Bayesian Hierarchical Mixture Models for High-Risk Births

James Holland Jones *
Department of Anthropology
Stanford University

Simon Jackman †
Department of Political Science
Stanford University

April 13, 2008

Abstract

Birthweight shows complex patterns of heterogeneity and has strong implications for infant mortality and later-life demographic outcomes. Using NCHS registration data from 1968-2005, we model the joint distribution of birthweight and gestational age as a two-component Gaussian mixture. The mixture has an intuitive interpretation: the first component represents the majority of the population and the second component represents a high-risk sub-population with lower mean birthweight and higher variance in both birthweight and gestational age. Using a Bayesian framework, we estimate the joint posterior distribution of the mixture model via MCMC simulation. The flexibility afforded by fitting the mixture model by the Gibbs sampler allows us to model the (binary) indicator for component membership as a function of covariates. Our interest focuses primarily in mother's and father's race, their interaction, and proxies of SES available from birth certificate information. The model fits well, though the posterior distributions of most of the coefficients in the hierarchical model have wide credible intervals. In particular, we find no strong evidence that race predicts component membership. We conclude with a discussion of potentially productive extensions to the model.

*Correspondence Address: Department of Anthropology, 450 Serra Mall, Building 50, Stanford, CA 94305-2034, USA; phone: +1-650-723-4824, fax: +1-650-725-0605; email: jhj1@stanford.edu

†Department of Political Science, Encina Hall, Stanford, CA 94305-6044, USA; phone: +1-650-723-4760; fax: +1-650-723-1808; e-mail: jackman@stanford.edu

1 Introduction

Low birthweight and young gestational age remain major risk factors for a variety of negative health outcomes, serving overwhelmingly as the most important predictors of perinatal mortality (Karn and Penrose 1951; Sappenfield et al. 1987; Wilcox 2001; Boardman et al. 2002; Alexander et al. 2008). While low birthweight exerts its greatest health impact early in life, it has also been implicated in a variety of health problems downstream in the life-cycle (Swamy et al. 2008). Low birthweight is universally recognized as a a major risk factor for negative health outcomes, however, high birthweight has also increasingly been implicated Van Valen and Mellin (1967); Wilcox and Russell (1983b). Gestational diabetes is a significant source of morbidity and mortality.

Distributions of birthweight (and gestational age) are characterized by long left tails. This feature of can be seen by comparing a kernel density estimate from an empirical birthweight distribution with a Gaussian density with the same mean and standard deviation as the empirical distribution, as shown in figure 1 for the California male birth cohort of 2000. The empirical density is more peaked around its mode and has narrower shoulders in addition to the heavy left tail. A number of population geneticists, including Karn and Penrose (1951) and Van Valen and Mellin (1967), have argued that strong stabilizing selection is exerted on human birthweight, showing that the probability of perinatal mortality is minimum just above the mean of the birthweight distribution.

Following the initial suggestion of Fryer et al. (1984) that birth cohorts may be composites of multiple heterogeneous subpopulations, Gage and co-workers (e.g., Gage and Therriault 1998; Gage 2000) have developed a rigorous methodology for understanding heterogeneity in birthweight and gestational age in birth cohorts using finite mixture models. Gage modeled cohort birthweight distributions as two-component Gaussian mixtures. These mixtures were either univariate, in the case of birthweight, or multivariate, in the case of birthweight and gestational age (Gage 2003). For example, Gage and Therriault (1998) disaggregated the total 1988 birth cohort New York state by race and sex and fit separate two-component Gaussian mixtures to each racial

group (white, black, Hispanic) and sex by direct maximization of the likelihood function. He fit three different variations on the Gaussian mixture: (1) a model with fixed common variances, (2) a model with free-to-vary variances, and (3) a model with no second Gaussian component (i.e., a single Gaussian distribution), and used likelihood ratio tests to assess the goodness of fit of the different models. The best-fitting model was consistently the two-component mixture with distinct means and variances.

The consistent finding from Gage's analysis (e.g., Gage and Therriault 1998; Gage 2000, 2003; Gage et al. 2004) is that birthweight distributions are characterized by (1) a first component that contains the majority of births and (2) a second, smaller component that has a lower mean and substantially higher variance than the first component. One appealing feature of the two-component mixture is the interpretability of this result. The first component represents the majority of births in the population (referred to as the "predominant component" in Gage and colleagues' work), while the second component represents the high-risk births. It is important to note that this high-risk component, because of its high variance, includes all very low birthweights, most low birthweights and all very high birthweights. Of course, an implication of this is that a fraction of births with normal birthweight and gestational age are, in fact, members of the high-risk component.

While they do not employ an explicit mixture model, Wilcox and Russell (1983a) use the logic of the two components to explain the paradox that while there are more female births less than 2500 grams, males have higher perinatal mortality. Wilcox and Russell (1983a) characterize the birthweight distribution by a Gaussian main component and a residual component for the excess probability mass in the left tail. Births in the predominant component reflect those of an orderly ontogenetic process, while those of the residual component are necessarily more heterogeneous and potentially pathological. Wilcox and Russell note that most perinatal deaths are concentrated in this residual component of the birthweight distribution. They further note that the female main component has a lower mean, meaning that a larger proportion of small births fall into this main component for females than for males. It is the component membership

that defines risk, not the birthweight per se. Risk thus becomes a latent trait, inferable only from the manifest trait of birthweight (and potentially covariates).

The mortality paradox noted by Wilcox and Russell (1983a) highlights the need to better classify high-risk births. Gage (2003) noted that using the bivariate distribution of birthweight and gestational age improves the ability to discern membership in the high-risk component. A second strategy for better separating the high-risk from the predominant components is to use covariates that are associated with high-risk birth. There are a number of covariates collected on birth certificates and reported to NCHS that may be of use in modeling high-risk births. Notable among these are mother and father's race, mother's education, and mother's marital status.

Race and birthweight are clearly related in complex ways. For example, African-Americans are substantially more likely than whites to have low birthweight births, while Hispanic's are at moderately increased risk (Alexander et al. 2008; Boardman et al. 2002; Reichman et al. 2008). Racial disparities in infant mortality have, if anything, increased in the United States since the 1970s. As Gage and Therriault (1998) note, however, there is no simple relationship between birthweight, race, and perinatal mortality risk. For example, black neonates have lower weight-specific mortality rates than whites. African-Americans hold a disproportionate share of both low birthweight births and infant deaths, yet small African-American babies are less likely to die than white counterparts. This phenomenon has been termed the "pediatric paradox." Gage and co-workers employ a logic similar to that used by Wilcox and Russell (1983a) in resolving the pediatric paradox. Specifically, they note that it is membership in the high-risk component that confers the mortality risk. With a lower mean for the main component, small African-American infants are less likely to be in the high-risk component. Using a differential frailty argument (Vaupel et al. 1979), Gage et al. (2004) suggest that a larger proportion of high-risk pregnancies in African-American women spontaneously abort, leaving the remaining population of neonates more robust than their size alone would predict.

Race as a causal variable is highly problematic. First, racial groupings have very little

biological meaning (Brown and Armelagos 2001). Following Lewontin's pioneering study, we know that the great of human genetic variation is contained at the level of the individual, with only a small fraction explained by the traditional racial categories (Lewontin 1972). While "race" is not a useful biological category, it is nonetheless a major means by which people mediate their behavior toward others. Race becomes important because people make it so. A second issue with race is that the way that racial categories are recognized and recorded by government agencies (e.g., the Census Bureau, National Center for Health Statistics) have changed significantly in recent years. Birth certificates from the early part of the NCHS series of publicly available birth microdata include classifications of infant race. More recent birth certificates do not record the infant's race, rather the race of the mother and the father separately. The detailed racial classification of both mother and father allow us to investigate whether mother's and father's race might have independent effects on birthweight.

In this paper we investigate the impact of mother's race, father's race, and the interaction between the two on birthweight. Our analysis will concentrate on a dichotomized variable for both mother and father's race: white vs. non-white. We will also investigate the possible effect of other covariates including mother's education and marital status. We will control for known contributers to small birthweight such as plural birth and mother's age.

1.1 Finite Mixture Models

Pearson (1894) pioneered the use of finite mixtures of Gaussian distributions in his analysis of Weldon's crab data. Mixture models are a flexible and powerful tool for understanding heterogeneous distributions. Any non-degenerate distribution can be written as a mixture of some other distributions. Some commonly used distributions have natural mixture interpretations. For example, a t-distribution is a scale mixture of normals, while a negative binomial is a gamma mixture of Poisson distributions.

A mixture model can be represented as

$$f(x) = \int_{\Theta} p(x|\theta)g(\theta)d\theta,$$

where $p(x)$ is the target distribution and $g(\theta)$ is the mixing distribution. In Bayesian inference, this is known as a predictive distribution (prior or posterior depending upon what $g(\theta)$ is). Interest frequently focuses on discrete $g(\theta)$, with support on a finite number of points and particular interest has focused on 2-component mixtures.

Mixture models have played a large role in understanding biodemographic heterogeneity. Wood and colleagues, for example, use a two-state hazards approach to model heterogeneity in fertility in a natural fertility population (Wood et al. 1994). Vaupel and Carey (1993) use a gamma mixture of Weibull distributions to model heterogeneity in medfly mortality. Jones (in prep) has used finite mixtures of Weibull distributions to model patterns of heterogeneity in frailty with changing forces of mortality.

In the current application, the use of mixture models is more than simply statistical convenience. The work of Wilcox and Russell and Gage and colleagues suggests the existence of latent classes in birthweight, the membership of which modifies the risk of negative outcome. Our work thus focuses on developing methods to estimate latent class membership from the data available from NCHS.

2 Methods

2.1 Data

We use birth microdata available from the National Center for Health Statistics (NCHS) from 1968-2005 (National Center for Health Statistics 2002). We subsetted the data to look only at live births from California. The model is fit to the joint distribution of birthweight (in kilograms) and gestational age (in weeks) for singleton births. Gestational age estimates available from NCHS come from three potential sources: (1) they are computed using dates of birth and last normal menses, (2) they are imputed from last normal menses period, or (3) they represent clinical

estimates made by the birth attendant. Covariates included in the analysis are: (1) mother's race re-coded as white/non-white (2) mother's age, (3) mother's marital status, (4) mother's education (in years), (5) birth order re-coded as first birth/non-first, (6) father's race recoded as white/non-white. NCHS data are remarkably complete. However, there are a few observations for which missing values are present. We handled missing data via case-wise deletion.

2.2 Mixture model

We will assume that our data $\mathbf{y} = (y_1, \ldots, y_N)$ has a density that is a mixture of M components. For biometric data such as birthweight, we will use Gaussian densities, which we denote $\phi(y; \theta)$, where $\theta = (\mu, \Sigma)$ contains the two parameters of the multivariate normal density (the mean and covariance matrix respectively). The probability of observing the birthweight y_i is thus:

$$f(y_i) = \pi_1 \phi(y_i; \theta_1) + \pi_2 \phi(y_i; \theta_2) + \ldots + \pi_M \phi(y_i | \theta_M)$$

where the π_i are the mixing proportions and

$$\sum_{i=1}^{M} \pi_i = 1.$$

Likelihood The likelihood for this model is thus given by,

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{m=1}^{M} [\pi_m \phi(y_i; \theta_m)]^{\zeta_{im}},$$

where

$$\zeta_{im} = \begin{cases} 1 & \text{if the } i\text{th observation is in the } m\text{th component} \\ 0 & \text{otherwise} \end{cases}.$$

The ζ_{im} are unobserved indicators that assign a particular observation y_i to component density $j \in 1, \ldots, M$. The binary indicator ζ_{im} suggests the possibility of modeling its probability

as as a function of exogenous covariates.

We will model $\pi_{im} = Pr(\zeta_{im} = 1)$ as a function of some set of covariates $\mathbf{X}$. These covariates are taken from the birth certificate data and include mother's race, mother's education, mother's marital status, father's race, birth plurality, and birth order.

Component probabilities are defined by:

$$p(z_i = j) = \frac{f(y_i|\theta_j)}{\sum_{k=1}^{J} f(y_i|\theta_k)}$$

To estimate the model, we use a Gibbs sampler. Let the vector of model parameters be θ and let θ be divided up into d sub-vectors $\theta = (\theta_1, \ldots, \theta_d)$. At each iteration, t sample, in turn, each θ_j^t from its conditional distribution, given all the other components of θ: $p(\theta_j|\theta_{-j}^{t-1}, y)$

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^t, \ldots \theta_d^{t-1})$$

Estimating mixtures of multivariate normal distributions turns out to be a non-trivial technical challenge and standard software (e.g., BUGS, JAGS) was unable to fit the model because of the complexity of fitting a covariance matrix that varies across mixture components. However, the fact that we are ultimately dealing Gaussian distributions means that we can take advantage of conjugacy and the sampler is fairly straightforward to write de novo. We thus wrote the Gibbs sampler in R. This solution is clearly not ideal for the scaling of the problem (e.g., using the full 200,000 births that occur in California in a year). As a result, we sub-sampled the full dataset and fit the model to a sample of 5000 births.

We adopted non-informative prior distributions for all the coefficients of the hierarchical model, using $\beta_i \sim \mathcal{N}(0, 100) \; \forall \; i$. Other parameters similarly received non-informative priors.

We ran the sampler with a 1000 iteration burn-in followed by a run of 5000 samples. The model converges remarkably rapidly as revealed by the trace plots in figures 6 and 7. For more details of the sampler, see the Appendix.

3 Results

Figure 2 plots a perspective plot of the bivariate density of birthweight and gestational age for the California boys' birth cohort in the year 2000. The surface is estimated using a 2-dimensional kernel density smoother. Figure 3 presents the same plot for the girls' birth cohort.

The MCMC model converged rapidly. Trace plots for the components of the covariance matrix are presented in figure 6, while trace plots for the coefficients of the hierarchical model are plotted in figure 7.

Consistent with the results of Gage and co-workers, we find that the two-component Gaussian mixture fits the data well. Figure 4 shows the 95% credibility ellipse from 4000 draws from the marginal bivariate distributions of birthweight and gestational age. The model estimated that approximately 85% of the births in the year 2000 fell into the main component, leaving 15% in the high-risk component. The posterior mean of birthweight in the main component was $\mu_{11} = 3.52$ kg, while the posterior mean of birthweight in the high-risk component was $\mu_{21} = 3.06$ kg. The posterior mean of gestational age in the main component was $\mu_{11} = 39.20$ kg, while the posterior mean of gestational age in the high-risk component was $\mu_{21} = 37.81$ kg. The variances for the main component were $\sigma^2_{1\cdot} = (0.20, 2.00)$ while the covariance between birthweight and gestational age was 0.16. The variances for the high-risk component were $\sigma^2_{2\cdot} = (0.71, 20.59)$ while the covariance between birthweight and gestational age was 2.2. These results are summarized in table 1.

Table 2 presents the results from the fitting of the hierarchical mixture model to the 2000 NCHS data for California boys. A notable feature of these results, in general, is that the credible intervals on the estimates are fairly wide, indicating that there is a large degree of uncertainty. Two coefficients (non-white mother and non-white father) had large posterior means but the 95% interior intervals of the posterior simulations cross zero. The only coefficient with a 95% interval that does not cross zero is mother's education, the posterior mean of which is $\beta_{med} = -0.05$.

4 Discussion

Using birth certificate data from NCHS (National Center for Health Statistics 2002), we have employed a Bayesian hierarchical model to fit a mixture of bivariate Gaussian distributions to the joint distribution of birthweight and gestational age for a sample of boys' singleton births from California in 2000. Gaussian mixture models – both univariate and bivariate – have been employed to describe the long left-tailed distributions of birthweight quite extensively (Gage and Therriault 1998; Gage 2000, 2003; Gage et al. 2004; Fang et al. 2007). We extend this work by fitting a mixture of bivariate mixtures in a hierarchical fashion, thus allowing us to include covariate information in the fitting of the mixture.

Gage et al. (2004) suggest estimating a parametric mixture of logistic regressions to look at the joint effect of birthweight and perinatal mortality. Estimate

$$f((x,y);\theta,\beta) = g(y|x;\beta,\theta)h(x;\theta)$$

where $f(x,y)$ is the joint density of x and y, $g(y)$ is the density of y which takes the form of a logistic regression with coefficients β and $h(x)$ is the density of x (the Gaussian mixture). y is status (dead/alive) while x is birthweight. Gage does not include covariates such as mother's race, education, etc. in the logistic regression mixing distribution.

This logistic mixture of normal distributions is one way of representing a classification problem. Using Bayesian methodology, we have presented an alternative approach to understanding the heterogeneity of birth cohorts. Our method uses information on the joint distribution of birthweight and gestational age along with a range of covariates commonly found on birth certificates to estimate the latent structure of the heterogeneous population. The next step in this analysis is to use the linked death data available from NCHS to investigate the predictive power of membership in the latent high-risk component on epinatal mortality.

Bayesian estimation of a mixture of multivariate normal distributions carries some technical challenges. In particular, fitting a covariance matrix that varies across mixture components

appears to be a task currently beyond the capabilities of BUGS. Nonetheless, the sampler was easily constructed in R and its performance was surprisingly good. The analysis in this paper uses a subsample of California births from the year 2000. The full dataset for California contains over 230,000 births per sex. We are currently exploring the scalability of our sampler to the full problem.

Only one of the seven covariates had a coefficient with a 95% credible interval that did not cross zero. It is possible that a larger sample of births would allow more precise estimation of the hierarchical model's parameters. Furthermore, it is also possible that a two component mixture of bivariate normals is not the best model for the joint distribution of birthweight and gestational age. One possibility is that a three-component mixture might fit better, a possibility we are exploring.

An interesting feature discernible from figure 4 is the apparent differences in the covariance structure of the two components. The high-risk component shows a stronger correlation between birthweight and gestational age than the normal component. While this may partially be an artifact of the much greater variance in birthweight in the high-risk component, posterior simulations from the covariance matrices suggest that the differences in correlations are real. Figure 5 shows 5000 posterior draws from the correlation matrix for (a) the high-risk component, (b) the normal component, and (c) their difference.

The interpretation of this result seem fairly straightforward. Conditional on a "normal" pregnancy going full term, there is modest natural variation in birthweight. Normal births thus show a small correlation between gestational age and birthweight in large measure because of the restricted variation in gestational age among normal births. On the other hand, a live birth following a disturbed pregnancy can occur at a wide range of gestational ages. Longer periods in utero allow for more growth (indeed potentially pathological growth in the case of gestational diabetes), leading to a strong correlation between gestational age and birthweight in the high-risk component.

Gage and colleagues have fit mixture models to different racial groups separately. In our

analysis, race enters as a covariate in the hierarchical model as both mother's and father's racial identity. While the distributions of the coefficients of both covariates had relatively large posterior means, the 95% credible regions of both coefficients crossed zero, raising the question of how important race is for component membership.

We have chosen to focus our discussion of the two-component mixture of bivariate Gaussian distributions. We also performed the analysis for a two-component univariate Gaussian distributions, focusing solely on birthweight. For this model, race, mother's education, and mother's marital status all had much stronger effects in predicting component membership. This raises the possibility that modeling the joint distribution of birthweight and gestational age is the wrong approach to the problem. Clearly, further exploration of this problem is necessary.

A natural question that arises in an analysis such as the one we have presented is: are the results simply a statistical convenience or do our results represent an underlying biodemographic reality? Both Wilcox and Russell (1983a) and Gage et al. (2004) present compelling evidence that the mixture model reflects underlying biology. Gage et al. (2004) suggest that the mixture interpretation explains the so-called "pediatric paradox," namely, that low birthweight black babies have lower epinatal mortality rates than low birthweight white babies despite the higher overall mortality rate of black neonates. The explanation that Gage and colleagues give for this apparent paradox is that African American infants are more likely to be born into the high-risk component, where they have lower mortality, but have higher mortality in the non-compromised component (where a larger number of births of both groups occur). Without the mixture model, the heterogeneity explanation does not work.

Future work on this project includes: (1) scaling of the problem to the full sample of births for both sexes, (2) inclusion of the full range of dates 1968-2005. In addition to these tasks, we will use the linked infant death data available from NCHS to investigate component membership and it's role in infant mortality. Finally, geocoding is available for California births, allowing us to link births to rich census information, in particular, socioeconomic status of the mother and father.

References

Alexander, G. R., M. S. Wingate, D. Bader, and M. D. Kogan (2008). The increasing racial disparity in infant mortality rates: Composition and contributors to recent US trends. *American Journal of Obstetrics and Gynecology 198*(1), 51.e1.

Boardman, J. D., D. A. Powers, Y. C. Padilla, and R. A. Hummer (2002). Low birth weight, social factors, and developmental outcomes among children in the United States. *Demography 39*(2), 353–368.

Brown, R. and G. Armelagos (2001). Apportionment of racial diversity: A review. *Evolutionary Anthropology 10*, 34–40.

Fang, F., H. Stratton, and T. B. Gage (2007). Multiple mortality optima due to heterogeneity in the birth cohort: A continuous model of birth weight by gestational age-specific infant mortality. *American Journal of Human Biology 19*(4), 475–486.

Fryer, J., R. Hunt, and A. Simons (1984). Biostatistical considerations: The case for using models. In F. Falkner (Ed.), *Prevention of Perinatal Mortality and Morbidity*, pp. 9–30. Basel, Switzerland: Karger.

Gage, T. B. (2000). Variability of gestational age distributions by sex and ethnicity: An analysis using mixture models. *American Journal of Human Biology 12*(2), 181–191.

Gage, T. B. (2003). Classification of births by birth weight and gestational age: an application of multivariate mixture models. *Annals of Human Biology 30*(5), 589–604.

Gage, T. B., M. J. Bauer, N. Heffner, and H. Stratton (2004). Pediatric paradox: Heterogeneity in the birth cohort. *Human Biology 76*(3), 327–342.

Gage, T. B. and G. Therriault (1998). Variability of birth-weight distributions by sex and ethnicity: Analysis using mixture models. *Human Biology 70*(3), 517–534.

Karn, M. N. and L. S. Penrose (1951). Birth weight and gestation time in relation to maternal age, parity and infant survival. *Annals of Eugenics 16*(2), 147–.

Lewontin, R. (1972). The apportionment of human genetic diversity. *Evolutionary Biology 6*, 381–398.

National Center for Health Statistics (2002). Data file documentations, natality, 2000.

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

Reichman, N. E., E. R. Hamilton, R. A. Hummer, and Y. C. Padilla (2008). Racial and ethnic disparities in low birthweight among urban unmarried mothers. *Maternal and Child Health Journal 12*(2), 204–215.

Sappenfield, W. M., J. W. Buehler, N. J. Binkin, C. J. R. Hogue, L. T. Strauss, and J. C. Smith (1987). Differences in neonatal and postneonatal mortality by race, birth-weight, and gestational-age. *Public Health Reports 102*(2), 182–192.

Swamy, G., T. Østbye, and R. Skjærven (2008). Association of preterm birth with long-term sur vival, reproduction, and next-generation preterm birth. *JAMA 299*(12), 1429–1436.

Van Valen, L. and G. Mellin (1967). Natural selection in natural populations 7. New York babies (Fetal Life Study). *Annals of Human Genetics, London 31*, 109–127.

Vaupel, J. W. and J. R. Carey (1993). Comppositional interpretations of medfly mortality. *Science 260*(1666-1667).

Vaupel, J. W., K. G. Manton, and E. Stallard (1979). Impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*(3), 439–454.

Wilcox, A. and A. F. Russell (1983a). Birthweight and perinatal mortality: I. on the frequency distribution of birthweight. *International Journal of Epidemiology 12*(3), 314–318.

Wilcox, A. and I. Russell (1983b). Birthweight and perinatal mortality II. on weight-specific mortality. *International Journal of Epidemiology 12*(3), 319–325.

Wilcox, A. J. (2001). On the importance – and the unimportance – of birthweight. *International Journal of Epidemiology 30*(6), 1233–1241.

Wood, J. W., D. J. Holman, A. I. Yashin, R. J. Peterson, M. Weinstein, and M. C. Chang (1994). A multistate model of fecundability and sterility. *Demography 31*(3), 403–426.

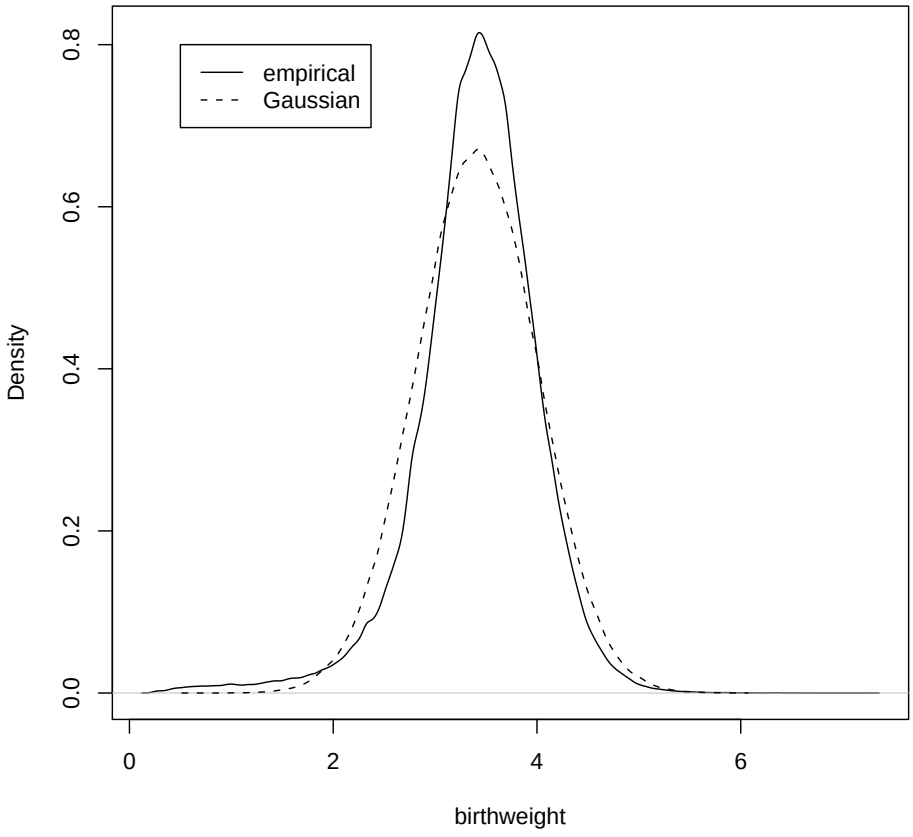

Figure 1: Comparison of the kernel density estimate for the California male birth cohort of 2000 to a Gaussian distribution with the same mean and standard deviation.

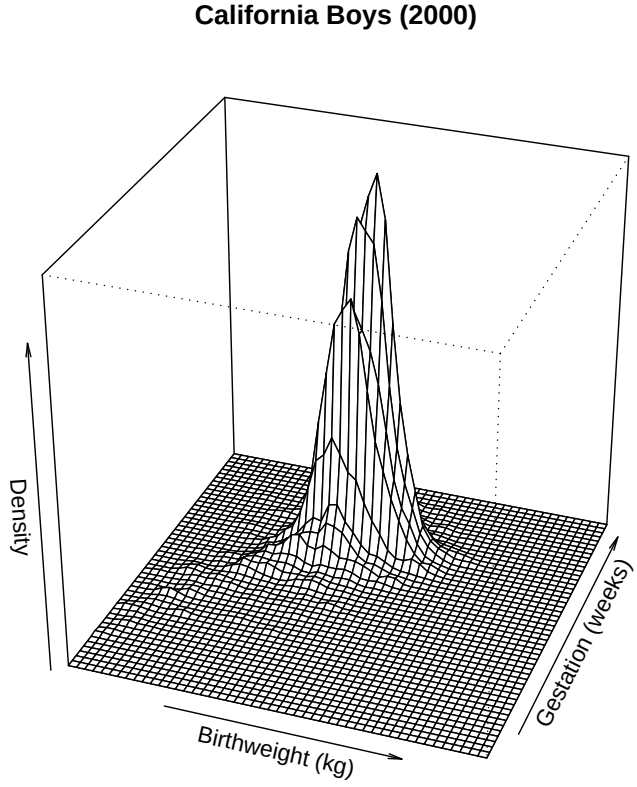

Figure 2: Perspective plot of the bivariate density of birthweight and gestational age for the California boys' birth cohort in the year 2000. The perspective plot is angled to emphasize the lower tail of the bivariate distribution.

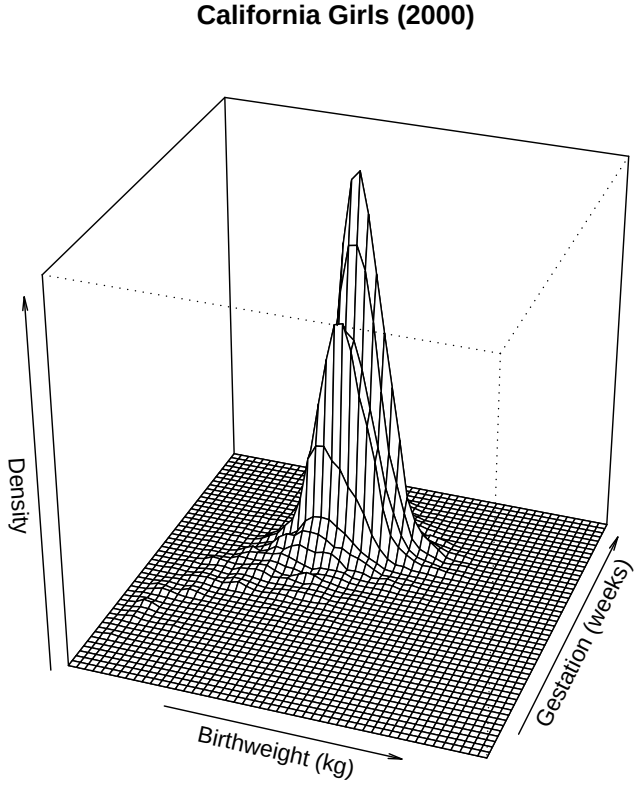

Figure 3: Perspective plot of the bivariate density of birthweight and gestational age for the California girls' birth cohort in the year 2000. The perspective plot is angled to emphasize the lower tail of the bivariate distribution.

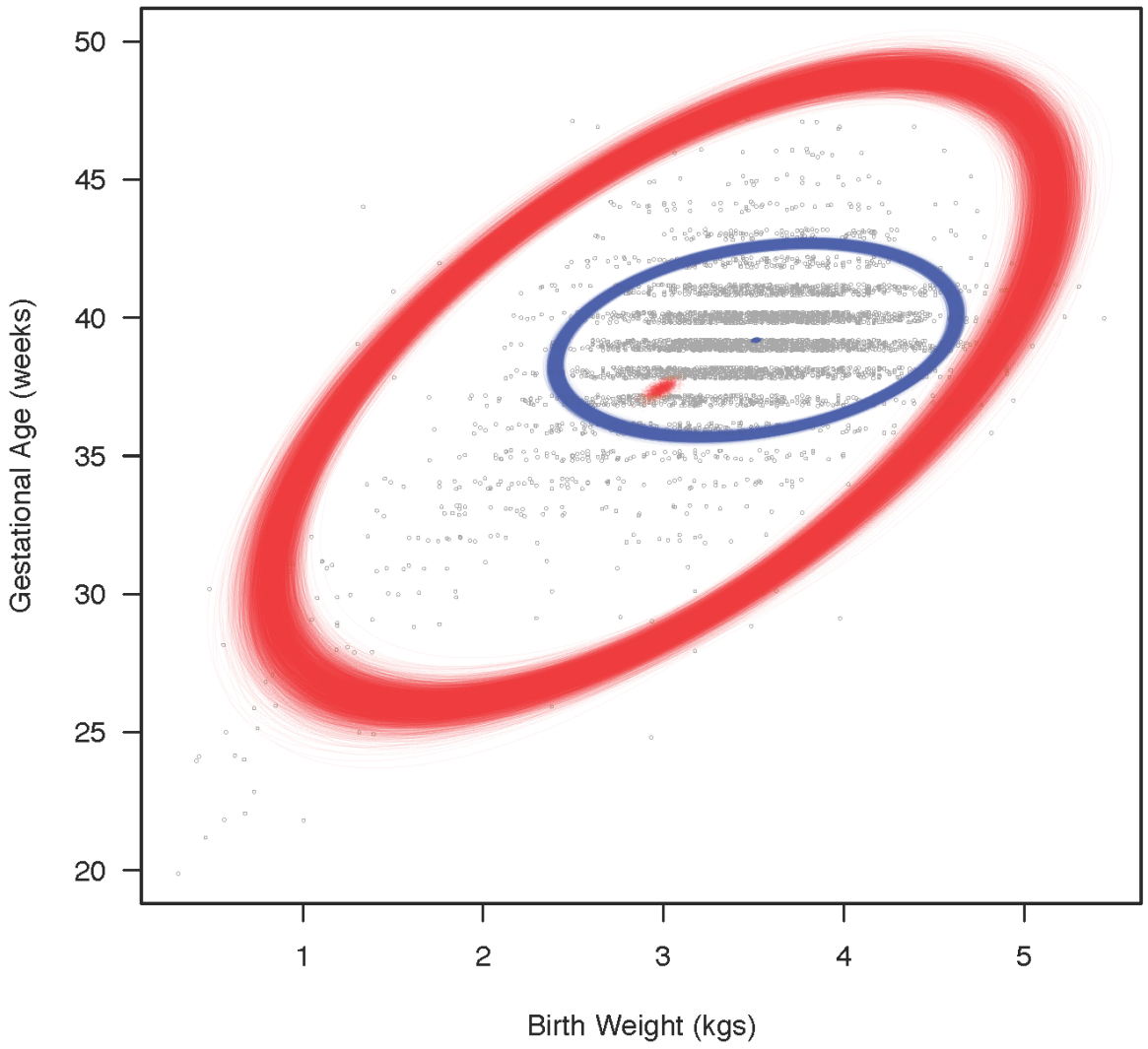

Figure 4: 95% credible intervals implied by covariance matrix and posterior means for two components based on 4000 draws from the posterior distributions. The predominant component is drawn in blue while the high-risk component is drawn in red. Draws of posterior means are given in the blue and red dots at the center of the ellipses.

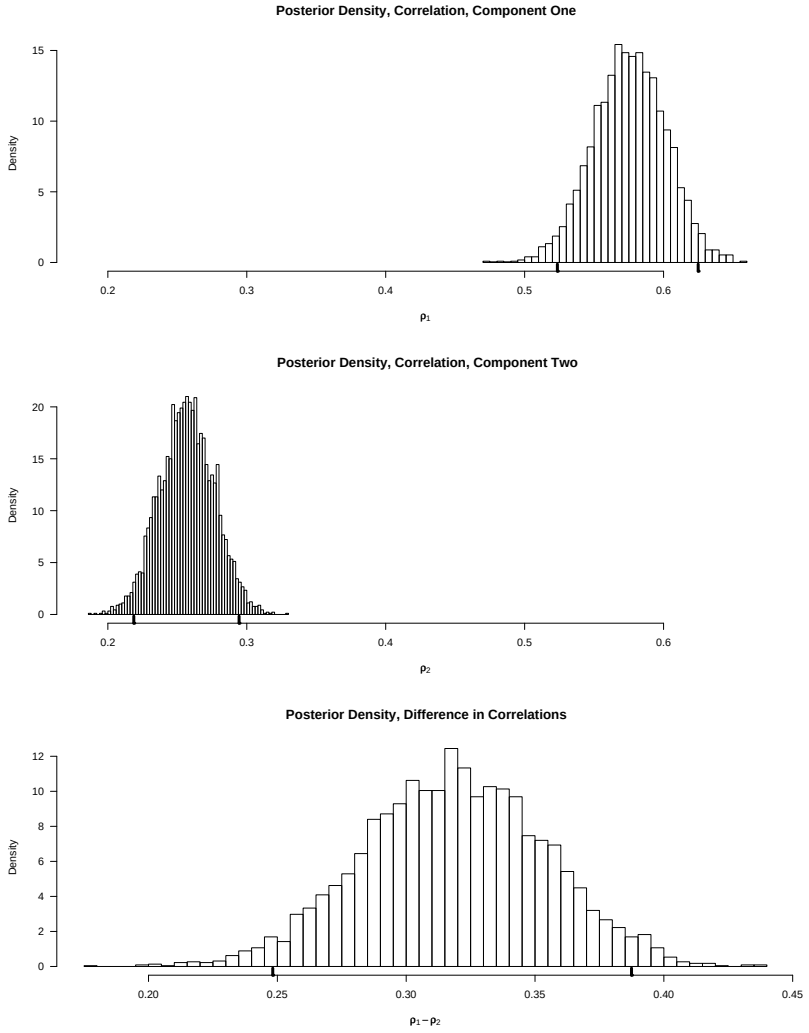
Posterior Density, Correlation, Component One
Density
ρ1
Posterior Density, Correlation, Component Two
Density
ρ2
Posterior Density, Difference in Correlations
Density
ρ1 − ρ2

Figure 5:

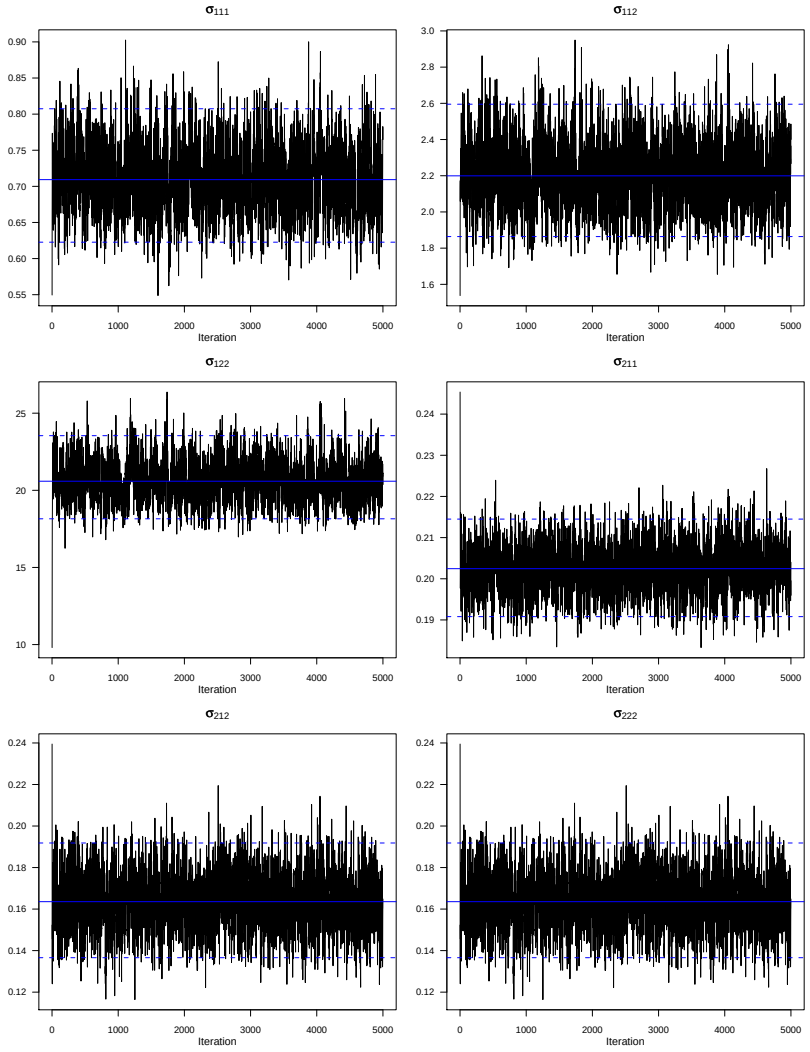

Figure 6: Trace plots for the components of the covariance matrix, $\mathbf{\Sigma}$.

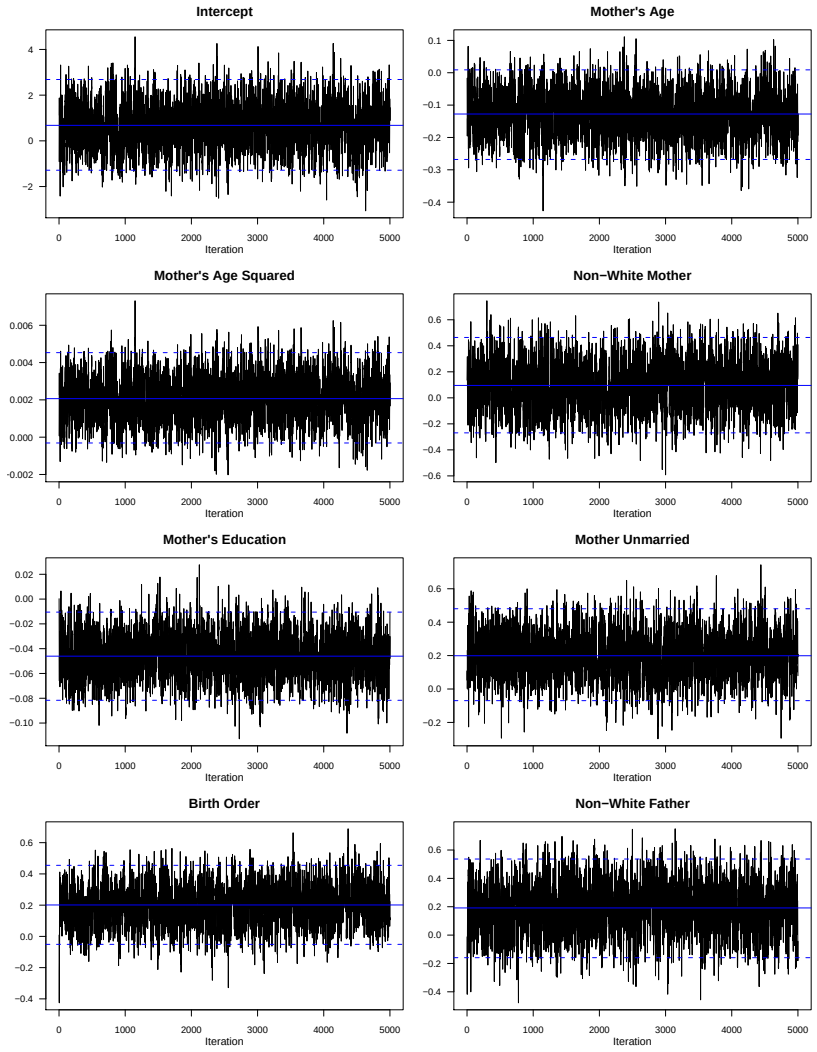

Figure 7: Trace plots for the coefficients β in the hierarchical model.

Parameter	Posterior Mean	Standard Deviation	2.5%	97.5%
μ Component 1, Birth Weight	3.06	0.04	2.96	3.14
μ Component 2, Birth Weight	3.52	0.01	3.51	3.54
μ Component 1, Gestational Age	37.81	0.19	37.43	38.18
μ Component 2, Gestational Age	39.20	0.03	39.15	39.26
$\sigma_{1,1}$, Component 1	0.71	0.05	0.62	0.81
$\sigma_{2,1}$, Component 1	2.20	0.19	1.86	2.60
$\sigma_{1,2}$, Component 1	2.20	0.19	1.86	2.60
$\sigma_{2,2}$, Component 1	20.59	1.39	18.15	23.54
$\sigma_{1,1}$, Component 2	0.20	0.01	0.19	0.21
$\sigma_{2,1}$, Component 2	0.16	0.01	0.14	0.19
$\sigma_{1,2}$, Component 2	0.16	0.01	0.14	0.19
$\sigma_{2,2}$, Component 2	2.00	0.09	1.84	2.18
Proportion in Component 1	0.18	0.01	0.16	0.21
Proportion in Component 2	0.82	0.01	0.79	0.84

Table 1: Posterior summaries for the mixture model.

Coefficient	Posterior Mean	Standard Deviation	2.5%	97.5%
Intercept	0.68	1.00	−1.29	2.68
Mother's Age	−0.13	0.07	−0.27	0.01
Mother's Age Squared	0.0021	0.0012	−0.0003	0.0045
Non-White Mother	0.10	0.19	−0.27	0.46
Mother's Education	−0.05	0.02	−0.08	−0.01
Mother Unmarried	0.20	0.14	−0.07	0.48
Birth Order	0.20	0.13	−0.05	0.45
Non-White Father	0.19	0.18	−0.16	0.54

Table 2: Posterior summaries for the coefficients for the hierarchical model.

Appendix: The MCMC Sampler

Model. Each observation $\mathbf{y}_i$, $i \in \{1, \ldots, n\}$ in component $j \in \{1, \ldots, J\}$ generates the likelihood contribution

$$p(\mathbf{y}_i|\zeta_{ij} = 1, \mu_j, \mathbf{\Sigma}_j) = \phi_M(\mathbf{y}_i|\mu_j, \mathbf{\Sigma}_j)$$

where ϕ_M is the multivariate (M-variate) normal density, $\mu_j \in \mathbb{R}^M$ is the mean of component j, and $\mathbf{\Sigma}_j$, the M-by-M covariance matrix for $\mathbf{y}$ in component j. Note that we work with $M = 2$ and $J = 2$. Conditional on the indicators ζ_{ij}, inference for μ_j and $\mathbf{\Sigma}_j$ is trivial, amounting to no more than conventional Bayesian inference for a multivariate normal mean and variance.

We use a hierarchical model for the unobserved component indicators $\zeta = \{\zeta_{ij}\}$. That is, with $j \in \{1, 2\}$ we have

$$\pi_i = \Pr(\zeta_{i1} = 1|\mathbf{x}_i) = F(\mathbf{x}_i\beta). \tag{1}$$

It is helpful to treat the ζ as additional parameters in the model, meaning that the full set of parameters is $\mathbf{\Theta} = \{\mu_1, \mu_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \zeta, \beta\}$. Note that conditional on ζ, the birth weight and gestational age data $\mathbf{y}$ are independent of the predictors of component membership $\mathbf{X}$, and the likelihood for this hierarchical model factors as

$$p(\mathbf{y}|\mathbf{\Theta}) = p(\mathbf{y}|\zeta, \mu_1, \mu_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2)p(\zeta|\mathbf{X}, \beta),$$

with the second term in the likelihood function given by equation 1. In a sense ζ acts like a set of auxiliary variables that facilitates computation of the likelihood function for $\mathbf{y}$, although here we will treat them as parameters.

We use vague normal priors for μ_1 and μ_2 and an improper uniform prior for both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. The hierarchical part of the model, equation 1, constitutes a prior for the ζ_{ij}. A prior over β completes the specification of the model: here we use a vague normal prior, with mean vector $\mathbf{0}$ and identity covariance matrix equal to 10^2 times an identity matrix. In summary then, our model is

$$\begin{aligned}
\mathbf{y}_i|\zeta_{ij} = 1 &\sim N(\mu_j, \mathbf{\Sigma}_j) && \text{(2a)}\\
\mu_j &\sim N(\eta_j, \mathbf{\Sigma}_{\mu_j}) && \text{(2b)}\\
p(\mathbf{\Sigma}_j) &\propto 1 && \text{(2c)}\\
\zeta_{i1} &\sim \text{Bernoulli}(\pi_i) && \text{(2d)}\\
\pi_i &= F(\mathbf{x}_i\beta) && \text{(2e)}\\
\beta &\sim N(\mathbf{b}_0, \mathbf{B}_0), && \text{(2f)}
\end{aligned}$$

and we set $\eta_j = \mathbf{0}$, $\mathbf{\Sigma}_{\mu_j} = 10^2 \cdot \mathbf{I}$, $\mathbf{b}_0 = \mathbf{0}$ and $\mathbf{B}_0 = 10^2 \cdot \mathbf{I}$, $j = 1, 2$.

Identification (invariance to label switching). To uniquely label the components of the mixtures, we also impose the constraint that that each element of μ_2 is greater than the corresponding element of μ_1. This constraint uniquely labels component "1" as the low-birth-weight/low-gestational-age component in these data. We impose this constraint via simple rejection sampling in the Gibbs sampling step for both μ_2 and μ_1. We initialize the sampler in a region of the parameter where this constraint is satisfied.

Posterior density. By Bayes Rule, and given the model structure described above, the posterior density for $\mathbf{\Theta}$ is

$$p(\mathbf{\Theta}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\zeta, \mu_1, \mu_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2)p(\mu_1)p(\mu_2)p(\mathbf{\Sigma}_1)p(\mathbf{\Sigma}_2)\, p(\zeta|\mathbf{X}, \beta)p(\beta). \quad (3)$$

We sample from this density using a Gibbs sampler, iteratively sampling from the following conditional densities:

1. $p(\mu_j|\mathbf{y}, \zeta, \mathbf{\Sigma}_j) = p(\mu_j)p(\mathbf{y}|\zeta, \mu_j, \mathbf{\Sigma}_j)$. Since $p(\mu_j)$ is the multivariate normal density in 2b and $p(\mathbf{y}|\zeta, \mu_j, \mathbf{\Sigma}_j)$ is the normal density in 2a, then we have the familiar precision-matrix weighted average of prior mean and likelihood

$$\begin{aligned}
\mu_j|\mathbf{y}, \zeta, \mathbf{\Sigma}_j &\sim N\left(\tilde{\mu}_j, \tilde{\mathbf{\Sigma}}_{\mu_j}\right) \quad \text{where} \\
\tilde{\mu}_j &= \tilde{\mathbf{\Sigma}}_{\mu_j}^{-1}(\mathbf{\Sigma}_{\mu_j}^{-1}\eta_j + \mathbf{V}_j^{-1}\bar{\mathbf{y}}_j), \\
\tilde{\mathbf{\Sigma}}_{\mu_j} &= \left(\mathbf{\Sigma}_{\mu_j}^{-1} + \mathbf{V}_j^{-1}\right)^{-1} \\
\bar{\mathbf{y}}_j &= \nu_j^{-1}\sum_{i=1}^{n}\mathbf{y}_i \cdot \zeta_{ij} \\
\mathbf{V}_j &= \sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}}_j)'\zeta_{ij}/\nu_j(\mathbf{y}_i - \bar{\mathbf{y}}_j) \\
\nu_j &= \sum_{i=1}^{n}\zeta_{ij}
\end{aligned}$$

   i.e., $\bar{\mathbf{y}}_j$ is the mean and $\mathbf{V}_j$ is the maximum likelihood estimate of the covariance matrix of $\mathbf{y}$ for those observations where $\zeta_{ij} = 1$, respectively.

2. $p(\mathbf{\Sigma}_j|\mathbf{y}, \mu_j, \zeta) = p(\mathbf{\Sigma}_j)p(\mathbf{y}|\zeta, \mu_j, \mathbf{\Sigma}_j)$ is the product of an improper, uniform prior for $\mathbf{\Sigma}_j$ and the multivariate normal density over $\mathbf{y}$ in component j. The resulting density for $\mathbf{\Sigma}_j$ is an inverse-Wishart density, with scale matrix parameter

$$\mathbf{S}_j = \sum_{i=1}^{n}(\mathbf{y}_i - \mu_j)'\zeta_{ij}(\mathbf{y}_i - \mu_j)$$

   and degrees of freedom parameter $\nu_j = \sum_{i=1}^{n}\zeta_{ij}$.

3. $p(\zeta_{ij}|\mathbf{y}_i, \mu_1, \mu_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{x}_i, \beta)$. Recall that $\zeta_{ij} \in \{0, 1\}$, and so this conditional distribution is a probability mass function. Also, in our case, we consider only $J = 2$ components. Thus, let π_i be the conditional probability that $\zeta_{i1} = 1$. By the discrete version of Bayes

Rule, we have

$$\begin{aligned}\pi_i &= \frac{\Pr(\zeta_{ij}=1)\, p(\mathbf{y}_i|\zeta_{ij}=1,\mu_1,\mathbf{\Sigma}_1)}{\sum_{k=1}^{J}\Pr(\zeta_{ik}=1)\, p(\mathbf{y}_i|\zeta_{ik}=1,\mu_k,\mathbf{\Sigma}_k)} \\ &= \frac{F(\mathbf{x}_i\beta)\,\phi(\mathbf{y}_i;\mu_1,\mathbf{\Sigma}_1)}{F(\mathbf{x}_i\beta)\,\phi(\mathbf{y}_i;\mu_1,\mathbf{\Sigma}_1)\;+\;(1-F(\mathbf{x}_i\beta))\,\phi(\mathbf{y}_i;\mu_2,\mathbf{\Sigma}_2)}\end{aligned}$$

where the term $F(\mathbf{x}_i\beta)$ comes from the hierarchical part of our model in 1. We then sample ζ_{i1} as the outcome of a Bernoulli trial with probability π_i. In the case of $J = 2$ components we simply set $\zeta_{i2} = 1 - \zeta_{i1}$.

4. $p(\beta|\mathbf{X},\zeta) = p(\beta)p(\zeta|\mathbf{X},\beta)$. The second term is the logistic regression likelihood,

$$p(\zeta|\mathbf{X},\beta) = \prod_{i=1}^{n} F(\mathbf{x}_i\beta)^{\zeta_{i1}}[1-F(\mathbf{x}_i\beta)]^{\zeta_{i2}} \tag{4}$$

while $p(\beta) \equiv N(\mathbf{b}_0,\mathbf{B}_0)$ is the vague multivariate normal prior in 2f. In a large sample like ours with $n = 5,000$ observations, the resulting density is well approximated by a multivariate normal density centered on the precision-matrix weighted average of the MLEs from the logistic regression of ζ_1 on $\mathbf{X}$ and the prior mean $\mathbf{b}_0 = \mathbf{0}$: i.e., approximately,

$$\beta|\mathbf{X},\zeta \sim N(\tilde{\mathbf{b}},\tilde{\mathbf{B}})$$

where

$$\begin{aligned}\tilde{\mathbf{b}} &= (\hat{\mathbf{V}}_\beta^{-1}+\mathbf{B}_0^{-1})^{-1}(\hat{\mathbf{V}}_\beta^{-1}\hat{\beta}+\mathbf{B}_0^{-1}\mathbf{b}_0), \\ \tilde{\mathbf{B}} &= (\hat{\mathbf{V}}_\beta^{-1}+\mathbf{B}_0^{-1})^{-1}\end{aligned}$$

and where $\hat{\beta}$ maximizes the likelihood in equation 4 and $\hat{\mathbf{V}}_\beta$ is the inverse of the Fisher information for $\hat{\beta}$.

Sampling from each of these conditional distributions constitutes a single iteration of the Gibbs sampler.

In this data set, the two components of the mixture are quickly resolved by the sampler, and the Gibbs sampler appears to rapidly converge on the joint posterior density in equation 3. We run the sampler for five thousand iterations, discarding the first 10% of the run as burn-in from arbitrary initial values, although it does appear that the sampler has settled on the joint posterior density well before then.