PSYCHOMETRIKA—VOL. 65, NO. 1, 93–119
MARCH 2000

BAYESIAN INFERENCE FOR FINITE MIXTURES OF GENERALIZED LINEAR MODELS WITH RANDOM EFFECTS

PETER J. LENK

THE UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

WAYNE S. DESARBO

PENNSYLVANIA STATE UNIVERSITY

We present an hierarchical Bayes approach to modeling parameter heterogeneity in generalized linear models. The model assumes that there are relevant subpopulations and that within each subpopulation the individual-level regression coefficients have a multivariate normal distribution. However, class membership is not known a priori, so the heterogeneity in the regression coefficients becomes a finite mixture of normal distributions. This approach combines the flexibility of semiparametric, latent class models that assume common parameters for each sub-population and the parsimony of random effects models that assume normal distributions for the regression parameters. The number of subpopulations is selected to maximize the posterior probability of the model being true. Simulations are presented which document the performance of the methodology for synthetic data with known heterogeneity and number of sub-populations. An application is presented concerning preferences for various aspects of personal computers.

Key words: Bayesian inference, consumer behavior, finite mixtures, generalized linear models, heterogeneity, latent class analysis, Markov chain Monte Carlo.

1. Introduction

Finite mixture or latent class models have been discussed in the statistical literature as early as the classic works of Newcomb (1886) and Pearson (1894). These semiparametric models assume that the sample of observations arises from a specified number of underlying subpopulations where the relative proportions of the subpopulations are unknown. The forms of the densities in each of these subpopulations are specified. However, subpopulation or class membership is not known a priori, so the density for a randomly selected observation is the convex sum of the component densities for the subpopulations. The primary inferential goals are to decompose the sample into its mixture components and to estimate the mixture probabilities and the unknown parameters of each component density. Everitt and Hand (1981) and Titterington, Smith, and Makov (1985) review the various types of distributions involved in such mixtures and discuss identification issues, as well as method of moments and maximum likelihood estimators.

DeSarbo and Cron (1988) propose a conditional mixture model that postulates separate regression functions within each of K subpopulations. Their procedure simultaneously partitions the population into K subpopulations and estimates the separate regression parameters per subpopulation. This model generalizes the Quandt (1972), Hosmer (1974), and Quandt and Ramsey (1978) stochastic switching regression models to more than two classes. DeSarbo and Cron use an EM algorithm (Dempster, Laird, & Rubin, 1977) to obtain maximum likelihood estimates of the K regression functions and posterior probabilities of an subject's memberships to the subpopulations. A large number of mixture regression models have since been developed (see Wedel & DeSarbo, 1994, for a review). Lwin and Martin (1989), De Soete and DeSarbo (1991) and Wedel and DeSarbo (1993) developed conditional mixture binomial probit and logit regression models. Kamakura and Russell (1989) and Kamakura (1991), respectively, develop conditional mixture

Requests for reprints should be sent to Peter J. Lenk, University of Michigan Business School, 701 Tappan Street, Ann Arbor MI 48109-1234.

0033-3123/2000-1/1996-0498-A $00.75/0

multinomial logit and probit regression models. Wang, Cockburn, and Puterman (1998); Wang, Puterman, Le, and Cockburn (1996); and Wedel, DeSarbo, Bult, and Ramaswamy (1993) proposed conditional univariate Poisson mixture regression models, and Wang and Puterman (in press) present a mixture of logistic regression models. DeSarbo, Ramaswamy, Reibstein, and Robinson (1993), DeSarbo, Wedel, Vriens, and Ramaswamy (1992), and Jones and McLachlan (1992) developed conditional multivariate normal regression mixtures.

An important aspect that has not been adequately addressed in these various finite mixture approaches concerns heterogeneity within each latent class or subpopulation. Traditional finite mixture specifications have been employed to implicitly model sample heterogeneity where each component density or latent class is often interpreted in many applications as separate subpopulations or response modes (e.g., segments of consumers). Although mixture models have seen a wide number of applications, accumulated empirical evidence suggests the need to reflect the diversity of characteristics, preferences, sensitivities, etc. within each component class (Allenby, Arora, & Ginter, 1998). That is, common coefficients for each subpopulation often do not accurately summarize the within-class variation.

In the presence of substantial, within-class heterogeneity in the coefficients, the finite mixture solution often requires an excessive number of latent classes or subpopulations to represent the heterogeneity adequately in the data, leading to over parameterization and many, relatively small, latent classes. An alternative formulation is a random effects model that assumes the subject-level coefficients are a random sample from a normal distribution. These models accommodate more extensive heterogeneity with fewer parameters than latent class models, provided that the normal assumption holds. (See Lenk, DeSarbo, Green, & Young, 1996, for a comparison and further references.) However, they may be inadequate in the presence of sizable subpopulations.

As a remedy, this paper proposes to extend the assumption of the traditional random effects model by using a finite mixture of normal distributions for the distribution of the coefficients. This model provides both the flexibility of the latent class model and the parsimony of the traditional random effects model. Indeed, both models are special cases of the proposed model: the latent class model corresponds to letting the within-class variances go to zero, and the traditional random effects model corresponds to using only one class or component.

The paper assumes that the dependent observations are from a generalized linear model (GLM; McCullagh & Nelder, 1983), which includes commonly used distributions such as binomial, Poisson, normal, and gamma. Wedel and DeSarbo (1995) recently proposed latent class, generalized linear models. Special cases of this framework are binomial probit and logit regression mixtures (DeSoete & DeSarbo, 1991; Wedel & DeSarbo, 1993), univariate Poisson regression mixtures (Wedel et al., 1993), and latent class analysis (Goodman, 1974). Zeger and Karim (1991) propose a Bayesian analysis of GLM which have random effects, and Breslow and Clayton (1993) propose an approximate Bayes procedure.

The paper assumes that the individual-specific regression parameters for the linear predictor in GLM are distributed across the population according to a finite mixture of multivariate normal distributions. Also, each member of the population has a different scale parameter. The heterogeneity in the individual-level scale parameters is described by a normal distribution (cf. Lenk, DeSarbo, Green, and Young 1996).

The paper uses Markov Chain Monte Carlo (MCMC; Gelfand & Smith, 1990; and Smith & Roberts, 1993) to approximate the Bayesian inference. Diebolt and Robert (1994) propose a MCMC for mixture models of univariate observations, and its model is not identified because permutations of the class labels, called "label switching", result in the same value of the likelihood function (Titterington, Smith, & Makov, 1985). Label switching results in misleading estimators when using MCMC procedures. Suppose that there are K subpopulations. A well designed Markov chain should explore the full parameter space, which includes $K!$ regions defined by permuting the mixture components' labels. When iterations are averaged over these visits, the MCMC estimator for a component's parameter converges to a weighted average of that pa-

rameter in all components where the weights are proportional to the number of iterations in each region. In contrast, the EM algorithm always moves in a direction that maximizes the likelihood function and does not face this problem. For a given starting point, EM terminates at one of the $K!$ modes and reports only the final results, ignoring previous iterations that may have had label switches. This paper identifies the model by ordering the mixture probabilities and modifies the standard MCMC algorithm to include these order restrictions.

An outstanding problem for mixture models is the choice of the number of mixture components. Currently, choice heuristics are based on information measures such as AIC (Akaike, 1973), consistent AIC or CAIC (Bozdogan, 1987), and BIC (Schwarz, 1978). These measures penalize the likelihood of the models where the penalty term is a function of the number of parameters, thus balancing fit with parsimony. This paper proposes computing the posterior probabilities for the number of mixture components. Jeffreys (1961, chap. V, VI) is frequently cited as the first instance of Bayesian hypothesis testing by using posterior probabilities of the hypotheses. BIC is a large sample approximation of the marginal distribution, which integrates the likelihood by the prior distribution of the parameters. Kass and Raftery (1995) review the vast literature on Bayes factors for model selection, and recent work by Carlin and Chib (1995), Chib (1995), Lewis and Raftery (1997), and Verdinelli and Wasserman (1995) considered their computation via Markov chain methods. We adapt the method of Gelfand and Dey (1994) to select the number of mixture components.

The next section illustrates the inadequacy of the latent class model in the presence of substantial, within-class heterogeneity in the coefficients, and introduces the mixture, random effects model in the special case of normal, linear regression. Section 3 presents the generalized linear model and discusses respective identification issues. Section 4 approximates the marginal distribution of the number of components. Section 5 summarizes two simulation studies using normal and Bernoulli 0/1 data. The simulations demonstrate that the Bayesian analysis recovers the true model and that the posterior probabilities indicate the correct number of mixture components. Section 5 also applies the proposed methodology to actual data collected on subjects' preferences for personal computer. Finally, we mention several areas for future research, as well as additional potential applications.

2. Latent Class Mixture Models and Misspecification

This section motivates the mixture, random effects model by highlighting some of the shortcomings of the latent class mixture model when there is substantial, within-class heterogeneity. The data consist of observations on n subjects or experimental units. There are multiple observations on each subject: subject i has m_i observations. Let Y_{ij} be the j-th dependent observation on subject i and $\mathbf{x}_{ij}$ be the corresponding $p \times 1$ vector of independent variables, which usually includes "1" for the intercept. Define

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{bmatrix} \text{ and } X_i = \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{im_i} \end{bmatrix} \text{ for } i = 1, \ldots, n$$

where $\mathbf{Y}_i$ is a $m_i \times 1$ vector of dependent observations, and X_i is a $m_i \times p$ design matrix for subject i.

The latent class model assumes that the population consists of K subpopulations and that there are separate regression models within each subpopulation. Suppose that the i-th subject or experimental unit belongs to class k. The latent class regression model specifies:

$$\mathbf{Y}_i = X_i\theta_k + \epsilon_{ik} \text{ for } i = 1, \ldots, n$$

where θ_k is a $p \times 1$ vector of regression parameters for latent class k, and ϵ_{ik} is a $m_i \times 1$ vector that has a multivariate normal distribution with mean 0 and covariance matrix $\sigma_k^2 I$ where I is the

identity matrix. The error terms are mutually independent across subjects. The proportion of the population who belongs to class k is ψ_k. If class membership is not known, the unconditional (on class membership) density of $\mathbf{Y}_i$ is a finite mixture model with K component densities:

$$f_i(\mathbf{y}_i) = \sum_{k=1}^{K} \psi_k q_{m_i}(\mathbf{y}_i | X_i\theta_k, \sigma_k^2 I), \tag{1}$$

where $q_{m_i}(\cdot|X_i\theta_k, \sigma_k I)$ is the m_i dimensional, multivariate normal density with mean $X_i\theta_k$ and covariance matrix $\sigma_k^2 I$.

Instead of assuming a common coefficient θ_k for all subjects in class k, the mixture, random effects model assumes that the regression coefficients are subject specific; these coefficients belong to one of K classes; and within a class the coefficients vary according to a normal distribution. For subject i:

$$\mathbf{Y}_i = X_i\beta_i + \epsilon_i \quad \text{for } i = 1, \ldots, n \tag{2}$$

where $\mathbf{Y}_i$ is a $m_i \times 1$ vector; X_i is a $m_i \times p$ design matrix; β_i is a $p \times 1$ vector of unknown regression coefficients, and ϵ_i is a $m_i \times 1$ vector of error terms that has a multivariate normal distribution with mean 0 and covariance matrix $\sigma_i^2 I$. The error terms are mutually independent. Further, we assume that the log error variances $\{\phi_i = \log(\sigma_i^2)\}$ are a random sample from a normal distribution with mean α and variance τ^2.

If subject i belongs to class k, then β_i has a normal distribution with mean θ_k and covariance matrix Λ_k, and the proportion of the population who belong to class k is ψ_k. If class membership is not known, then the regression coefficients are a random sample from a mixture distribution with the following density:

$$g(\beta_i) = \sum_{k=1}^{K} \psi_k q_p(\beta_i | \theta_k, \Lambda_k), \tag{3}$$

where $q_p(\cdot|\theta_k, \Lambda_k)$ is the p dimensional, multivariate normal density with mean θ_k and covariance matrix Λ_k. The unconditional mean and covariance of β_i are:

$$E(\beta_i) = \theta = \sum_{k=1}^{K} \psi_k \theta_k \tag{4}$$

$$Var(\beta_i) = \Lambda = \sum_{k=1}^{K} \psi_k(\Lambda_k + \theta_k\theta_k') - \theta\theta'. \tag{5}$$

After integrating over β_i, the marginal distribution of $\mathbf{Y}_i$ is also a mixture model:

$$f_i(\mathbf{y}_i) = \sum_{k=1}^{K} \psi_k q_{m_i}(\mathbf{y}_i | X_i\theta_k, \sigma_i^2 I + X_i\Lambda_k X_i'). \tag{6}$$

The covariance matrix for component k in (6) allows for a nonzero covariance structure, while the latent class model in (1) assumes that observations for subject i are mutually independent. The means for both models are the same.

To demonstrate the inadequacy of the latent class model in the presence of substantial within-class heterogeneity, we simulated data according to (2) and (3). There were 100 "subjects" and 10 observations per subject. First, we independently generated an independent variable $\mathbf{x}_{ij}$ from a normal distribution with mean two and standard deviation one. Each subject is then independently assigned to one of three classes with probabilities $\psi_1 = .2$, $\psi_2 = .3$, and $\psi_3 = .5$. Conditional on class assignment, the intercept β_{1i} and slope β_{2i} were generated from

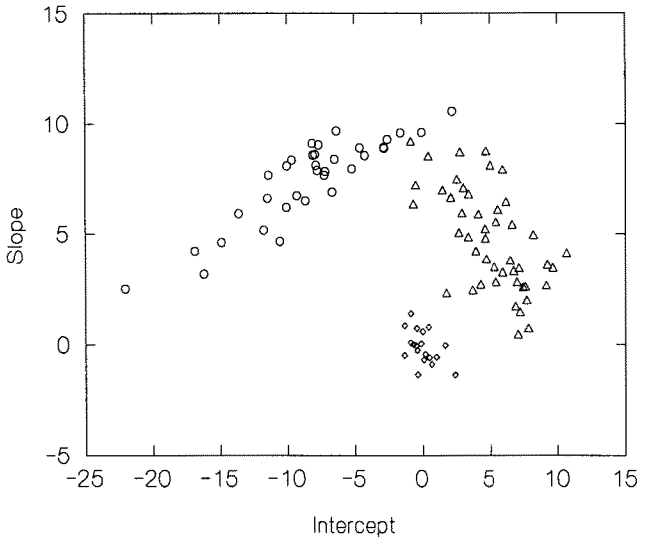

FIGURE 1.
True Regression Coefficients for a Simulated Data Set

a bivariate normal distribution. Figure 1 plots the true, individual-level slopes versus the true intercepts. Next, for each subject an error variance was generated from a lognormal distribution. Finally, the dependent variables were constructed with normally distributed errors.

The top half of Table 1 reports the simulation parameters, and the bottom half reports the results for the finite mixture regression model (1) that uses three latent classes. This model is estimated with the EM algorithm (DeSarbo & Cron, 1988), which is an iterative, maximum likelihood method. Subjects are assigned to classes based on their posterior probability of membership given the current parameters estimates from the EM algorithm. Next, the parameters are re-estimated based on the current assignment. The procedure is repeated until the likelihood function no longer increases. This solution severely distorts the sizes of the derived classes and biases the estimated class coefficients θ_k. In addition, ignoring the within-class heterogeneity in the regression coefficients inflates the error variances.

TABLE 1.
Parameters for the Simulated data and the Latent Class Regression Estimates

Component	One	Two	Three
Size	14	33	53
Intercepts' Mean	0	−10	5
Slopes' Mean	0	7	5
Intercepts' Variance	1	25	9
Slopes' Variance	1	4	5
Intercept-Slope Covariance	0	9	−5
Error Variances' Mean	9.49	9.49	9.49
Error Variances' Standard Deviation	7.64	7.64	7.64
		Latent Class Regression	
Size	20	36	44
Intercept	3.62	−7.50	4.10
Slope	7.80	3.30	4.60
Error Variance	19.72	43.01	21.58

One response to the inadequate representation of the heterogeneity provided by the three class solution is to increase the number of support points. We fitted models with one to 12 latent classes. The 11 support-point solution minimized three common information criterion: AIC (Akaike, 1973), consistent AIC or CAIC (Bozdogan, 1987), and BIC (Schwarz, 1978). These information criterion unambiguously indicate a 11 class solution, which has 43 parameters: 11 intercepts, 11 slopes, 11 error variances, and 10 mixture probabilities. Many of the classes are very small and some only have three members, resulting in large standard errors. In addition, if these models were used to identify subpopulations, 11 classes would be difficult to interpret, especially since the data have only one independent variable.

Another approach would be to use a random effects model with K equal to one in (3). Using (4) and (5) for the example, we would fit a model where β_i has a normal distribution with mean and covariance matrix:

$$\theta = \begin{bmatrix} -0.5 \\ 4.6 \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} 54.45 & -6.00 \\ -6.00 & 9.94 \end{bmatrix}.$$

We can see in Figure 1 that the mean θ is in a region without subjects, and the 95% ellipsoid would contain unusually many subjects along its north-west to north-east boundary and unusually few subjects in the south-west quadrant.

The mixture, random effects model is more parsimonious than the latent class model and more flexible than the random effect model with one component. Both are special cases of the mixture, random effects model. By setting $\Lambda_k = 0$ in the marginal density in (6), one obtains the marginal density for the latent class model in (1). Also, setting K equal to one results in the traditional random effects model.

3. Finite Mixtures of Generalized Linear Models

The j-th dependent variable for the i-th subject, Y_{ij}, is from the generalized linear model (McCullagh & Nelder, 1983) with density:

$$f(y_{ij}|\beta_i) = \exp\left[\frac{y_{ij}h(\mathbf{x}_{ij}'\beta_i) - b[h(\mathbf{x}_{ij}'\beta_i)]}{a(\phi_i)} + c(y_{ij}, \phi_i)\right] \tag{7}$$

$$\text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m_i$$

where $\mathbf{x}_{ij}$ is a $p \times 1$ vector of independent variables; β_i is a $p \times 1$ vector of regression coefficients; $h(\mathbf{x}_{ij}'\beta_i) = \xi_{ij}$ is the natural parameter, and ϕ_i is the scale parameter that may depend on the subject. The functions a, b, and h are univariate and real-valued. For the normal distribution, $h(\mathbf{x}_{ij}'\beta_i)$ is the mean; $a(\phi_i) = \exp(\phi_i)$ is the variance; $b[h(\mathbf{x}_{ij}'\beta_i)] = \frac{1}{2}h(\mathbf{x}_{ij}'\beta_i)^2$; and $c(y_{ij}, \phi_i) = -\frac{1}{2}(y_{ij}^2 \exp(-\phi_i) + \ln(2\pi) + \phi_i)$. We will assume that the observations are mutually independent. The mean and variance of Y_{ij} are $b_1(\xi_{ij})$ and $b_2(\xi_{ij})a(\phi_i)$ respectively, where $b_1 = \frac{d}{d\xi}b(\xi)$ is the first derivative of b, and $b_2 = \frac{d^2}{d\xi^2}b(\xi)$ is the second derivative of b. We will assume that the regression coefficients follow the mixture model in Equation (3). The subject specific scale parameters $\{\phi_i\}$ form a random sample from a normal distribution with mean α and variance τ^2.

As mentioned in the Introduction, the above model is not identified because permutations of the class labels result in the same value of the likelihood function. Titterington, Smith, and Makov (1985) remark that there does not exist one set of parameter restriction that will identify the model for all possible choices of parameters in the multivariate setting. This paper identifies the model by ordering the mixture probabilities: $\psi_1 < \cdots < \psi_K$. If none of the true mixture probabilities are equal, then this ordering identifies the model. If two or more components have the same mixture probabilities, then this restriction is inappropriate. However, simulation studies using equal probabilities indicate that the algorithm is not adversely affected because parameter

uncertainty masks the equality of the mixture probabilities. If one believed that some probabilities are equal, then Titterington, Smith, and Makov recommend additional restrictions such as ordering the intercepts or variances. Clearly, there may be situations were combinations of these restrictions will not, in theory, identify the model.

The prior distributions for the remaining parameters are mutually independent and have the following specification: ψ has a Dirichlet distribution constrained to the region $\psi_1 < \cdots < \psi_K$, and the distributions for θ_k and Λ_k for $k = 1, \ldots, K$ are multivariate normals and Inverted Wishart, respectively. α has a normal distribution, and τ^2 has an inverse gamma distribution. K has a discrete probability function on the integers $1, \ldots, M$ where M is specified by the researcher. These prior distributions were selected for three reasons: they facilitate the posterior analysis; they are fairly flexible families; and their prior parameters can be selected so that the posterior analysis is relatively insensitive to the prior for data sets with a moderate number of subjects and observations per subject. Lenk, DeSarbo, Green, and Young (1996) illustrate this point with a hierarchical Bayes linear regression model by randomly deleting observations within subjects.

Appendix A provides details about the joint distribution of the data and the unknown parameters. Appendix B presents the prior parameters used in the empirical examples, and Appendix C describes the MCMC algorithm, which is an iterative method of generating random deviates from the posterior distribution of the parameters. The basic idea is to generate random deviates from a Markov chain such that its stationary distribution is the posterior distribution.

4. Model Selection

The number of mixture components can be selected by choosing the model with the largest posterior probability. If the number of components are a priori equally likely, then choosing the model with the largest Bayes factor is an equivalent procedure. Both procedures require computing the marginal density of the data given the number of mixture components. The marginal density integrates the likelihood function times the prior density over the parameter space.

We use the method of Gelfand and Dey (1994) to approximate the marginal density from the output of the Markov chain. For the model with K components, indicate all of the parameters by Ω_K. The marginal density of the data given K components is

$$\begin{aligned} f_K(Y) &= \int_{\Omega_K} f_K(Y|\Omega_K) p_K(\Omega_K) d\Omega_K \\ &= \left\{ E\left[\frac{g_K(\Omega_K)}{f_K(Y|\Omega_K) p_K(\Omega_K)} \right] \right\}^{-1}, \end{aligned}$$

where f_K is the density of the data given the parameters for model K; p_K is the prior density of the parameters; g_K is an arbitrary density on the support of Ω_K, and the expectation is with respect to the posterior distribution of Ω_K. The MCMC approximation is

$$\tilde{f}_K(Y) = \left[\frac{1}{U-B} \sum_{u=B+1}^{U} \frac{g_K\left(\Omega_K^{(u)}\right)}{f_K\left(Y|\Omega_K^{(u)}\right) p_K\left(\Omega_K^{(u)}\right)} \right]^{-1},$$

where $\Omega_K^{(u)}$ is the value of Ω_K on the iteration u of the Markov chain, and the last $U - B$ iterations of U iterations are used. If g_K is the posterior density of Ω_K, then the approximation is exact. However, we only know the posterior density to a normalizing constant, and the unknown normalizing constant is exactly the quantity that we need to compute. Consequently, one needs to specify a g_K that is completely known. The estimated, posterior probabilities are $\tilde{P}(K|Y) \propto p(K)\tilde{f}_K(Y)$ where $p(K)$ is the prior probability for K mixture components. Inde-

pendent Markov chains are run for each of the mixture models with 1 to M components. In the empirical work of this paper, the prior probabilities are equally likely.

Kass and Raftery (1995) recommend that g_k should be close to the posterior density. We specify g_K either by using the property that posterior distributions are asymptotically normal or else by using the fact that distributions are conjugate given the other parameters. For example, if the class membership and $\{\beta_i\}$ were know, then Λ_K would have an Inverted Wishart distribution. The parameters of g_K are estimated from the output of the Markov chain by the method of moments. For example, $g_K(\beta_i)$ is assumed to be the normal density. On iteration u of MCMC, the draw $\beta_i^{(u)}$ is saved. Then the mean and covariance of these random deviates are used to estimate the mean and covariance of $g_K(\beta_i)$. Appendix D provides further details about the choice of g_K.

5. Empirical Studies

Section 5.1 reports two simulation experiments using normal and Bernoulli data. Section 5.2 applies the methodology to analyze empirical preference data for personal computers.

5.1. Simulations

The purposes of the simulations are two-fold. First, they verify that for a known number of mixture components the MCMC procedure recovers the unknown parameters. Second, they demonstrate that the posterior probabilities of the models indicate the correct model. The first simulation generates data from a linear regression model with normal error, and the second generate 0/1 data from a logistic regression model.

Fifty data sets are generate for the simulation study of the linear regression model. As in section 2, (2) has a slope and intercept, and the true model has three mixture components. The parameters for the simulated data are the same as section 2 except that α is -1 and τ^2 is 4. Each data set consists of 100 subjects and 10 observations per subject. The procedure was initialized by randomly assigning subjects to groups. The MCMC ran for 2000 iterates and utilized the last 1000 for estimation.

Figure 2 graphs the MCMC iterations for the means and standard deviations of the regression coefficients, the scale parameters, and the mixture probabilities from one of the simulated data sets. The initial, transitory period in the graphs is due to the algorithm searching for the best classification of subjects, after which the procedure quickly settles into the stationary distribution.

Despite much recent work, convergence diagnostics for MCMC remains an open question, which is beyond the scope of this paper. See, for example, Gelman and Rubin (1992), Geyer (1992), Polson (1996), and Roberts and Polson (1994). As a practical matter, researchers frequently plot the MCMC iterations to verify convergence. For example, the plots in Figure 2 seem to indicate that the chain has converged to its stationary distribution by iteration 1000. The convergence issue with MCMC is similar to that for maximum likelihood estimation: the chain can become stuck in a region of the parameter space corresponding to a local mode of the likelihood function. Then convergence diagnostics based on the chain may falsely signal convergence. Additional safeguards are to run multiple chains from different starting points to verify that they result in similar answers and to perform simulation studies where the true parameters are known. We used these last two methods, along with visual inspection of the random deviates plotted against iteration, to decide that the chain has run sufficiently long.

Table 2 reports the results for 50 simulated data sets for the three component solution. The Bayesian analysis is compared to an ad-hoc, three-stage procedure, which is sometimes used by practitioners. First, individual-level maximum likelihood (ML) estimates of the coefficients are obtained. Second, subjects are clustered based on their ML estimates. The clustering is performed by nearest neighbor agglomerative clustering (Seber 1984, pp. 360 to 361). Third, the means and

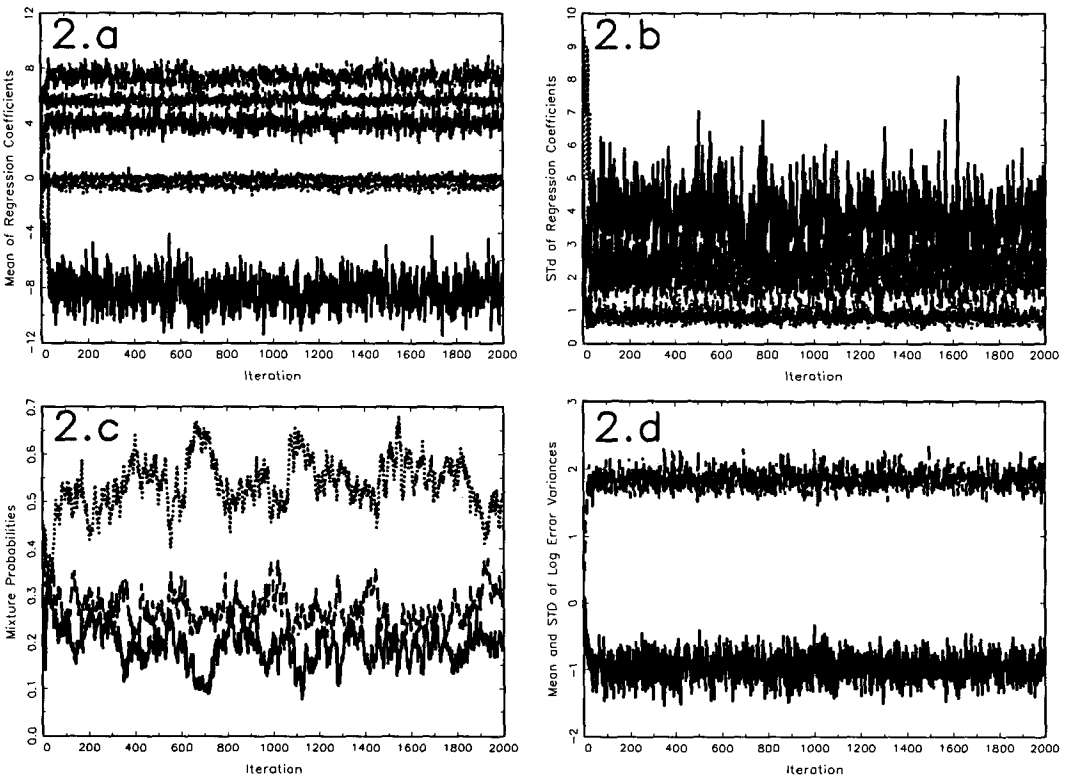

FIGURE 2.
Iterations from the MCMC Sampler for the Three Class Solution. (2.a) mean of the regression coefficents, (2.b) standar deviation of the regression coefficents, (2.c) mixture probabilities, and (2.d) mean and variance of the log error variances.

covariance matrices for a cluster are estimated from the estimated regression coefficients for subjects assigned to that cluster.

Table 2 first shows that the estimated mixture probabilities from MCMC are close to their true values and are more accurate than those obtained from the three stage, clustering algorithm. The expected number of subjects in the classes are 20, 30, and 50. On average, 19 of the 20 subjects expected in class one are correctly classified; approximately 0.6 of the subjects in class one are assigned to class two and 0.02 to class three. The numbers do not add to 20 because there are not 20 subjects in class one on every simulation. Likewise, 27 of the 30 subjects in class two are correctly classified, as are 48 of the 50 subjects in class three. In contrast, the clustering algorithm has much higher misclassification counts.

Table 2 then reports the estimated parameters of each mixture component. The estimators depend on the classification of the observations. The MCMC estimators are the posterior means. The posterior standard deviations provide a measure of the uncertainty about the parameters: they are used in the same way that standard errors are used. The posterior means tend to be within one or two posterior standard deviations across the simulations. Also, averaging over simulations, they are within one or two simulation standard deviations. The three-stage cluster estimates are less accurate, which reflects their larger misclassification rates.

The reason for the inferior performance of the ad hoc three-stage procedure is that it treats the estimated regression coefficients as if they were the true parameters and ignores their sampling variation. Invariably, with 100 subjects some of the estimated regression coefficients are very inaccurate, which results in a poor cluster solution. Other simulations indicate that when the number of observations per subject is large relative to the number of regressors, the ad hoc three-stage procedure performs nearly as well as the Bayesian model.

TABLE 2.
Classification Rates and Parameter Estimates from the Simulation Study of the Mixture, Normal Regression Model with Three Components

Mixture Probabilities						
	Hierarchical Bayes			Cluster the MLEs		
Component	One	Two	Three	One	Two	Three
True	0.200	0.300	0.500	0.200	0.300	0.500
Estimate	0.198	0.310	0.492	0.164	0.284	0.552
	(0.004)	(0.004)	(0.006)	(0.011)	(0.011)	(0.017)

Classification Rates						
Number	Hierarchical Bayes			Cluster the MLEs		
Assigned to	One	Two	Three	One	Two	Three
One	19.327	0.531	0.146	6.240	7.900	2.260
	(0.632)	(0.413)	(0.038)	(1.370)	(1.153)	(0.700)
Two	0.558	26.988	2.077	4.540	14.560	9.320
	(0.438)	(0.931)	(0.789)	(1.210)	(1.587)	(1.961)
Three	0.155	2.301	47.916	9.260	7.360	38.560
	(0.038)	(0.681)	(1.076)	(1.409)	(1.299)	(2.212)

Parameter Estimates								
Component	True	Posterior Mean	Posterior STD	Cluster the MLEs	True	Posterior Mean	Posterior STD	Cluster the MLEs
		Intercepts' Mean				Intercepts' Variance		
One	0	−0.209	0.288	−5.987	1	1.238	0.599	28.222
		(0.220)	(0.016)	(1.067)		(0.189)	(0.106)	(11.033)
Two	−10	−9.400	0.995	−4.072	25	23.413	7.311	16.164
		(0.378)	(0.030)	(0.918)		(1.129)	(0.349)	(2.375)
Three	5	4.599	0.503	2.291	9	9.640	2.503	22.510
		(0.281)	(0.012)	(0.457)		(0.224)	(0.093)	(3.109)
		Slopes' Mean				Slopes' Variance		
One	0	0.146	0.243	4.177	1	1.016	0.438	2.388
		(0.133)	(0.008)	(0.467)		(0.054)	(0.032)	(0.284)
Two	7	6.806	0.389	5.439	4	3.700	1.127	4.562
		(0.156)	(0.009)	(0.341)		(0.149)	(0.046)	(0.704)
Three	5	5.108	0.359	4.548	5	5.280	1.252	7.600
		(0.069)	(0.006)	(0.166)		(0.189)	(0.039)	(0.572)
		Intercepts and Slopes' Covariance				Mean Log Error Variance		
One	0	0.085	0.360	−2.308	−1	−1.007	0.204	−1.352
		(0.093)	(0.044)	(1.452)		(0.031)	(0.002)	(0.032)
Two	9	8.047	2.632	−0.430		Variance Log Error Variance		
		(0.466)	(0.110)	(0.885)	4	3.973	0.604	4.243
Three	−5	−5.114	1.557	−4.032		(0.077)	(0.011)	(0.075)
		(0.290)	(0.047)	(0.745)				

The means across 50 simulations are reported. The simulation standard errors for the means are reported in parenthesis.

TABLE 3.
Performance of Individual-level Estimates from the Simulation Study of the Mixture, Normal Regression Model with Three Components

	Intercept		Slope		Log Error Variance	
	RMSE	CORR	RMSE	CORR	RMSE	CORR
Posterior Mean	0.908	0.992	0.428	0.990	0.521	0.967
	(0.035)	(0.001)	(0.017)	(0.001)	(0.005)	(0.001)
Maximum Likelihood	1.184	0.986	0.541	0.984	0.632	0.967
	(0.051)	(0.001)	(0.024)	(0.002)	(0.007)	(0.001)

"RMSE" is the root mean squared error between the true values and their estimators, and "CORR" is the correlation. The performance measures are averaged across the 50 simulated data sets. Simulation standard errors are in parentheses.

Table 3 reports the performance of the individual-level parameter estimates. The Bayes estimates have smaller RMSE and larger correlations with the true individual-level parameters than do the individual-level ML estimates. Table 4 summarizes the model selection criterion for the simulation study. The models are estimated with one to five components. Each of the five models are assumed to be equally likely. The posterior probabilities correctly identify 48 of the 50 simulations as having three clusters.

The second simulation study generated 50 data sets from a logistic regression model. The dependent variables $\{Y_{ij}\}$ take the values zero or one with the following probabilities:

$$P(Y_{ij} = 1) = \frac{\exp(\mathbf{x}'_{ij}\beta_i)}{1 + \exp(\mathbf{x}'_{ij}\beta_i)} \text{ and } P(Y_{ij} = 0) = 1 - P(Y_{ij} = 1).$$

Logistic regression is in the exponential family with

$$h(\mathbf{x}'_{ij}\beta_i) = \mathbf{x}'_{ij}\beta_i; \quad a(\phi_i) = 1; \quad b[h(\mathbf{x}'_{ij}\beta_i)] = \ln[1 + \exp(\mathbf{x}'_{ij}\beta_i)]; \text{ and } c(y_{ij}, \phi_i) = 0.$$

Each of the 50 data sets has 100 "subjects" and 40 observations per subject. The predictors are 2×1 vectors whose first element is a "1", and the second element is drawn from a standard normal distribution. The true model has two mixture components, which represent 40% and 60% of the data.

The true parameters along with the simulation results are given in Table 5. The simulation indicates that the procedure correctly classifies subjects and accurately estimates the parameters of the model. The model selection criterion are given in Table 6. The posterior probabilities correctly identified all 50 data sets.

TABLE 4.
Model Choice for the Simulation Study of the Mixture, Normal Regression Model

	Number of Mixture Components				
	One	Two	Three	Four	Five
Log P(Data)	−1730.4	−1707.5	−1673.7	−1687.8	−1694.3
	(87.0)	(88.3)	(88.3)	(86.5)	(87.5)
Posterior Probability	0.000	0.000	0.958	0.022	0.020
	(0.000)	(0.000)	(0.185)	(0.124)	(0.141)
Choice Counts	0	0	48	1	1

"Choice Counts" is the number of times in 50 simulated data sets that the posterior probability was maximum for the corresponding mixture model. The measures are averaged across the 50 simulated data sets. Simulation standard errors are in parentheses.

TABLE 5.
Classification Rates and Parameter Estimates from the Simulation Study of the Logistic Regression Model

Classification Rates					
Probability			Classification Counts		
Component	One	Two	Component	One	Two
True	0.400	0.600	One	39.826	1.352
Estimate	0.403	0.597		(1.057)	(0.995)
	(0.006)	(0.006)	Two	1.174	57.648
				(0.977)	(1.341)

Parameter Estimates						
Component	True	Posterior Mean	Posterior STD	True	Posterior Mean	Posterior STD
		Intercepts' Mean			Intercepts' Variance	
One	0	−0.026	0.070	0.01	0.078	0.028
		(0.022)	(0.001)		(0.002)	(0.001)
Two	−1	−1.000	0.064	0.04	0.082	0.029
		(0.021)	(0.001)		(0.003)	(0.001)
		Slopes' Mean			Slopes' Variance	
One	−1	−0.993	0.083	0.01	0.092	0.039
		(0.043)	(0.001)		(0.002)	(0.001)
Two	1	0.988	0.071	0.04	0.089	0.034
		(0.042)	(0.001)		(0.002)	(0.001)
		Intercepts and Slopes' Covariance				
One	0.000	−0.003	0.023			
		(0.002)	(0.001)			
Two	−0.028	−0.021	0.024			
		(0.002)	(0.001)			

The means across 50 simulations are reported. The simulation standard error for the means are reported in parenthesis.

TABLE 6.
Model Choice for the Simulation Study of the Logistic Regression Model

	Number of Mixture Components			
	One	Two	Three	Four
Log P(Data)	−2373.8	−2319.5	−2336.8	−2349.9
	(36.4)	(38.4)	(38.7)	(39.1)
Posterior Probability	0.000	1.000	0.000	0.000
	(0.000)	(1e−5)	(1e−5)	(0.000)
Choice Counts	0	50	0	0

"Choice Counts" is the number of times in 50 simulated data sets that the posterior probability was maximum for the corresponding mixture model. The measures are averaged across the 50 simulated data sets. Simulation standard errors are in parentheses.

5.2. Preferences for Personal Computers

The subjects for the study were 170 MBA students attending The University of Michigan Business School in 1994. Each student evaluated descriptions of 20 hypothetical personal computers. They were asked if they would consider purchasing the described computer. "Yes" was coded as "1", and "no" was coded as "0". The personal computers were described by 13 binary

factors in a conjoint analysis framework. After an initial analysis, the following factors were retained in the analysis: CPU, CD-ROM, software, and price. There were slow or fast speeds of CPU, the absence or presence CD-ROM, the absence or presence of bundled software, and low or high price levels. The lower level was coded as "−1", and the high level was coded as "1". The Bayesian mixture logistic regression was fitted with one to four components. MCMC was ran for 11,000 iterations, and the last 1000 iterates were used in the analysis. The Bayesian model selection procedure selected the two component solution, which had a posterior probability of one.

Table 7 reports the posterior means and standard deviations of the parameters for the models with one to three components. The four component model is not reported due to space limitations. The four component solution has class probabilities of 0.029, 0.091, 0.317, and 0.563. The three and four component solutions each have two moderately sized classes, while the remaining classes are very small. This behavior frequently occurs when the model is over fitted: the small classes usually consist of subjects whose estimated coefficients are outliers relative to their true component.

Table 7 then reports the posterior means of the component means $\{\theta_k\}$ for the regression coefficients. For the two component solution, subjects' mean coefficients within each class are nearly the same for CPU, CD-ROM, and software. The major distinguishing factors are the mean intercept and price coefficient. Holding the computer's features and price constant, the first group is more likely, on average, to purchase the computer because they have significantly larger inter-

TABLE 7.
Estimated Mixture Logistic Regression Model for Computer Preference Study

Components	One	Two		Three		
Class	1	1	2	1	2	3
Probability	1.000	0.291	0.709	0.118	0.371	0.511
	(0.000)	(0.030)	(0.030)	(0.021)	(0.027)	(0.032)
Coefficients' Means (θ)						
Intercept	−1.829	−0.673	−2.463	−0.076	−1.611	−2.322
	(0.139)	(0.116)	(0.123)	(0.173)	(0.163)	(0.174)
CPU	0.463	0.427	0.533	0.478	0.925	0.159
	(0.057)	(0.087)	(0.086)	(0.164)	(0.151)	(0.080)
CD-ROM	0.618	0.525	0.688	0.871	0.554	0.712
	(0.056)	(0.120)	(0.075)	(0.198)	(0.125)	(0.101)
Software	0.168	0.229	0.165	0.578	0.133	0.075
	(0.063)	(0.085)	(0.089)	(0.191)	(0.111)	(0.084)
Price	−1.870	−0.632	−2.488	−1.087	−2.557	−1.442
	(0.126)	(0.100)	(0.129)	(0.149)	(0.181)	(0.186)
Coefficients' Variances (Λ)						
Intercept	1.159	0.358	0.431	0.107	0.157	0.520
	(0.226)	(0.126)	(0.108)	(0.081)	(0.189)	(0.220)
CPU	0.104	0.088	0.208	0.056	0.364	0.041
	(0.045)	(0.042)	(0.080)	(0.032)	(0.146)	(0.017)
CD-ROM	0.155	0.240	0.103	0.191	0.048	0.195
	(0.056)	(0.118)	(0.051)	(0.173)	(0.023)	(0.080)
Software	0.066	0.100	0.075	0.062	0.053	0.062
	(0.020)	(0.045)	(0.037)	(0.039)	(0.026)	(0.027)
Price	0.897	0.108	0.412	0.070	0.277	0.553
	(0.207)	(0.057)	(0.114)	(0.045)	(0.132)	(0.251)

Parameter estimates are the posterior means, and the posterior standard deviations are in parentheses.

cepts. The second group is very sensitive to price. Next, Table 7 gives the within-class variances (diagonal of Λ_K) of the regression coefficients. For the two component solution, subjects in class one have more dispersion in their coefficients for CD-ROM and software, while the subjects in class two have more dispersion in their coefficients for the intercept, CPU, and price. The extent of the within-class variances indicates that an ordinary latent class model would need more than two classes to describe the heterogeneity in the coefficients, as will be documented.

Table 8 presents the within component correlation matrices for the two class solution. In class one, subjects' coefficient for intercept, CPU, CD-ROM, and software are positively correlated, and they are negatively correlated with price. In contrast, the preference for CPU is negatively correlated with the preferences for CD-ROM and software in class two, and preferences for CD-ROM are negative correlated with the intercept and positively correlated with the price coefficient.

As a basis of comparison to the MCMC based procedure illustrated above, we performed a latent class probit analysis utilizing the methodology in Wedel and DeSarbo (1995), a generalization of De Soete and DeSarbo (1991). Both the logit and probit models can be derived from random utility models (Luce, 1959; and McFadden, 1974) with the same deterministic component and different error structures. The probit model has normally distributed errors, while the logit model has Type II extreme value errors. When the deterministic component is linear in its parameters, the coefficients in the probit and logit models are the ratio of the coefficients in the deterministic component and the scaling factor for the error distributions. Consequently, the estimated coefficients from the probit and logit models are not directly comparable on a ratio scale because the scaling factors differ in the two models. However, the sign of the coefficients have the same meaning in both models: if a coefficient is positive, then the selection probability increases with the independent variable.

The latent class probit analysis was performed with one to five support points or classes. The goodness of fit heuristics for model selection are presented in Table 9. According to these information theory based heuristics, four classes appears to summarize most parsimoniously the structure in this data. In general, these information heuristics need not provide the same model choices as the posterior probabilities; although, BIC was developed as an asymptotic approxi-

TABLE 8.
Within Component Correlations for the Two Component Solution of the Computer Preference Study

	Class One				Class Two			
	Intercept	CPU	CD-ROM	Software	Intercept	CPU	CD-ROM	Software
CPU	0.58				0.19			
CD ROM	0.63	0.27			−0.41	−0.60		
Software	0.71	0.50	0.59		0.19	−0.45	0.19	
Price	−0.71	−0.53	−0.52	−0.60	−0.86	−0.14	0.35	−0.30

TABLE 9.
Goodness of Fit Heuristics for the Latent Class Probit Model of the Computer Preference Study

Number of Classes	Logarithm of Likelihood	Number of Parameters	BIC	CAIC
1	−3032.93	5	6106.51	6111.51
2	−1979.32	11	4048.09	4059.08
3	−1861.87	17	3861.98	3861.98
†4	−1717.94	23	3622.91	3645.91
5	−1713.35	29	3662.52	3691.52

†Denotes solution with minimum BIC and minimum CAIC.

TABLE 10.
Estimated Latent Class Probit Model for the Computer Preference Study

Class	One	Two	Three	Four
Probability	0.05	0.75	0.06	0.14
Intercept	1.175	−1.116	0.278	−0.451
CPU	0.673	0.531	0.599	0.596
CD ROM	0.581	0.543	0.614	0.686
Software	0.617	0.553	0.666	0.561
Price	−1.707	−1.021	−1.892	−1.490

mation to the posterior probabilities. Table 10 presents the parameter estimates for each of the four classes, which consists of a large class, a medium class, and two small classes. According to this solution, most of the heterogeneity appears to lie with respect to the intercept and price sensitivity coefficients. As demonstrated in the simulation data, it evidently takes more classes to account for the heterogeneity in the parameters when one does not permit within-class variability as compared to our proposed finite mixture model where only two classes were necessary. The analysis of the latent class probit model is not directly comparable to that of the mixture, logit model for at least three reasons: the link functions for the two models are different; the model choice criterion are different, and the inference methods are different—Bayesian versus maximum likelihood. Despite these difference, the qualitative results reinforce the finding that that latent class models require more classes to describe the heterogeneity in the subject-level parameters than mixture models.

6. Discussion

We have presented a finite mixture, random effects, generalized linear model where the individual-level coefficients for members of a class are a random sample from a normal distribution. We show that both the traditional latent class and random effects models are special cases. If there is only one class, then the proposed model simplifies to the usual random effects model, and it becomes the latent class model as the within-class variances approach zero.

A simulation study demonstrated that in the presence of substantial within-class heterogeneity, the ordinary latent class approach tends to result in an excessive number of classes. The large number of estimated classes results in many, very small groups and a large number of model parameters. Although random effects models require fewer parameters, they lack the flexibility of latent class models: they cannot describe non-normal heterogeneity, such as multi-modal distributions. The finite mixture, random effects model combines the flexibility of classical latent class models with the parsimony of random effects models.

We have outlined the numerical procedure for coefficient estimation that uses recent developments in Bayesian inference and Markov chain Monte Carlo. This approach also provides a method of selecting the number of mixture components via the computation of their posterior probabilities. A simulation study indicated that the method successfully identifies the known structure of the simulated data and that it is superior to an ad hoc three-stage procedure that clusters the individual-level maximum likelihood estimates. The paper presented an application concerning an analysis of revealed preferences for personal computers. Two latent classes were identified—the larger class is more price sensitive and less likely to purchase a computer with given features.

Further work needs to be accomplished in this area of research:

- More extensive Monte Carlo analyses should be undertaken in order to fully test the performance of the proposed hierarchical Bayes numerical approach;

- More extensive model comparison tests with ordinary latent class methods, maximum likelihood procedures, and naive multistage procedures, (e.g. clustering individual-level estimators) should be fruitful ground for additional research;
- Alterations in the Markov chain Monte Carlo estimation procedure need to be tested for potentially more efficient numerical methods for this particular framework;
- The convergence properties of the algorithm need to be further explored in a more systematic fashion;
- Additional empirical applications with comparisons with traditional methods must be explored in future research.

Appendix A: Joint Distribution

The joint distribution of the data and unknown parameters in hierarchical Bayes models is given by a series of condition distributions. Notation is problematic because one runs out of symbols for the different distributions. We will adopt the "bracket" notation in Gelfand and Smith (1990): "$[Y|X]$" means the conditional density of Y given X. The arguments provide the context that resolves the ambiguity of using "$[\cdot]$" for different density functions.

The conditional distributions of the hierarchical model are:

1. There are n subjects and m_i observations from subject i. The distribution for observation j of subject i is from the exponential family:

$$[y_{ij}|\beta_i, \phi_i] = \exp\left[\frac{y_{ij}h(\mathbf{x}'_{ij}\beta_i) - b[h(\mathbf{x}'_{ij}\beta_i)]}{a(\phi_i)} + c(y_{ij}, \phi_i)\right]$$
$$\text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m_i$$

   where $\mathbf{x}_{ij}$ is a $p \times 1$ vector of covariates; β_i is a $p \times 1$ vector of regression coefficients, and ϕ_i is the scale parameter. All observations are assumed to be mutually independent.
2. Define the random variables z_i to be $z_i = k$ if subject i belongs to class k for $k = 1, \ldots, K$. The distribution of z_i are the mixture probabilities:

$$\prod_{i=1}^{n}[z_i|\psi] = \prod_{i=1}^{n}\prod_{k=1}^{K}[z_i = k|\psi]$$
$$= \prod_{k=1}^{K}\psi_k^{n_k}$$
$$n_k = \sum_{i=1}^{n} I(z_i = k) \text{ for } k = 1, \ldots, K.$$

   where $I(z_i = k)$ is one if $z_i = k$ and zero otherwise, and n_k is the number of subjects assigned to class k.
3. If class membership is not known, the subject-specific regression coefficients, β_i, are mutually independent and identically distributed from the mixture of K multivariate normal distributions:

$$\left[\beta_i \mid (\psi_k, \theta_k, \Lambda_k)_{k=1}^{K}\right] = \sum_{k=1}^{K}\psi_k(2\pi)^{-\frac{p}{2}}|\Lambda_k|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k)\right]$$
$$\text{for } i = 1, \ldots, n$$

where θ_k is the mean vector of component k; Λ_k is the covariance matrix of component k, and ψ_k is the mixture probability of component k. Conditional on the class memberships $\{z_i\}$, the regression coefficients are random samples from normal distributions:

$$[\beta_i|\theta_k, \Lambda_k, z_i = k] = (2\pi)^{-\frac{p}{2}}|\Lambda_k|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k)\right] \text{ for } i = 1, \ldots, n.$$

4. The $\{\theta_k\}$ are mutually independent with a p dimensional normal distribution with prior mean vectors $\{u_{0,k}\}$ and prior covariance matrices $\{V_{0,k}\}$:

$$[\theta_k] = (2\pi)^{-\frac{p}{2}}|V_{0,k}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\theta_k - u_{0,k})'V_{0,k}^{-1}(\theta_k - u_{0,k})\right] \text{ for } k = 1, \ldots, K.$$

5. The $\{\Lambda_k\}$ are mutually independent $p \times p$ random covariance matrices from Inverted Wishart distributions with prior shaper parameters $\{f_{0,k}\}$ and prior scale parameters $\{G_{0,k}\}$:

$$[\Lambda_k] = c\frac{|G_{0,k}|^{f_{0,k}/2}}{|\Lambda_k|^{(f_{0,k}+p+1)/2}} \exp\left[-\frac{1}{2}\text{tr}(\Lambda_k^{-1}G_{0,k})\right] \text{ for } k = 1, \ldots, K$$

$$c^{-1} = 2^{f_{0,k}p/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma[(f_{0,k} + 1 - i)/2],$$

where $f_{0,k} \geq p$, and $G_{0,k}$ is a $p \times p$ positive definite matrix.

6. The mixture probabilities ψ have an ordered Dirichlet distribution with prior parameters $w_{0,1}, \ldots, w_{0,K}$:

$$[\psi] = \frac{\prod_{k=1}^{K}\psi_k^{w_{0,k}-1}}{\int_{S_K}\prod_{k=1}^{K}s_k^{w_{0,k}-1}ds_1ds_2\cdots ds_{K-1}}$$

$$S_K = \left\{(\psi_1, \ldots, \psi_K) : 0 \leq \psi_1 < \cdots < \psi_K \text{ and } \sum_{k=1}^{K}\psi_k = 1\right\}.$$

7. The subject-specific scale parameters $\{\phi_i\}$ are mutually independent from a normal distribution with mean α and standard deviation τ:

$$[\phi_i|\alpha, \tau^2] = \left(2\pi\tau^2\right)^{-\frac{1}{2}} \exp\left[-\frac{(\phi_i - \alpha)^2}{2\tau^2}\right] \text{ for } i = 1, \ldots, n.$$

8. The mean α of the subject-level scale parameters $\{\phi_i\}$ has a normal distribution with prior mean a_0 and prior variance d_0^2:

$$[\alpha] = \left(2\pi d_0^2\right)^{-\frac{1}{2}} \exp\left[-\frac{(\alpha - a_0)^2}{2d_0^2}\right].$$

9. The variance τ^2 of the subject-level scale parameters $\{\phi_i\}$ has an inverse gamma distribution with prior shape parameter $r_0/2$ and prior scale parameter $s_0/2$:

$$[\tau^2] = \frac{\left(\frac{s_0}{2}\right)^{\frac{r_0}{2}}}{\Gamma\left(\frac{r_0}{2}\right)}\left(\tau^2\right)^{-\left(\frac{r_0}{2}+1\right)} \exp\left[-\frac{s_0}{2\tau^2}\right].$$

The joint density of the data and unknown parameters is given by:

$$\left\{\prod_{i=1}^{n}\prod_{j=1}^{m_i}[y_{ij}|\beta_i,\phi_i]\right\}\times\left\{\prod_{i=1}^{n}\prod_{k=1}^{K}[\beta_i\mid\theta_k,\Lambda_k,z_i=k][z_i=k|\psi]\right\}$$
$$\times\left\{\prod_{k=1}^{K}[\theta_k][\Lambda_k]\right\}\times[\psi]\times\left\{\prod_{i=1}^{n}[\phi_i|\alpha,\tau^2]\right\}\times[\alpha][\tau^2].$$

Appendix B: Prior Parameters

The MCMC procedure of this paper requires proper prior distributions for the parameters. This section gives these prior parameters for the empirical examples. The prior parameters were set to be nearly noninformative. That is, the prior standard deviation was selected so that the range of variability in the prior distribution is much larger than the anticipated range of variability in the actual parameters. The consequence of this choice is that the prior distribution is fairly flat in the region where the likelihood function has most of its mass, and the posterior analysis gives more weight to the likelihood than the prior. These choices of the prior parameters are for illustration purposes. If a researcher had information based on previous studies, then that information could be introduced via the prior parameters.

The prior parameters are:

1. The heterogeneity in the scale parameters ϕ_i is described by a normal distribution with mean α and variance τ^2. The prior distribution of α has a normal distribution with prior mean $a_0 = 0$ and prior variance $d_0^2 = 10$. The prior distribution for τ^2 is an inverse gamma distribution with prior shape $r_0/2 = 1/2$ and prior scale $s_0/2 = 1/2$.
2. For each mixture component θ_k has a normal distribution with prior mean $u_{0,k} = 0$ and prior covariance matrix $V_{0,k} = 100I$, where I is the identity matrix.
3. Λ_k has an Inverted Wishart distribution with prior shape parameter $f_{0,k} = p + 1$ where p is the number of regression coefficients and prior scale parameter $G_{0,k} = pI$.
4. The mixture probabilities are from an ordered Dirichlet distribution with prior parameters $w_{0,k} = 1$.

A sensitivity analysis for the simulation study and computer survey indicated that the posterior distributions are insensitive to these prior parameters when there is a moderate number of members in each mixture component. If the number in class k is small, then the posterior distribution for Λ_k is sensitive to the choice of $f_{0,k}$ and $G_{0,k}$. The above choice for these parameters imply that large values of the diagonal elements of Λ_k are probable. If a class has a small number of members, the posterior means tend to overestimate the true values of Λ_k. One alternative, in this case, is to reconsider the likely range for the diagonal elements of Λ_k and to use a more informative prior that puts less mass on very large values.

Appendix C: Markov Chain Monte Carlo Algorithm

This section describes the MCMC procedure assuming that the number of mixture components is known to be K. Those readers who are not familiar with MCMC should see Gelfand and Smith (1990) or Smith and Roberts (1993) for an introduction.

Following Diebolt and Robert (1994), the key to analyzing latent class models is to generate a variable z_i for subject i that indicates class membership.

The sequence of draws from the full, conditional distributions follows:

1. The full conditional distribution of z_i is:

$$\begin{aligned}
&[z_i = k|\text{All other parameters}] \\
&\quad = [\beta_i|z_i = k, \theta_k, \Lambda_k][z_i = k] \\
&\quad \propto \psi_k|\Lambda_k|^{-1} \exp\left[-\frac{1}{2}(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k)\right].
\end{aligned}$$

We randomly generate z_i from the integers 1 to K where the probability that $z_i = k$ is:

$$\psi_{i,k} = \frac{|\Lambda_k|^{-1/2} \exp\left[-0.5(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k)\right]\psi_k}{\sum_{j=1}^{K}|\Lambda_j|^{-1/2} \exp\left[-0.5(\beta_i - \theta_j)'\Lambda_j^{-1}(\beta_i - \theta_j)\right]\psi_j}. \tag{8}$$

2. Given that $z_i = k$, we generate β_i. For the important, special case of linear regression with normally distributed errors given by (2), the full conditional distribution of β_i is:

$$\begin{aligned}
&[\beta_i|\text{All other parameters}] \\
&\quad \propto [\mathbf{y}_i|\beta_i, \sigma_i][\beta_i|z_i = k, \theta_k, \Lambda_k] \\
&\quad \propto \exp\left[-\frac{1}{2\sigma_i^2}(\mathbf{y}_i - X_i\beta_i)'(\mathbf{y}_i - X_i\beta_i) - \frac{1}{2}(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k)\right] \\
&\quad \propto \exp\left[-\frac{1}{2}(\beta_i - b_{i,k})'D_{i,k}^{-1}(\beta_i - D_{i,k})\right] \\
&D_{i,k} = (X_i'X_i/\sigma_i^2 + \Lambda_k^{-1})^{-1} \\
&b_{i,k} = D_{i,k}(X_i'\mathbf{y}_i/\sigma_i^2 + \Lambda_k^{-1}\theta_k).
\end{aligned}$$

Thus, we generate β_i from a multidimensional normal distribution with mean vector $b_{i,k}$ and covariance matrix $D_{i,k}$.

For the nonnormal, generalized linear model in (7), the full conditional distribution is:

$$\begin{aligned}
&[\beta_i|\text{All other parameters}] \\
&\quad \propto [\mathbf{y}_i|\beta_i, \phi_i][\beta_i|z_i = k, \theta_k, \Lambda_k] \\
&\quad \propto \exp\left[\sum_{j=1}^{m_i}\left\{\frac{y_{ij}h(\mathbf{x}_{ij}'\beta_i) - b[h(\mathbf{x}_{ij}'\beta_i)]}{a(\phi_i)}\right\} - \frac{1}{2}(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k)\right].
\end{aligned}$$

β_i is generated with the aid of a Metropolis step. One possibility would be to use a symmetric random walk (Gelman, Roberts, & Gilks, 1996) for a jump distribution. This class of procedures is simple to implement but does not use the structure of the problem. Instead, we propose a jump distribution that uses local information about the posterior distribution.

The log density of β_i, ignoring terms that do not depend on β_i, is

$$T(\beta_i) = \frac{1}{a(\phi_i)}\sum_{j=1}^{m_i}\left[y_{ij}h(\mathbf{x}_{ij}'\beta_i) - b[h(\mathbf{x}_{ij}'\beta_i)]\right] - \frac{1}{2}(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k).$$

We will use a quadratic approximation of T to generate a candidate value of β. The approximation will be centered around an approximate mode, which is updated on each iteration.

Assume that the maximum of T exists and that the Hessian is negative definite in a neighborhood of $\hat{\beta}_i$. The gradient or vector of partial derivatives of T is

$$\nabla T(\beta_i) = \frac{1}{a(\phi_i)} \sum_{j=1}^{m_i} \left[y_{ij} h_1(\mathbf{x}'_{ij}\beta_i) - b_1[h(\mathbf{x}'_{ij}\beta_i)] h_1(\mathbf{x}'_{ij}\beta_i) \right] \mathbf{x}_{ij} - \Lambda_k^{-1}(\beta_i - \theta_k),$$

where h_1 and b_1 are the first derivatives of h and b. The Hessian or matrix of second derivatives is:

$$H(\beta_i) = \frac{1}{a(\phi_i)} \sum_{j=1}^{m_i} \Big[y_{ij} h_2(\mathbf{x}'_{ij}\beta_i) - b_2[h(\mathbf{x}'_{ij}\beta_i)][h_1(\mathbf{x}'_{ij}\beta_i)]^2 - b_1[h(\mathbf{x}'_{ij}\beta_i)] h_2(\mathbf{x}'_{ij}\beta_i) \Big] \mathbf{x}_{ij}\mathbf{x}'_{ij} - \Lambda_k^{-1},$$

where h_2 and b_2 are the second derivatives of h and b. The quadratic approximation of T about the vector $\hat{\beta}$ is:

$$T(\beta_i) \approx T(\hat{\beta}_i) + \nabla T(\hat{\beta}_i)'(\beta_i - \hat{\beta}_i) + \frac{1}{2}(\beta_i - \hat{\beta}_i)' H(\hat{\beta}_i)(\beta_i - \hat{\beta}_i)$$

$$\approx c - \frac{1}{2} \left\{ \beta_i - [\hat{\beta}_i - H(\hat{\beta}_i)^{-1} \nabla T(\hat{\beta}_i)] \right\}' \left[-H(\hat{\beta}_i) \right] \left\{ \beta_i - [\hat{\beta}_i - H(\hat{\beta}_i)^{-1} \nabla T(\hat{\beta}_i)] \right\},$$

where c is a constant that does not depend on β.

If $\hat{\beta}_i$ is the mode of T, then $\nabla T(\hat{\beta})$ is 0. In general, the mode is not known, and its estimate is updated on each iteration. Suppose that $\hat{\beta}_i^{(u)}$ is the estimate of the mode on iteration u. Then it is updated on iteration $u+1$ with Newton-Raphson step:

$$\hat{\beta}_i^{(u+1)} = \hat{\beta}_i^{(u)} - H\left[\hat{\beta}_i^{(u)}\right]^{-1} \nabla T\left[\hat{\beta}_i^{(u)}\right].$$

The quadratic approximation of T on iteration $u+1$ becomes:

$$T(\beta) \approx c + \frac{1}{2}\left[\beta_i - \hat{\beta}_i^{(u+1)}\right]' H\left[\hat{\beta}_i^{(u)}\right]\left[\beta_i - \hat{\beta}_i^{(u+1)}\right].$$

Define $V^{(u)} = -H\left[\hat{\beta}_i^{(u)}\right]^{-1}$. On iteration $u+1$, the Metropolis step (Chib & Greenberg, 1995; Hastings, 1970; Tanner, 1993) then generates a candidate β_i^c from a normal distribution with mean $\hat{\beta}_i^{(u+1)}$ and covariance matrix $V^{(u)}$. Let $\beta_i^{(u)}$ be the value of β_i from the previous iteration of the chain. Then, $\beta_i^{(u+1)}$ is set to β_i^c with log probability (Hastings):

$$\min \Big\{ 0, T(\beta_i^c) - T(\beta_i^{(u)}) - \frac{1}{2}\left[\beta_i^{(u)} - \hat{\beta}_i^{(u+1)}\right]' \left[V^{(u)}\right]^{-1} \left[\beta_i^{(u)} - \hat{\beta}_i^{(u+1)}\right] + \frac{1}{2}\left[\beta_i^c - \hat{\beta}_i^{(u+1)}\right]' \left[V^{(u)}\right]^{-1} \left[\beta_i^c - \hat{\beta}_i^{(u+1)}\right] \Big\};$$

else $\beta_i^{(u+1)}$ is set to $\beta_i^{(u)}$. After an initial, transitory period, the values of $\hat{\beta}^{(u)}$ and $V^{(u)}$ stabilize. Then, the MCMC will run faster if $\hat{\beta}^{(u)}$ and $V^{(u)}$ are not updated on each iteration.

It can be shown that the above procedure results in a reversible Markov chain with a transition matrix that depends on the iteration through the approximate mode and Hessian. However, the posterior distribution satisfies the detailed balanced equations for each iteration, so that it is the limiting distribution of the Markov chain. Although this algorithm is general, specific cases, such as the multinomial-probit (Albert & Chib, 1993), have specialized algorithms that are more efficient.

3. Define $I(z_i = k)$ to be the indicator function, which is one if subject i is assigned to class k and zero otherwise. Suppose that n_k subjects are assigned to class k. Then the full conditional distribution of θ_k is:

$$
\begin{aligned}
&[\theta_k|\text{All other parameters}] \\
&\quad \propto \prod_{i=1}^{n}[\beta_i|z_i = k, \theta_k, \Lambda_k][\theta_k] \\
&\quad \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n} I(z_i = k)(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k) - \frac{1}{2}(\theta_k - u_{0,k})'V_{0,k}^{-1}(\theta_k - u_{0,k})\right] \\
&\quad \propto \exp\left[-\frac{1}{2}(\theta_k - u_{n,k})'V_{n,k}^{-1}(\theta_k - u_{n,k})\right] \\
&V_{n,k} = (n_k\Lambda_k^{-1} + V_{0,k}^{-1})^{-1} \\
&u_{n,k} = V_{n,k}(n_k\Lambda_k^{-1}\bar{\beta}_k + V_{0,k}^{-1}u_{0,k}) \\
&n_k = \sum_{i=1}^{n} I(z_i = k) \\
&\bar{\beta}_k = n_k^{-1}\sum_{i=1}^{n}\beta_i I(z_i = k).
\end{aligned}
$$

We then generate θ_k from a normal distribution with mean vector $u_{n,k}$ and covariance matrix $V_{n,k}$.

4. With the definitions in the previous item, the full conditional distribution of Λ_k is:

$$
\begin{aligned}
&[\Lambda_k|\text{All other parameters}] \\
&\quad \propto \prod_{i=1}^{n}[\beta_i|z_i = k, \theta_k, \Lambda_k][\Lambda_k] \\
&\quad \propto |\Lambda_k|^{-(n_k+f_{0,k}+p+1)/2}\exp\left[-\frac{1}{2}\text{tr}\left\{\Lambda_k^{-1}\left[\sum_{i=1}^{n} I(z_i = k)(\beta_i - \theta_k)(\beta_i - \theta_k)'\right]\right\}\right] \\
&\qquad \times \exp\left[-\frac{1}{2}\text{tr}\left\{\Lambda_k^{-1}G_{0,k}\right\}\right] \\
&\quad \propto |\Lambda_k|^{-(f_{n,k}+p+1)/2}\exp\left[-\frac{1}{2}\text{tr}\left\{\Lambda_k^{-1}G_{n,k}\right\}\right] \\
&f_{n,k} = f_{0,k} + n_k \\
&G_{n,k} = G_{0,k} + \sum_{i=1}^{n} I(z_i = k)(\beta_i - \theta_k)(\beta_i - \theta_k)'.
\end{aligned}
$$

We generate Λ_k from an Inverted Wishart distribution with shape parameter $f_{n,k}$ and scale matrix $G_{n,k}$.

5. The full conditional distribution of the mixture probabilities ψ is:

$$[\psi|\text{All other parameters}] \propto \prod_{i=1}^{n}\prod_{k=1}^{K}[z_i = k|\psi][\psi]$$

$$\propto \prod_{k=1}^{K} \psi_k^{w_{n,k}} I(\psi_1 < \cdots < \psi_K)$$

$$w_{n,k} = w_{0,k} + n_k$$

$$n_k = \sum_{i=1}^{n} I(z_i = k),$$

where n_k is the number of subjects assigned to class k. Thus, the full conditional distribution of the mixture probabilities is an ordered Dirichlet distribution.

One method of generating an standard Dirichlet distribution is to first generate K random deviate from appropriate gamma distributions, and set the probabilities to the ratio of each deviate to their sum of the deviates. A similar method can be used for the ordered Dirichlet distribution. The algorithm first generates ordered gamma random deviates from the density:

$$[X] \propto \prod_{k=1}^{K} x_k^{w_{n,k-1}} \exp(-x_k) I(X_1 \leq \cdots \leq X_K).$$

The ordered Dirichlet is obtained from $\psi_j = x_j / \sum_{k=1}^{K} x_k$.

Random deviates are generated from the ordered gamma distribution with "slice sampling" (Polson, 1996, p. 307, Example 5, and Damien, Wakefield, & Walker, 1999), which is a Markov chain method to decompose a complex density into the product of uniform and exponential densities. Slice sampling introduces K uniform random variables, V_k, such that their joint density with the ordered gamma random variables is:

$$[X, V] \propto I(v_1 \leq x_1^{w_{n,1}-1}) \exp(-x_1) \prod_{k=2}^{K} I(v_k \leq x_k^{w_{n,k}-1}, v_{k-1} \leq v_k) \exp(-x_k).$$

The marginal density of X is the ordered gamma distribution. Slice sampling first generates V given X and then X given V. Given X, the conditional distribution of V_k is uniform on $[0, X_k^{w_{n,k}-1}]$, so $V_k = X_k^{w_{n,k}-1} U_k$ where U_k is uniform on $[0, 1]$. Given V_1, the conditional density of X_1 is:

$$[X_1 | V_1] \propto \exp(-x_1) \text{ for } x_1 \geq v_1^{\frac{1}{w_{n,1}-1}}.$$

Given V_k and X_{k-1}, the conditional density of X_k is

$$[X_k | V_k, X_{k-1}] \propto \exp(-x_k) \text{ for } x_k \geq \max\left(v_k^{\frac{1}{w_{n,k}-1}}, x_{k-1}\right).$$

These conditional distributions are truncated, exponential distributions and are easily generated by inverting their cumulative distributions function.

In practice, this method sometimes causes the Markov chain to stall because the current values of n_k are inconsistent with the ordering of the ψ. If this state persists for consecutive iterations during the transitory stage of the Markov chain, then the algorithm for this paper reorders the classes according to n_k and continues.

6. The scale parameter ϕ_i has full conditional distribution:

$$[\phi_i | \text{All other parameters}]$$

$$\propto [\mathbf{y}_i | \beta_i, \phi_i][\phi_i | \alpha, \tau]$$

$$\propto \exp\left[\sum_{j=1}^{m_i} \left\{ \frac{y_{ij} h(\mathbf{x}_{ij}' \beta_i) - b[h(\mathbf{x}_{ij}' \beta_i)]}{a(\phi_i)} + c(y_{ij}, \phi_i) \right\} - \frac{(\phi_i - \alpha)^2}{2\tau^2} \right].$$

The scale parameter ϕ_i is generated with the aid of a Metropolis step. Up to a constant, the natural logarithm of the full conditional is:

$$S(\phi_i) = \sum_{j=1}^{m_i} \left[\frac{y_{ij} h(\mathbf{x}'_{ij}\beta_i) - b[h(\mathbf{x}'_{ij}\beta_i)]}{a(\phi_i)} + c(y_{ij}, \phi_i) \right] - \frac{1}{2\tau^2}(\phi_i - \alpha)^2.$$

S is approximated by a quadratic function in the same fashion that T was in item 2. The first and second derivatives are:

$$\frac{d}{d\phi_i} S(\phi_i) = -SE \frac{a_1(\phi_i)}{a(\phi_i)^2} + \sum_{j=1}^{m_i} c_1(y_{ij}, \phi_i) - \frac{1}{\tau^2}(\phi_i - \alpha)$$

$$\frac{d^2}{d\phi_i^2} S(\phi_i) = -SE \left[\frac{a_2(\phi_i)}{a(\phi_i)^2} - 2\frac{a_1(\phi_i)^2}{a(\phi_i)^3} \right] + \sum_{j=1}^{m_i} c_2(y_{ij}, \phi_i) - \frac{1}{\tau^2}$$

$$SE = \sum_{j=1}^{m_i} (y_{ij} h(\mathbf{x}'_{ij}\beta_i) - b[h(\mathbf{x}'_{ij}\beta_i)])$$

$$a_1(\phi_i) = \frac{d}{d\phi_i} a(\phi_i) \text{ and } a_2(\phi_i) = \frac{d^2}{d\phi_i^2} a(\phi_i)$$

$$c_1(y_{ij}, \phi_i) = \frac{\partial}{\partial \phi_i} c(y_{ij}, \phi_i) \text{ and } c_2(y_{ij}, \phi_i) = \frac{\partial^2}{\partial \phi_i^2} c(y_{ij}, \phi_i).$$

Next, expand S about $\hat{\phi}_i$, the maximum of S. Define

$$v = -\left[\frac{d^2}{d\phi_i^2} S(\hat{\phi}_i) \right]^{-1}.$$

Then

$$S(\phi_i) \approx S(\hat{\phi}_i) + \left[\frac{d}{d\phi_i} S(\hat{\phi}_i) \right](\phi_i - \hat{\phi}_i) - \frac{1}{2v}(\phi_i - \hat{\phi}_i)^2$$

$$\approx c - \frac{1}{2v} \left(\phi_i - \left[\hat{\phi}_i + v \frac{d\, S(\hat{\phi}_i)}{d\phi_i} \right] \right)^2.$$

Because $\hat{\phi}_i$ is not known, let $\hat{\phi}_i^{(u)}$ be its estimate on iteration u. This estimate can be updated on iteration $u + 1$ by:

$$\hat{\phi}_i^{(u+1)} = \hat{\phi}_i^{(u)} + v^{(u)} \frac{d\, S\left(\hat{\phi}_i^{(u)}\right)}{d\phi_i}$$

$$v^{(u)} = -\left[\frac{d^2}{d\phi_i^2} S\left(\hat{\phi}_i^{(u)}\right) \right]^{-1},$$

in which case

$$S(\phi_i) \approx c - \frac{1}{2v^{(u)}}(\phi_i - \hat{\phi}_i^{(u+1)})^2.$$

On iteration $u + 1$, we generate a candidate ϕ_i^c from a normal distribution with mean $\hat{\phi}_i^{(u+1)}$ and variance $v^{(u)}$. If $\phi_i^{(u)}$ is the present value on iteration u, then set $\phi_i^{(u+1)} = \phi_i^c$ with log

probability:

$$\min\left\{0, S(\phi_i^c) - S(\phi_i^{(u)}) - \frac{1}{2v^{(u)}}\left[\phi_i^{(u)} - \hat{\phi}_i^{(u+1)}\right]^2 + \frac{1}{2v^{(u)}}\left[\phi_i^c - \hat{\phi}_i^{(u+1)}\right]^2\right\};$$

else retain $\phi_i^{(u)}$ for $\phi_i^{(u+1)}$.

7. The full conditional distribution of α is:

$$\begin{aligned}
[\alpha|\text{All other parameters}] &\propto \prod_{i=1}^{n}[\phi_i|\alpha, \tau][\alpha] \\
&\propto \exp\left[-\frac{1}{2\tau^2}\sum_{i=1}^{n}(\phi_i - \alpha)^2 - \frac{1}{2d_0^2}(\alpha - a_0)^2\right] \\
&\propto \exp\left[-\frac{1}{2d_n^2}(\alpha - a_n)^2\right] \\
d_n^2 &= (n\tau^{-2} + d_0^{-2})^{-1} \\
a_n &= d_n^2\left(\tau^{-2}\sum_{i=1}^{n}\phi_i + d_0^{-2}a_0\right).
\end{aligned}$$

We thus generate α from a normal distribution with mean a_n and variance d_n^2.

8. The full conditional distribution of τ^2 is:

$$\begin{aligned}
[\tau^2|\text{All other parameters}] &\propto \prod_{i=1}^{n}[\phi_i|\alpha, \tau^2][\tau^2] \\
&\propto \left(\tau^2\right)^{-n/2-r_0/2-1}\exp\left[-\frac{\sum_{i=1}^{n}(\phi_i - \alpha)^2}{2\tau^2} - \frac{s_0}{2\tau^2}\right] \\
&\propto \left(\tau^2\right)^{-r_n/2-1}\exp\left[-\frac{s_n}{2\tau^2}\right] \\
r_n &= r_0 + n \\
s_n &= s_0 + \sum_{i=1}^{n}(\phi_i - \alpha)^2.
\end{aligned}$$

We generate τ^2 from an inverse gamma distribution with shape parameter $r_n/2$ and scale parameter $s_n/2$.

The algorithm, after an initial transitory period, produces dependent draws from the posterior distribution of the parameters. These draws can then be used to estimate posterior expectations. For instance, let Ω be a parameter of interest, and let $\Omega^{(u)}$ be the draw on the u-th iteration of MCMC. Then the posterior mean of a function g of Ω can be estimated by:

$$E[g(\Omega)|\text{ Data }] = \frac{1}{U - B}\sum_{u=B+1}^{U} g(\Omega^{(u)}),$$

where the last U—B of the U iterations are used.

Open questions are the selection of U and B and appropriate convergence criterion, which is beyond the scope of this paper. See, for example, Gelman and Rubin (1992) and Geyer (1992) along with a discussion, Polson (1996), and Roberts and Polson (1994).

Appendix D: Marginal Densities

This appendix provides details about the choice of g_K and estimates of its parameters for section 4. K is the number of components in the mixture distribution. g_K consists of the product of the following densities:

1. $g_K(\beta_i)$ is a multivariate normal density, and its mean and covariance matrices are estimated from the MCMC draws of β_i;
2. $g_K(z_i)$ is a multinomial distribution with probabilities ψ;
3. $g_K(\theta_k)$ is the multivariate normal density, and its mean and covariance matrix are estimated from the MCMC draws of θ_k;
4. $g_K(\phi_i)$ is the normal density, and its mean and variance are estimated form the MCMC draws of ϕ_i;
5. $g_K(\alpha, \ln(\tau^2))$ is a bivariate normal density, and its mean and covariance matrix are estimated from the MCMC draws of α and $\ln(\tau^2)$;
6. $g_K(\Lambda_k)$ is an Inverted Wishart density with h_k degrees of freedom and scale matrix H_k so that $E(\Lambda_k^{-1}) = h_k H_k^{-1}$. If one knew that n_k subjects belonged to class k, then the posterior degrees of freedom for Λ_k would be $f_{0,k} + n_k$ where $f_{0,k}$ is the prior degrees of freedom. Because subject classifications are not known, a reasonable choice of h_k is $f_{0,k} + \tilde{\psi}_k n$ where $\tilde{\psi}_k$ is the MCMC estimate of the posterior mean of ψ_k. A method of moments estimator of H_k is $h_k \left(\tilde{\Lambda}_k^{-1}\right)^{-1}$ where $\tilde{\Lambda}_k^{-1}$ is the MCMC estimator of the posterior mean of Λ_k^{-1};
7. $g_K(\psi)$ is a Dirichlet density with parameters v_k. Define $v = \sum_{k=1}^K v_k$. Then $E(\psi_j) = v_j/v$; $\text{Var}(\psi_j) = (v+1)^{-1}[E(\psi_j) - E(\psi_j)^2]$, and $\sum_{k=1}^K \text{Var}(\psi_k) = (v+1)^{-1}[1 - \sum_{k=1}^K E(\psi_k)^2]$. A method of moments estimator of v_k can be based on:

$$v = \frac{1 - \sum_{k=1}^K E(\psi_k)^2}{\sum_{k=1}^K Var(\psi_k)} - 1$$

$$v_j = vE(\psi_j)$$

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In N. Petrov & F. Csadki (Eds.), *Proceedings of the Second International Symposium of Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of American Statistical Association, 88*, 669–679.

Allenby, G. M., Arora, N., & Ginter, J. L. (1998). On the Heterogeneity of Demand. *Journal of Marketing Research, 37*(November), 384–389.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extension. *Psychometrika, 52*, 345-370.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of American Statistical Association, 88*, 8–25.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, 57*, 473–484.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of American Statistical Association, 90*, 1313–1321.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician, 49*, 327–335.

Damien, P., Wakefield, J. C., & Walker, S. (1999). Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables. *Journal of the Royal Statistical Society, Series B, 61*, 331-334.

Diebolt, J., & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B, 56*, 362–375.

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

DeSarbo, W. S., & Cron W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification, 5*, 248–282.

DeSarbo, W. S., Wedel, M., Vriens, M., & Ramaswamy, V. (1992). Latent class metric conjoint analysis. *Marketing Letters, 3*, 273–288.

DeSarbo, W. S., Ramaswamy, V., Reibstein, D. J., & Robinson, W. T. (1993). A latent pooling methodology for regression analysis with limited time series of cross sections: A PIMS data application. *Marketing Science, 12*, 103–124.

De Soete, G., & DeSarbo, W. S. (1991). A latent class probit model for analyzing pick any/*n* data. *Journal of Classification, 8*, 45–63.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman and Hall.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B, 56*, 501–514.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling–based approaches to calculating marginal densities. *Journal of American Statistical Association, 85*, 398–409.

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. M. F. Smith (Eds.), *Bayesian statistics 5* (pp. 599–607). Cambridge: Oxford University Press.

Gelman, A., & Rubin, D. B. (1992). Iterative simulation using single and multiple sequences. *Statistical Science, 7*, 457–511.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science, 7*, 473–511.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identified and unidentified models. *Biometrika, 61*, 215–231.

Hastings, W. K (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97–109.

Hosmer, D. W. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines. *Communications in Statistics, Series B, 3*, 995–1006.

Jeffreys, H. (1961). *Theory of probability* (3rd. ed.). Cambridge: Oxford University Press.

Jones, P. N., & McLachlan, G. J. (1992). Fitting finite mixture models in a regression context. *Australian Journal of Statistics, 43*, 233–240.

Kamakura, W. A. (1991). Estimating flexible distributions of ideal-points with external analysis of preference. *Psychometrika, 56*, 419–448.

Kamakura, W. A., & Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research, 26*, 379–390.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of American Statistical Association, 90*, 773–795.

Lenk, P. J., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science, 15*(2), 173–191.

Lewis, S. M. & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace—Metropolis estimator. *Journal of American Statistical Association, 92*, 648–655.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: John Wiley & Sons.

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

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. New York: Chapman and Hall.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. In P. Zarembda (Ed.), *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics, Series B, 8*, 343–366.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions, Series A, 185*, 71–110.

Polson, Nicholas G. (1996). Convergence of Markov chain Monte Carlo algorithms. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. M. F. Smith (Eds.), *Bayesian statistics 5* (pp. 297–312). Cambridge: Oxford University Press.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of American Statistical Association, 67*, 306–310.

Quandt, R. E., & Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regression. *Journal of American Statistical Association, 73*, 730–738.

Roberts, G. O., & Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society, Series B, 56*, 377–384.

Seber, G. A. F. (1984). *Multivariate observations*. New York: John Wiley & Sons.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, 55*, 3–23.

Tanner, M. A. (1993). *Tools for statistical inference* (Lecture Notes in Statistics 67). 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.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of American Statistical Association, 90*, 614–618.

Wang, P. M., Cockburn, I. M., & Puterman, M. L. (1998). "Analysis of Patent Data—A Mixed Poisson Regression Model. *Journal of Business and Economic Statistics, 16*, 27–41.

Wang, P. M., & Puterman, M. L. (in press). Mixed logistic regression models. *Journal of Agricultural, Biological, and Environmental Statistics*.

Wang, P. M., Puterman, M. L., Le, N., & Cockburn, I. (1996). Mixed Poisson regression with covariate dependent rates. *Biometrics, 52*, 381–400.

Wedel, M., & DeSarbo, W. S. (1993). A latent class binomial logit methodology for the analysis of paired comparison data: An application reinvestigating the determinants of perceived risk. *Decision Science, 24*, 1157–1170.

Wedel, M., & DeSarbo, W. S. (1994). A Review of recent developments in latent structure regression models. In R. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352-388), London: Blackwell Publishing.

Wedel, M., & DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification, 12*, 21–55.

Wedel, M., DeSarbo, W. S., Bult, J. R., & Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data with an application to direct mail. *Journal of Applied Econometrics, 8*, 397–411.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of American Statistical Association, 86*, 79–679.

Manuscript received 21 AUG 1995
Final version received 22 MAR 1999