinomial distribution. A noninformative prior is given by setting β_k equal to one for all k. The noninformative prior gives equal weight to all multinomial probabilities. When a noninformative prior is used the Bayes modal estimate of $\boldsymbol{\pi}$ is equal to the maximum likelihood estimate.

The Bayes modal estimate of $\boldsymbol{\pi}$ is the mode of the Dirichlet posterior distribution. If $n_k + \beta_k > 1$ for all k then the posterior distribution will have a mode. The value of π_k at the mode of the posterior distribution is (Bernardo and Smith, 1994):

$$\frac{n_k + \beta_k - 1}{n + n_\beta - K} \,. \tag{12}$$

If the missing data sufficient statistics (n_k) were known the Bayes modal estimates of the multinomial probabilities would be given by Equation 12.

Since the multinomial distribution is a member of the regular exponential family of distributions the E step of the EM algorithm consists of computing the expected values of the missing data sufficient statistics (the n_k) over the distribution of the missing data conditioned on the observed data and provisional values of the parameters (Dempster, Laird, & Rubin, 1977). The E step of the EM algorithm at iteration $s = 0, 1, \ldots$ is the same as the E step used in computing maximum likelihood estimates using the EM algorithm given by Equation 5.

In the M step the expected values of the sufficient statistics computed in the E step are substituted for the sufficient statistics in the formula to compute complete data Bayes modal estimates of the parameters. The M step at iteration s consists of using the $n_k^{(s)}$ computed in the E step to compute

$$\pi_k^{(s+1)} = \frac{n_k^{(s)} + \beta_k - 1}{n + n_\beta - K} \,. \tag{13}$$

The only difference between the EM algorithms for computing maximum likelihood and Bayes modal estimates is that maximum likelihood uses Equation 6 for the M step calculation, whereas for Bayes modal estimates Equation 13 is used for the M step calculation. The E steps and M steps are repeated until convergence occurs. One convergence criterion is the relative difference in values of the observed data posterior evaluated at the parameter estimates in consecutive iterations (the observed data posterior evaluated at the parameter estimates will always increase on consecutive iterations). Another convergence criterion is the relative difference in parameter estimates in consecutive iterations.

A potential problem that can exist in computing the Bayes modal estimates is that the mode of the posterior will not exist if $n_k^{(s)} + \beta_k < 1$ for at least one k. The posterior mode may or may not exist if at least one $\beta_k < 1$, which implies that the prior distribution approaches infinity at one or more of the boundaries,

4

and has no mode. It can be assured that the posterior mode exists if $\beta_k \geq 1$ for all k. In cases where n_β is small it may not be feasible to have $\beta_k \geq 1$ for all k. When $\beta_k < 1$ for some k an alternative to using the mode of the posterior distribution given in Equation 12 (which may not exist) in the M step is to use the posterior mean given by

$$\pi_k^{(s+1)} = \frac{n_k^{(s)} + \beta_k}{n + n_\beta} . \tag{14}$$

Using the posterior mean (Equation 14) rather than the posterior mode (Equation 13) in the M step will guarantee the M step calculation can always be performed when one or more of the β_k are less than one. When using the posterior mean in the M step the algorithm is no longer an EM algorithm, it is not assured that the algorithm will converge, and if it does converge it is unclear what the properties of the estimates are. The EM algorithm using Equation 13 for the M step results in a series of estimates that converge to the mode of the observed data posterior distribution. When Equation 14 is used in place of Equation 13 in the M step it not the case that a sequence of estimates is generated that converge to the mean of the observed data posterior distribution.

Experience in using Equation 14 in the M step (some of which is reported below) indicates that the resulting sequence of estimates seems to converge and provide a reasonable answer. Consequently, when some of the β_k are less than one using Equation 14 in the M step may be a reasonable, though ad hoc, method of obtaining estimates.

Examples

This section uses simulated data to present some examples of the Bayes modal estimates presented in the previous section. The data used are the responses of 83,690 examinees to a form of the ACT Reading test (40 dichotomously scored items). The program EM1 (Zeng, 1995; Zeng, 1997) was used with these data to estimate three-parameter logistic model item parameters for the 40 items. Two discrete latent variable distributions were specified as population distributions and for each of the two distributions the estimated item parameters were used to generate item responses for 100 simulated examinees. The simulated observations were used to compute maximum likelihood and Bayes modal estimates of the latent variable distribution. The item parameters used to simulate the observations were also used in estimating the latent variable distributions (the population item parameters were used when estimating the latent variable distributions). The estimated distributions are compared to the population distributions from which the data were generated.

The first population distribution used was a discrete version of a standard normal distribution (mean zero, standard deviation 1) using 50 equally spaced points from -3 to 3 (inclusive). This latent variable distribution was used along with the item parameters to generate 100 simulated sets of responses to the 40 items ($n = 100$). The simulated item response data were used to compute a maximum likelihood estimate of the latent variable distribution and six Bayes modal estimates of the latent variable distribution. Both the maximum likelihood and Bayes modal estimates were computed over 20 equally spaced points from -3 to 3 ($K = 20$).

Fewer discrete points were used for the estimated distribution than were specified for the population distribution in order to increase the chances that a posterior mode would exist in the M step calculation. For a given prior sample size using fewer discrete points in the latent variable distribution (smaller values of K) will increase the values of β_k and increase the chance that a posterior mode will exist.

Three of the Bayes modal estimates used a discrete version of a standard normal distribution as the prior means (β_k/n_β) with prior samples sizes (n_β) of 25, 50 and 100. The other three Bayes modal estimates used priors in which the prior means were all equal (the prior means formed a uniform distribution) with prior sample sizes of 25, 50, and 100. Even though the prior means were equal this is not a noninformative Dirichlet prior since the $\beta_k = \beta > 1$ for all k. Setting the prior means equal with $\beta > 1$ indicates the most likely distribution is thought to be a uniform multinomial (a noninformative prior would indicate that all possible multinomial distributions are considered equally likely).

Figure 1 presents the population distribution, maximum likelihood estimates and Bayes modal estimates. The top plot in Figure 1 contains the population distribution, maximum likelihood estimate, and the three Bayes modal estimates using the normal prior means. The bottom plot in Figure 1 contains the population distribution, maximum likelihood estimate, and the three Bayes modal estimates using the uniform prior

5

means (the population and maximum likelihood estimates are the same in the two plots in Figure 1). Since the population distribution and estimated distributions use different numbers of discrete points the probabilities plotted were adjusted so the frequency polygons shown had the same area for all distributions.

The second population distribution was a discrete version of a four-parameter beta distribution with shape parameters 4 and 2, lower limit -3.5, and upper limit 3.05, using 50 equally spaced points from -3 to 3. The limits of the beta distribution are outside the limits of the discrete population distribution so that the lowest and highest points of the discrete distribution have non-zero probability. This distribution is negatively skewed and flatter than the standard normal used for the first population distribution. One hundred simulated sets of item responses were generated using this population distribution ($n = 100$). The simulated data were used to compute maximum likelihood and Bayes modal estimates of the latent variable distribution using 20 equally spaced points from -3 to 3 ($K = 20$).

Six Bayes modal estimates were calculated. For three of the Bayes modal estimates a discrete version of a standard normal distribution (mean 0 and standard deviation 1) was used for the prior means with prior sample sizes of 25, 50 and 100. The other three Bayes modal estimates used as the prior means a discrete version of a normal distribution with mean 1 and standard deviation 1.5 with prior sample sizes of 25, 50 and 100. Since the discrete points used were from -3 to 3 the normal prior with a mean of 1 was truncated for high latent variable values more than for low latent variable values so that the distribution was negatively skewed. Results for the second population distribution are presented in Figure 2. The top plot in Figure 2 gives results for the prior means based on a normal distribution with mean 1 and standard deviation 0, and the bottom plot gives results for the prior means based on a normal with mean 1 and standard deviation 1.5.

The posterior mean (Equation 13) was used in the M step calculation of the Bayes estimates shown in Figures 1 and 2 for all cases except the three estimates computed using the uniform prior means for the first population distribution. The posterior mean was used in the M step calculation in these cases due to at least one of the parameters of the Dirichlet prior being less than one which could result in the posterior mode not existing. In all cases in which the prior mean was used the observed posterior evaluated at the parameter estimates increased at each iteration, as would be the case if the posterior mode were used, and the estimates converged (the relative difference in consecutive values of the observed posterior eventually became smaller than a cutoff of 10^{-5}). The term "Bayesian estimates" will be used when referring to estimates that may be computed using the posterior mean or mode in the M step calculation.

Figures 1 and 2 show how the Bayesian estimates are closer to the maximum likelihood estimates for smaller values of the prior sample size and are closer to the prior means for larger values of the prior sample size.

Application to Computing Average Domain Scores

This section presents a small simulation study that examines the performance of Bayesian estimates of the latent variable distribution as used in estimating average domain scores for groups of examinees. A domain score is a mean item score over a domain of items. An average domain score is the mean domain score for a group of examinees. Pommerich and Nicewander (1996) discuss several methods using item response theory to estimate the average domain scores when examinees take only a subset of items in the domain. The method presented in this paper uses an estimate of the latent variable distribution along with estimates of the item parameters for all items in the domain to compute an estimate of the average domain score.

For dichotomous items the average domain score is the mean of the expected item scores (p-values) for the group of interest. The p-value of item j is given by

$$\sum_{k=1}^{K} \Pr(Y_j = 1 \mid q_k, \delta_j)\pi_k \,, \tag{15}$$

where $\Pr(Y_j = 1 \mid q_k, \delta_j)$ is the probability of answering item j correctly given $\Theta = q_k$ with the item parameters for item j given by δ_j. The average of the p-values over the J items in the domain (the average domain score) is

$$\frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{K} \Pr(Y_j = 1 \mid q_k, \delta_j)\pi_k \,. \tag{16}$$

6

To compute an average domain score requires estimates of $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$.

This section presents a small simulation study in which it is assumed estimates of $\boldsymbol{\Delta}$ are available for all items in the domain, and Bayesian and maximum likelihood estimates of $\boldsymbol{\pi}$ are obtained using responses to a sample of examinees to a subset of items in the domain. Simulation is used to assess the performance of different methods of estimating $\boldsymbol{\pi}$ on estimates of the average domain score using Equation 16.

The data used to obtain item parameters are the responses of 83,690 examinees to a form of the ACT Mathematics test (60 dichotomously scored items). The program EM1 (Zeng, 1995; Zeng, 1997) was used with these data to estimate three-parameter logistic model item parameters for the 60 items. The 60 items were considered the domain and the 10 items given by the first item and every sixth item thereafter (items $1, 7, 13, \ldots, 55$) were considered administered items. Estimates of $\boldsymbol{\pi}$ were obtained using only the 10 administered items. The 10 administered items were slightly easier than the 60 items in the domain.

Simulations were performed using two population latent variable distributions and two sample sizes (25 and 100). The first population distribution was a discrete version of a standard normal distribution on 50 equally spaced points in the interval -3 to 3 (inclusive). The second population distribution was a discrete version of a normal distribution with a mean of -.5 and standard deviation of 1 on 50 equally spaced points from -3 to 3 (inclusive). The second population distribution had a slight positive skew since the midpoint of the interval [-3, 3] is not the mean of the distribution. A plot of the two population distributions is presented in Figure 3.

For each of the four conditions (two population distributions by two sample sizes) 1000 samples of the appropriate sample size were drawn from the appropriate population distribution. Thus, there were a total of 250,000 observations simulated ($[25 + 25 + 100 + 100] \times 1000$). Responses to the administered items were simulated for each observation. For each simulated sample the average domain score was estimated using ten methods. The first method used the average of the p-values for the 10 administered items as the estimated average domain score (this will be referred to as the observed average domain score). The second method used maximum likelihood to estimate the latent variable distribution ($\boldsymbol{\pi}$). The maximum likelihood estimates of $\boldsymbol{\pi}$ were used in Equation 16 to compute an estimate of the average domain score. The remaining eight methods used Bayes modal estimates of $\boldsymbol{\pi}$ to compute estimates of the average domain score given by Equation 16. The eight Bayes modal estimates consisted of a combination of two sets of prior means by four prior sample sizes (25, 50, 100, and 500).

One of the sets of prior means consisted of a discrete version of a standard normal distribution using 20 equally spaced points between -3 and 3 (inclusive). The second set of prior means was a discrete version of a four-parameter beta distribution with shape parameters 3 and 3, lower limit of -3.2, and upper limit of 3.2, using 20 equally spaced points between -3 and 3 The lower and upper limits of the beta distribution are outside the lower and upper limits of the prior means so that the lowest and highest prior means are greater than zero. The two sets of prior means are presented in Figure 4.

For each of the four conditions the ten estimates of the average domain score were compared based on the root mean square error (RMSE), absolute bias, and standard deviation computed over the 1000 replications. Let d_t be the domain score based on the population distribution (the true domain score). Then the RMSE is given by

$$\sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (d_i - d_t)^2}, \tag{17}$$

where d_i is the average domain score as computed using data from replication i. The absolute bias is given by

$$|\bar{d} - d_t|, \tag{18}$$

where

$$\bar{d} = \sum_{i=1}^{1000} d_i. \tag{19}$$

The standard deviation is given by

$$\sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (d_i - \bar{d})^2}. \tag{20}$$

7

The absolute bias is a measure of systematic error in a domain score estimate and the standard deviation is a measure of the random error in a domain score estimate. The RMSE is a combined measure of systematic and random error. The square of RMSE equals the sum of the squared standard deviation and squared absolute bias.

In all cases the posterior mean (Equation 14) was used in the M step rather than the posterior mode (Equation 13). This was due to the fact that for the priors chosen $\beta_k < 1$ for some of the k, so that the posterior mode may not exist. Unlike the two examples presented earlier there were cases in which the observed data posterior evaluated at the estimate did not increase from one iteration to the next. Consequently, the criterion used to judge convergence of the EM algorithm was the relative difference in the estimates of π from one iteration to the next. Iterations were continued until the relative difference between π_k in the current and previous iteration was less than 10^{-4} for all k.

Table 1 contains values of RMSE, absolute bias, and standard deviation as estimated from the simulation. The top portion of Table 1 give results for a sample size of 25, and the bottom portion of Table 1 give results for a sample size of 100. For each sample size the results are given for each of the two population distributions.

The RMSE of the maximum likelihood estimate is smaller than the observed estimates in every condition. This is primarily due to the maximum likelihood estimate having a much smaller bias than the observed estimate (the observed estimate is biased due to the administered items being somewhat easier than the group of all items in the domain). The Bayesian estimates have smaller RMSE than the maximum likelihood estimates for the population with mean 0. This is true for both sample sizes, both sets of prior probabilities, and all prior sample sizes. The advantage of the Bayesian estimates over the maximum likelihood estimate is greater for the sample size of 25 than for the sample size of 100.

The Bayesian estimates have larger RMSE than the maximum likelihood estimates for the population with mean -0.5. This is due to the large bias of the Bayesian estimates in this case. Both sets of prior means are centered at zero, but the mean of the population distribution is -0.5.

Increasing the prior sample size generally results in an increase in the bias and a decrease in the standard deviation for the Bayesian estimates. An exception is the case where the prior and population distributions are the same (normal prior means with the zero mean population distribution). When the prior and population distributions match, the bias, in addition to standard deviation, decreases with increasing prior sample size.

Discussion

This paper has presented details of using the EM algorithm to calculate Bayes modal estimates of a latent variable distribution in item response models. An advantage of the algorithm presented is its simplicity. The M step computation of the posterior mode is given by a simple formula (Equation 13). An iterative computational procedure is not needed for the M step calculation. In comparison, the M step calculation used for Bayes modal estimates of item parameters is much more complex (e.g., Zeng, 1997).

This paper discussed computation of estimates of the probabilities of a discrete latent variable distribution (π) assuming known values of the item parameters. It is possible in either maximum likelihood or Bayes modal estimation to use the EM algorithms discussed in this paper to compute an estimate of π at the same time an estimate of Δ is also being computed. It can be seen in the complete data likelihood given in Equation 2 and the complete data posterior given in Equation 7 that the parameters π and Δ can be independently maximized. Therefore, in the M step where both π and Δ are estimated the M step calculations for π discussed in this paper still apply independently of the M step calculation needed for Δ. There will be additional calculations needed in the E step when both π and Δ are being estimated since the $n_k^{(s)}$ will not be a complete set of sufficient statistics for Δ.

Some examples compared Bayesian estimates with various priors to maximum likelihood estimates. The examples illustrated how as the prior sample size is increased the estimated distribution is closer to the prior distribution and further from the maximum likelihood estimate. A small simulation study investigated the performance of Bayesian estimates of the latent variable distribution in estimating average domain scores. Bayesian estimates of the latent variable distribution resulted in more accurate estimates of the average domain score than maximum likelihood estimates when the prior was not too different from the population distribution (beta or normal prior means centered at zero with the population distribution based on a normal distribution with mean zero), although the Bayesian estimates were less accurate when the prior was not

as close to the population distribution (beta or normal prior means centered at zero with the population distribution based on a normal distribution with mean -0.5).

Bayesian estimates will tend to have reduced variance at the expense of some bias when compared to maximum likelihood estimates. As shown in the simulation as long as the bias in the Bayesian estimates is not too great they can contain less total error than maximum likelihood estimates (which have greater variance than the Bayesian estimates). The Bayesian estimates have the most potential to perform favorably relative to the maximum likelihood estimate for small sample sizes. For small sample sizes the variance of the maximum likelihood estimate will be large enough that the Bayesian estimates may have lower overall error (the smaller variance of the Bayesian estimates may compensate for their larger bias). Even with smaller samples sizes Bayesian estimates can perform worse that the maximum likelihood estimate if the bias of the Bayesian estimates is too large. This occurred in the simulation for the population distribution with mean -.05 where even with a sample size of 25 the standard deviation of the ML estimate was less than the bias of any of the Bayesian estimates.

A useful extension of the methods discussed in this paper would be empirical Bayes estimates of a discrete latent variable distribution. An example where empirical Bayes estimates may be useful is a situation in which latent variable distributions need to be estimated for each school in a group of schools (e.g., estimating average domain scores for a group of schools). The empirical Bayes approach would use data for all schools to estimate a prior that would be used to calculate Bayesian estimates for individual schools.

A problem that was noted for the Bayes modal estimates presented in this paper is that the mode of the Dirichlet prior will not exist if $n_k^{(s)} + n_\beta < 1$ in the M step calculation. This situation existed for almost all the Bayesian estimates that were computed in this paper. In those cases, the mean of the Dirichlet prior was substituted for the mode in the M step calculation. This seemed to work well for the examples considered in this paper (the algorithm always converged and the estimates seemed reasonable). Even though the approach of substituting the mean for the mode in the M step worked well for the examples in this paper it should be used very cautiously since there is no theoretical justification that guarantees the algorithm converges in this case. In addition, the properties of the estimates are unknown. For instance, the estimates produced are not necessarily the mean of the observed data posterior. It may be possible to use other more complicated Bayesian computational procedures to calculate the mean of the observed data posterior distribution (Tanner, 1996).

References

Bernardo, J. M., & Smith A. F. M. (1994). *Bayesian Theory*. New York: John Wiley & Sons.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: The MIT Press.

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

Harwell, M. R., Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement*, *15*, 375-389.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika*, *46*, 79-92.

Holland, P. W. (1990). On the sampling foundations of item response theory models. *Psychometrika*, *55*, 577-601.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177-195.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3*. (2nd ed.). Mooresville IN: Scientific Software.

Pommerich, M., & Nicewander, W. A. (1996). *Estimating average domain scores*. Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta, Canada.

Tanner, M. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (3rd ed.). New York: Springer-Verlag.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: John Wiley & Sons.

Tsutakawa, R. K., & Lin H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, *51*, 251-267.

Tsutakawa, R. K. (1992). Prior distribution for item response curves. *British Journal of Mathematical and Statistical Psychology*, *45*, 51-74.

Woodruff, D. J., & Hanson, B. A. (1996). *Estimation of item response models using the EM algorithm for finite mixtures*. ACT Research Report 96-6. Iowa City, IA: American College Testing.

Zeng, L. (1995). *EM1 — An IRT calibration system* [Computer program]. Iowa City, IA: Author.

Zeng, L. (1997). Implementation of marginal Bayesian estimation with four-parameter beta prior distributions. *Applied Psychological Measurement*, *21*, 143-156.

Table 1. Errors in Average Domain Score Estimates

Sample Size = 25

Estimate	Population Mean 0			Population Mean -0.5		
	RMSE	Bias	s.d.	RMSE	Bias	s.d.
Observed	0.0526	0.0341	0.0401	0.0523	0.0356	0.0382
ML	0.0393	0.0026	0.0392	0.0337	0.0023	0.0336
Bayes Normal 25	0.0152	0.0008	0.0152	0.0432	0.0413	0.0127
Bayes Normal 50	0.0094	0.0005	0.0094	0.0513	0.0507	0.0079
Bayes Normal 100	0.0053	0.0003	0.0053	0.0575	0.0573	0.0045
Bayes Normal 500	0.0012	0.0001	0.0012	0.0642	0.0642	0.0010
Bayes Beta 25	0.0194	0.0106	0.0162	0.0511	0.0492	0.0137
Bayes Beta 50	0.0165	0.0129	0.0103	0.0621	0.0615	0.0087
Bayes Beta 100	0.0157	0.0146	0.0059	0.0707	0.0705	0.0051
Bayes Beta 500	0.0165	0.0164	0.0014	0.0802	0.0802	0.0012

Sample Size = 100

| Estimate	Population Mean 0			Population Mean -0.5		
	RMSE	Bias	s.d.	RMSE	Bias	s.d.
Observed	0.0382	0.0328	0.0195	0.0408	0.0362	0.0188
ML	0.0195	0.0005	0.0195	0.0168	0.0011	0.0168
Bayes Normal 25	0.0139	0.0004	0.0139	0.0233	0.0202	0.0116
Bayes Normal 50	0.0108	0.0003	0.0108	0.0317	0.0304	0.0090
Bayes Normal 100	0.0075	0.0002	0.0075	0.0419	0.0414	0.0063
Bayes Normal 500	0.0022	0.0000	0.0022	0.0590	0.0590	0.0018
Bayes Beta 25	0.0148	0.0042	0.0142	0.0263	0.0234	0.0121
Bayes Beta 50	0.0131	0.0068	0.0113	0.0369	0.0357	0.0096
Bayes Beta 100	0.0126	0.0097	0.0080	0.0498	0.0493	0.0068
Bayes Beta 500	0.0149	0.0147	0.0024	0.0728	0.0728	0.0021
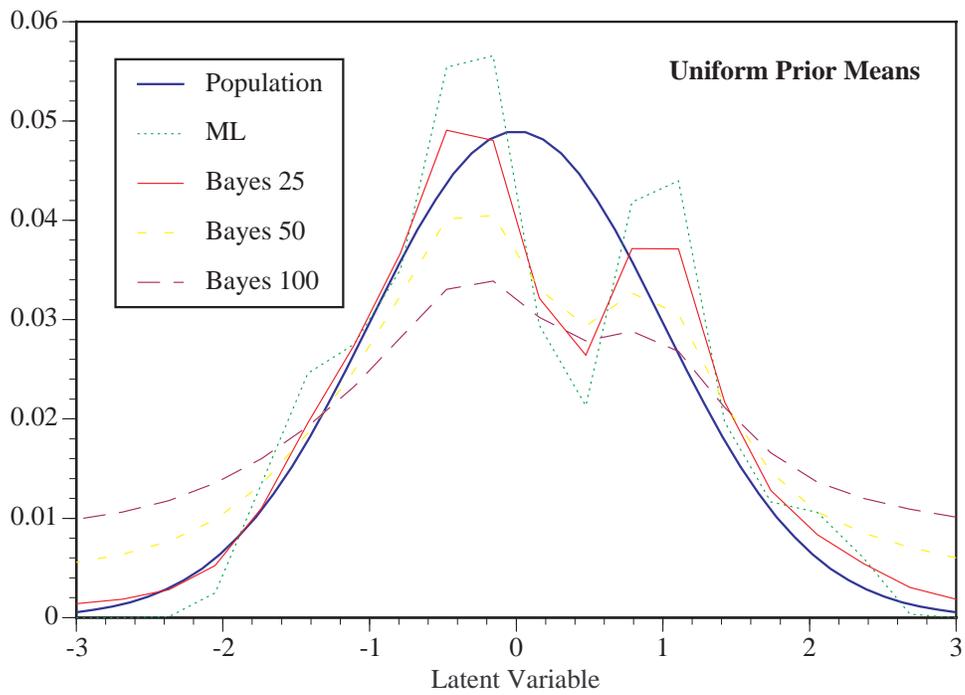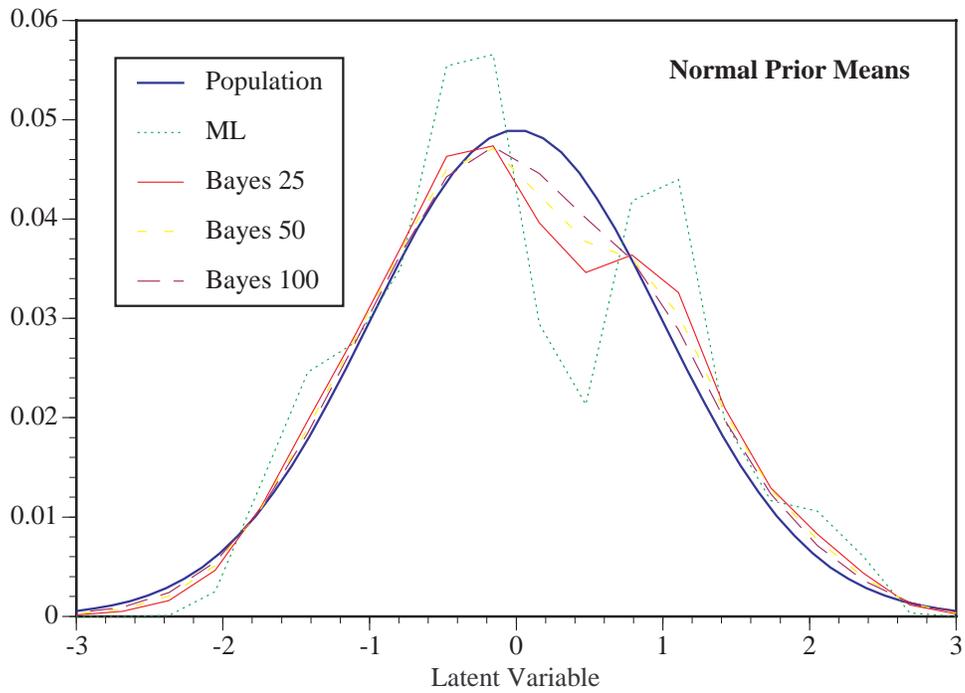
Figure 1. Estimated Distributions using Sample from First Population Distribution (Discrete version of Normal).
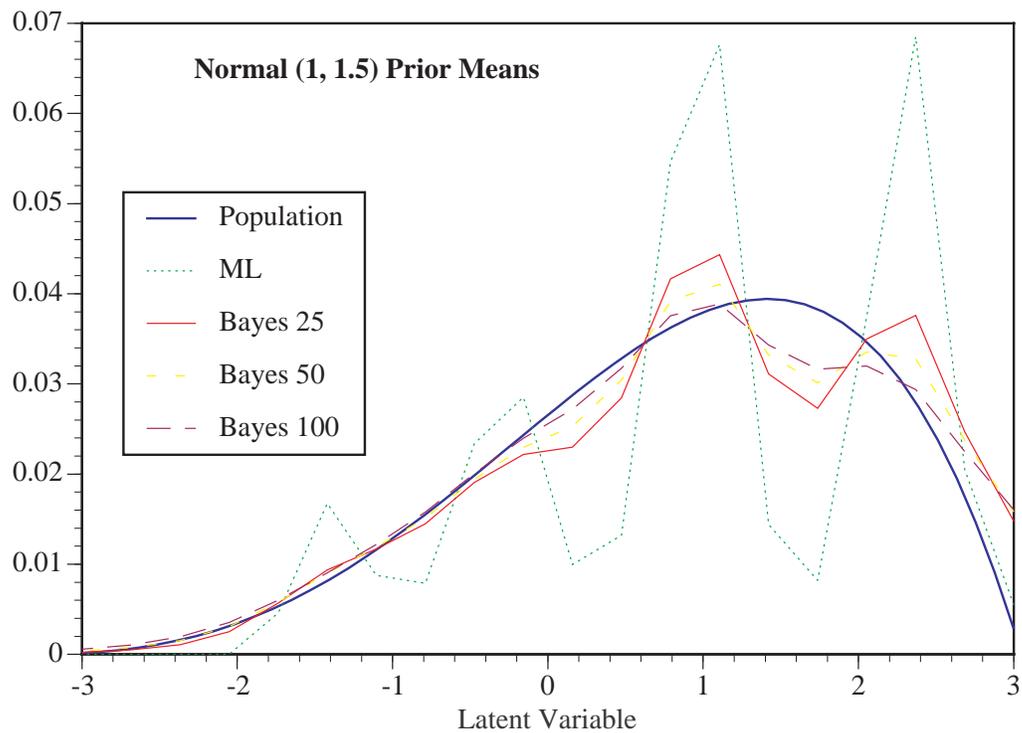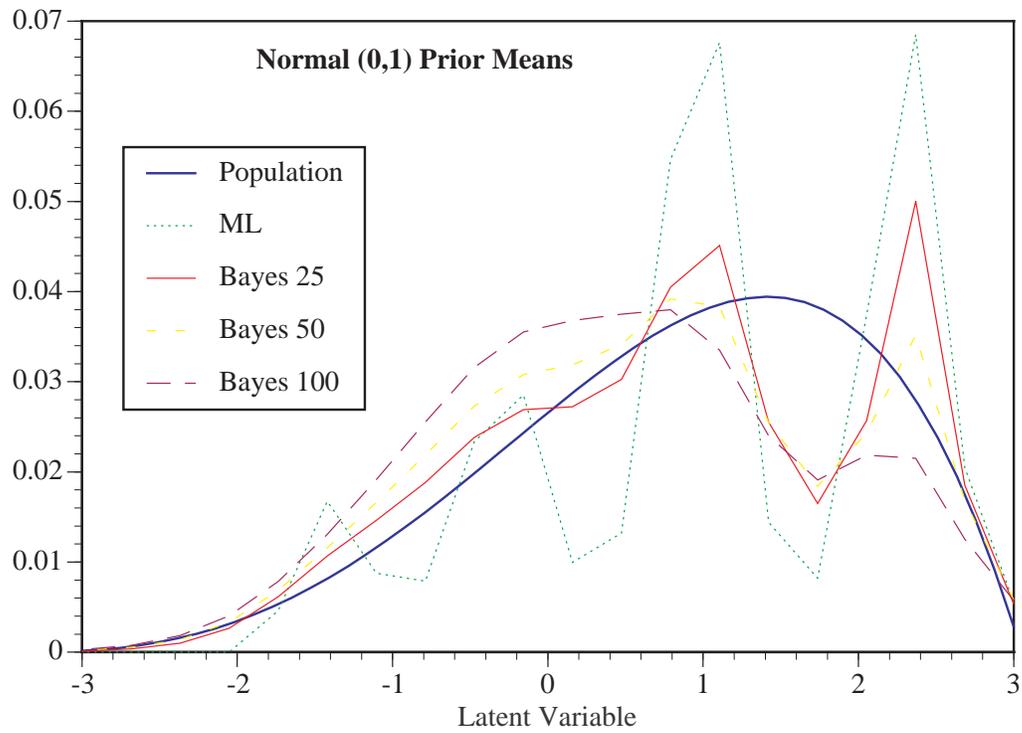
Figure 2. Estimated Distributions using Sample from Second Population Distribution (Discrete Version of Beta).
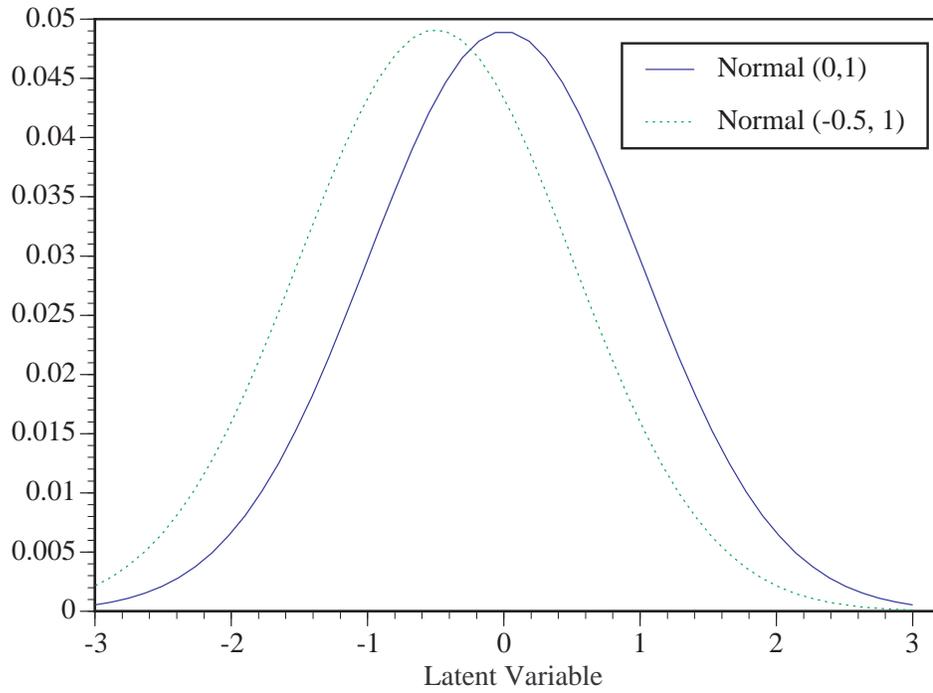
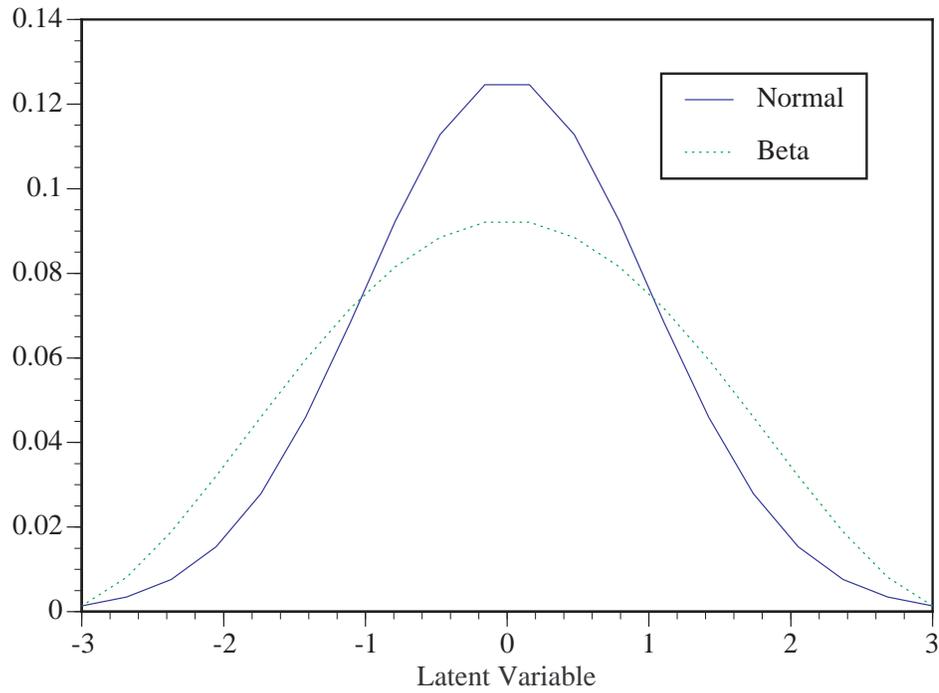Figure 3. Population Distributions used for Average Domain Score Simulations.

Figure 4. Prior Means used for Bayes Modal Estimates in Average Domain Score Simulations.