

Bayesian Graduation Using Constrained Bernoulli-Mixture Models

William S. Cleveland and Chuanhai Liu*

Statistical Research Department, Bell Laboratories, Lucent Technologies

* the contact author for correspondence

Address: Room 2C-262

600 Mountain Avenue

Murray Hill, NJ 07974

U.S.A.

TEL: (908) 582-3986

FAX: (908) 582-3340

EMAIL: liu@research.bell-labs.com

Bayesian Graduation Using Constrained Bernoulli-Mixture Models

William S. Cleveland and Chuanhai Liu

Statistical Research Department, Bell Laboratories, Lucent Technologies

abstract

Broffitt (1988) showed that Bayesian graduation method is useful for statistical inference about the age-specific probability of death for people insured under a certain policy. It is commonly assumed that the age-specific probability of death, as a function of age, is increasing convex. However, it is realized that it can be important to model the populations of different risk levels. In this paper, we develop a class of Bernoulli-mixture models, where insured people come from two hidden populations; each population has its own age-specific probability of death that is increasing convex; and the age-specific ratio of the sample sizes of the two populations is monotone. It is shown that the EM algorithm (Dempster, Laird, and Rubin, 1977) and the DA algorithm (Tanner and Wong, 1987) can be used to obtain maximum likelihood estimation and Bayesian estimation, respectively. The numerical results show that modeling the underlying different risk groups leads to more reliable inference.

Key words and phrases: Data-augmentation algorithm; Expectation-maximization algorithm; Posterior-predictive checking.

1 Introduction

Broffitt (1988) demonstrated advantages of Bayesian graduation method that makes use of the prior knowledge on the age-specific probability of death, that is, the age-specific probability of death, as a function of age, is increasing convex. This interesting modeling strategy has stimulated some research interest in statistical modeling and computation. Carlin (1992) implemented the Gibbs sampler (Geman and Geman, 1984; and Gelfand and Smith, 1990) for posterior simulation of the posterior distribution. Gelman, Meng, and Stern (1996) provided an alternative computational method for the model and focused on model checking. Gelman (1996) investigated the sensitivity of the specifications of prior distribution for the parameters. Liu (2000) implemented the EM algorithm (Dempster, Laird, and Rubin, 1977) for maximum likelihood estimation and the Data Augmentation (DA) algorithm (Tanner and Wong, 1987) for Bayesian estimation.

Section 2 gives a brief review of the constrained binomial and Poisson models used in the literature for the mortality data: the age-specific probability of death, as a function of age, is increasing convex. As discussed by many authors, this constrained binomial and Poisson models are questionable for people at later ages. Using posterior-predictive checking (Rubin, 1984; Gelman, Meng, and Stern, 1996) via rescaled residual plots, we notice that the major lack-of-fit occurs for ages around 46 rather than for later ages.

Suggestions are that the insured people of same age have possibly different risk levels and that the ratio of the number of insured people of the same age from high-risk group to that from low-risk group increases as age increases. To investigate this, in Section 3 we consider modeling the hidden risk groups. We develop a class of Bernoulli-mixture models, where insured people come from two hidden populations; each population has its own age-specific probability of death that is increasing convex; and the age-specific ratio of the sample sizes of the two populations is monotone. Section 4 describes a data augmentation scheme for fitting the model using EM the EM algorithm and the DA algorithm. Sections 5 and 6 provide the EM algorithm and the DA algorithm for maximum likelihood estimation and Bayesian estimation, respectively. Section 7 presents the numerical results. Section 8 concludes with a few remarks.

2 The data and a brief review of the previous data analysis

The data in Table 1 were taken from Broffitt (1988) and considered by many others, *e.g.*, Carlin (1992) and Gelman et al. (1996). For each age, t_i ($i = 1, \dots, m = 30$), Table 1 gives N_i , the number of people insured under a certain policy, and n_i , the number of insured who died. People who joined or left the policy in the middle of the year are counted as half. For simplicity, we round the number N_i and ignore the fraction.

Following Gelman et al. (1996), we assume that the observed deaths at each age, n_i , follow independent binomial distributions, *i.e.*,

$$n_i | (\theta_i, N_i) \stackrel{\text{ind}}{\sim} \text{BINOMIAL}(N_i, \theta_i) \quad (i = 1, \dots, m = 30). \quad (1)$$

The objective is to estimate the probability of death at age t_i , θ_i , for all $i = 1, \dots, m$ under the assumption that the θ_i , as a function of age t , is increasing and convex over the observed range ($35 \leq t \leq 64$). Because N_i was in the hundreds or thousands and the rates were very low, Gelman et al. (1996) used the Poisson approximation for computational/mathematical convenience. For notational convenience, we call the increasing convex binomial model the Single Risk Group Model (SRGM).

The increasing convex constraints on θ can be written as

$$\theta_i = \sum_{j=0}^m C_{ij} \alpha_j,$$

where C is the $m \times (m + 1)$ matrix

$$C = (C_{ij}) = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & \frac{1}{m-1} & 0 & \dots & 0 \\ 0 & 1 & \frac{2}{m-1} & \frac{1}{m-2} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & \frac{m-1}{m-1} & \frac{m-2}{m-2} & \dots & 1 \end{bmatrix} \quad (2)$$

and

$$\alpha_j \geq 0, \quad \sum_{j=0}^m \alpha_j = 1. \quad (3)$$

Liu (2000) developed the EM algorithm for maximum likelihood estimation, and the DA algorithm for Bayesian estimation of the constrained binomial model. Liu (2000) also provided the EM algorithm and the DA algorithms for maximum likelihood estimation and Bayesian estimation of the Poisson model, respectively. As Gelman et al. (1996) suggested, the Poisson model is a good approximation to the binomial model.

The maximum likelihood estimates $\hat{\theta}_i$ of θ_i in the binomial model (1) is displayed in Figure 1 (a). For checking the goodness of fit, the standardized residuals

$$\frac{n_i/N_i - \hat{\theta}_i}{\sqrt{\hat{\theta}_i(1 - \hat{\theta}_i)/N_i}} \quad (i = 1, \dots, m) \quad (4)$$

are displayed in Figure 1 (b), which shows that the observation at age 46 appears to be an outlier.

Because of the difficulty in counting the number of free parameters (Gelman et al., 1996) and the adequacy of the large-sample results, the Bayesian approach is useful. Figure 1 (c) displays the Bayesian estimates (with the prior distribution $\text{DIRICHLET}(\alpha; \kappa)$, where $\kappa_0 = \kappa_1 = 1/2$ and $\kappa_j = (m - j + 1)/[m(m - 1)]$ for $j = 2, \dots, m$) in terms of traces of nine draws and the point-wise median from the posterior distribution, which, as suggested by a referee, is more useful than, for example, point-wise 95% bounds because the entire curve is of interest. Let n_i^* be a draw of the post-predicted death counts corresponding to n_i with N_i insured people, that is,

$$n_i^* | (\theta_i, N_i) \sim \text{BINOMIAL}(N_i, \theta_i),$$

then

$$E(n_i^*/N_i | Y_{\text{obs}}) = E(\theta_i | Y_{\text{obs}})$$

and

$$\text{Var}(n_i^*/N_i) = \text{Var}(\theta_i | Y_{\text{obs}}) + E(\theta_i(1 - \theta_i) | Y_{\text{obs}})/N_i.$$

Thus, the point-wise 95% posterior-predictive (HPD) intervals $(-1.96, 1.96)$ of

$$\frac{n_i^*/N_i - E(\theta_i | Y_{\text{obs}})}{\sqrt{\text{Var}(\theta_i | Y_{\text{obs}}) + E(\theta_i(1 - \theta_i) | Y_{\text{obs}})/N_i}}$$

for the standardized residuals

$$\frac{n_i/N_i - E(\theta_i|Y_{\text{obs}})}{\sqrt{\text{Var}(\theta_i|Y_{\text{obs}}) + E(\theta_i(1 - \theta_i)|Y_{\text{obs}})/N_i}} \quad (i = 1, \dots, m), \quad (5)$$

which are displayed in Figure 1 (d), provide a way of posterior-predictive checking (Rubin, 1984; Gelman et al., 1996). We see that the lack-of-fit happens around age $t = 46$ and possibly for later ages. Surprisingly, this has not been noticed in the previous analysis of the data in the literature, where the attention has been paid to the fit of the model to the data for the later ages.

3 A class of Bernoulli-mixture models with increasing and increasing convex constraints

It is well known that the so-called over-dispersion problem can happen for binomial and Poisson models (*e.g.*, Efron, 1978, 1986; Lambert and Roeder, 1995). The lack-of-fit of SRGM to the data and the discussion of both Gelman et al. (1996) and Broffitt (1988) suggest that there exist possibly systematic change of the proportions of different risk groups over the observed range of ages. Gelman et al. (1996, p. 750) noted that “*Even if the assumption of convex mortality rate is true in the natural population, it is very likely that the insurance company has screened out some high-risk older people, and thus destroy the convexity for later ages*”. Taking this into account, we extend the binomial model by partitioning the underlying sampling population into two latent sub-populations/groups: one with high-risk of death and the other with low-risk. Based on the prior knowledge on mortality rates (Broffitt, 1988, p. 115): “*It is well accepted that, for ages 30 and above, human mortality rates increase with age*”, we assume that for each of the two groups, the mortality rate curve is increasing convex.

Since the group affiliation of individuals is unknown, we need to model the group affiliation parameter ω_t . Let $N_{l,t}$ and $N_{h,t}$ be the (unobservable) numbers of insured people from the low-risk and high-risk populations at age t , respectively. Then $N_t = N_{l,t} + N_{h,t}$. Suppose

that $N_{l,t}$ and $N_{h,t}$ follow independent Poisson distributions. Then

$$N_{l,t}|(N_t, \omega_t) \sim \text{BINOMIAL}(N_t, \omega_t),$$

where ω_t ($0 \leq \omega_t \leq 1$) is the fraction of the insured low-risk people. We assume that the fractions ω_t , as a function of the age t , is increasing, meaning that the older a person in a high-risk group the more likely the person to be “screened” out. Nevertheless, ω_t can be modeled when information on ω_t is available.

Let $\theta_{l,t}$ and $\theta_{h,t}$ be the mortality rates of the low-risk and high-risk populations, respectively. The death probability for each insured person at age t is then

$$\theta_t = \omega_t \theta_{l,t} + (1 - \omega_t) \theta_{h,t},$$

that is, the observed death count n_t for insured people at age t given N_t is the sum of the N_t independent Bernoulli-mixture variables, and hence follows the distribution

$$n_t|N_t \stackrel{\text{ind}}{\sim} \text{BINOMIAL}(N_t, \omega_t \theta_{l,t} + (1 - \omega_t) \theta_{h,t}). \quad (6)$$

In summary, we have the model in (6) for the mortality data, where as a function of age t over the observed range, ω_t is increasing and both the low-risk death rate $\theta_{l,t}$ and the high-risk death rate $\theta_{h,t}$ are increasing and convex.

4 A complete-data model

The constraints on the low-risk fraction ω_t of insured people and the death rates in both low-risk and high-risk populations make it difficult to estimate these parameters from the observed data. Here, we augment data to create the complete-data. For notational convenience, we use $i = 1$ to m to index the observed age range.

The increasing constraints on the probabilities ω_i 's are that both $\gamma_1 = \omega_1$ and the first-order differences $\gamma_i = \omega_i - \omega_{i-1}$ ($i = 2, \dots, m$) are positive and that $\omega_m = \gamma_1 + \dots + \gamma_m < 1$. Let $\gamma_0 = 1 - \sum_{i=1}^m \gamma_i$ and write the equalities

$$\omega_1 = \gamma_1, \quad \omega_2 = \gamma_1 + \gamma_2, \quad \dots, \quad \omega_m = \gamma_1 + \gamma_2 + \dots + \gamma_m$$

in a slightly more general form

$$\omega_i = \sum_{j=0}^m M_{i,j} \gamma_j, \quad (7)$$

where $0 \leq M_{i,j} \leq 1$ for $i = 1$ to m and $j = 0$ to m , $\gamma_j > 0$ for $j = 0$ to m , and $\sum_{j=0}^m \gamma_j = 1$. Let y_{ij} be the binary indicator that insured person j belongs to the low-risk population. Then Equation (7) suggests immediately the following simple mixture model for y_{ij} with essentially no constraints, except for the trivial conditions $\gamma_j > 0$ for all $j = 0, \dots, m$ and $\sum_{j=1}^m \gamma_j = 1$. For $i = 1$ to m and $j = 1$ to N_i ,

$$x_{ij} \equiv (x_{ij}(0), \dots, x_{ij}(m))' \sim \text{MULTINOMIAL}_{m+1}(1, (\gamma_0, \dots, \gamma_m)),$$

and

$$y_{ij} | (x_{ij}(k) = 1) \sim \text{BERNOULLI}(M_{ik}),$$

where $x_{ij} = (x_{ij}(0), \dots, x_{ij}(m))$ is a vector of $(m+1)$ binary indicators with $\sum_{k=0}^m x_{ij}(k) = 1$.

As given in (2) and (3), the death probabilities $\theta_{l,i}$ with increasing convex constraints can also be written in the form of (7), that is,

$$\theta_{l,i} = \sum_{j=0}^m L_{i,j} \alpha_j \quad (i = 1, \dots, m), \quad (8)$$

where $0 \leq L_{i,j} \leq 1$ for $i = 1$ to m and $j = 0$ to m , $\alpha_j > 0$ for $j = 0$ to m , and $\sum_{j=0}^m \alpha_j = 1$. To be more specific, $L_{i,j}$ is given as

$$L_{i,j} = C_{ij} = \begin{cases} 0, & \text{if } j = 0 \text{ or } j > i; \\ 1, & \text{if } j = 1; \\ \frac{i-j+1}{m-j+1}, & \text{otherwise.} \end{cases}$$

Accordingly, we augment data for the death indicator n_{ij} of person j when in low-risk group (*i.e.*, $y_{ij} = 1$) by taking

$$z_{ij} | (y_{ij} = 1) \sim \text{MULTINOMIAL}_{m+1}(1, (\alpha_0, \dots, \alpha_m)),$$

and

$$n_{ij} | (y_{ij} = 1, z_{ij}(k) = 1) \sim \text{BERNOULLI}(L_{ik})$$

where $z_{ij} = (z_{ij}(0), \dots, z_{ij}(m))$ is a vector of $(m + 1)$ binary indicators with $\sum_{k=0}^m z_{ij}(k) = 1$. Similarly, we have for the high-risk group

$$\theta_{h,i} = \sum_{j=0}^m H_{i,j} \beta_j \quad (i = 1, \dots, m), \quad (9)$$

where $0 \leq H_{i,j} = L_{i,j} \leq 1$ for $i = 1$ to m and $j = 0$ to m , $\beta_j > 0$ for $j = 0$ to m , and $\sum_{j=0}^m \beta_j = 1$. Similarly, we augment data for the death indicator n_{ij} of person j when in high-risk group (*i.e.*, $y_{ij} = 0$) by letting

$$z_{ij}|(y_{ij} = 0) \sim \text{MULTINOMIAL}_{m+1}(1, (\beta_0, \dots, \beta_m))$$

and

$$n_{ij}|(y_{ij} = 0, z_{ij}(k) = 1) \sim \text{BERNOULLI}(H_{ik}).$$

This completes our complete data

$$Y_{\text{com}} = \{(x_{ij}, y_{ij}, z_{ij}, n_{ij}) : i = 1, \dots, m; j = 1, \dots, N_i\}. \quad (10)$$

It is easy to show that the complete-data model preserves the observed-data model. First, we have for all $i = 1, \dots, m$ and $j = 1, \dots, N_i$

$$\begin{aligned} y_{ij}|\theta &\sim \text{BERNOULLI}(\omega_i), \\ n_{ij}|(y_{ij} = 1) &\sim \text{BERNOULLI}(\theta_{l,i}), \\ n_{ij}|(y_{ij} = 0) &\sim \text{BERNOULLI}(\theta_{h,i}), \\ n_{ij}|\theta &\sim \text{BERNOULLI}(\omega_i \theta_{l,i} + (1 - \omega_i) \theta_{h,i}). \end{aligned}$$

Let

$$n_i = \sum_{j=1}^{N_i} n_{ij}$$

be the number of 1s in the N_i Bernoulli-mixture variables n_{ij} for $i = 1, \dots, m$. Then we have for all $i = 1, \dots, m$

$$n_i \stackrel{\text{ind}}{\sim} \text{BINOMIAL}(N_i, \omega_i \theta_{l,i} + (1 - \omega_i) \theta_{h,i}).$$

Therefore, the complete-data model for (10) preserves the the observed-data model.

5 Maximum likelihood estimation using the EM algorithm

The complete-data model consists of three multinomial-binomial hierarchical models, which involve the parameters γ 's, α 's, and β 's, respectively. The multinomial-binomial hierarchical model belongs to the exponential family. The complete-data sufficient statistics for γ 's, and thereby for ω 's, are

$$S_x(0) = \sum_{j=1}^{N_i} \sum_{i=1}^m x_{ij}(0), \quad \dots, \quad S_x(m) = \sum_{j=1}^{N_i} \sum_{i=1}^m x_{ij}(m), \quad \text{and} \quad N = \sum_{i=1}^m N_i;$$

and the complete-data maximum likelihood estimates of γ 's are

$$\hat{\gamma}_0 = \frac{S_x(0)}{N}, \quad \dots, \quad \text{and} \quad \hat{\gamma}_m = \frac{S_x(m)}{N}.$$

The complete-data sufficient statistics for α 's are

$$S_{yz}(0) = \sum_{j=1}^{N_i} \sum_{i=1}^m y_{ij} z_{ij}(0), \quad \dots, \quad S_{yz}(m) = \sum_{j=1}^{N_i} \sum_{i=1}^m y_{ij} z_{ij}(m), \quad \text{and} \quad S_y = \sum_{j=1}^{N_i} \sum_{i=1}^m y_{ij};$$

and the complete-data maximum likelihood estimates of α 's are

$$\hat{\alpha}_0 = \frac{S_{yz}(0)}{S_y}, \quad \dots, \quad \text{and} \quad \hat{\alpha}_m = \frac{S_{yz}(m)}{S_y}.$$

The complete-data sufficient statistics for β 's are

$$S_{(1-y)z}(k) = \sum_{j=1}^{N_i} \sum_{i=1}^m (1 - y_{ij}) z_{ij}(k) \quad (k = 0, \dots, m) \quad \text{and} \quad S_{1-y} = \sum_{j=1}^{N_i} \sum_{i=1}^m (1 - y_{ij});$$

and the complete-data maximum likelihood estimates of β 's are

$$\hat{\beta}_k = \frac{S_{(1-y)z}(k)}{S_{1-y}} \quad (k = 0, \dots, m).$$

The EM algorithm for the exponential family is straightforward. The E-step of EM computes the conditional expectations of the complete-data sufficient statistics, given the observed data and the current estimates of the parameters. The M-step updates the estimates of the parameters by replacing the complete-data sufficient statistics in the complete-data maximum likelihood estimators with their conditional expectations obtained in the E-step.

For each i , there are N_i insured people, of which n_i died. In other words, n_i death indicators n_{ij} take the value of 1 and the other n_{ij} 's take value of 0. Given the current estimates of the parameters, from the Bayes theorem we have

$$y_{ij}|(n_{ij} = 1) \sim \text{BERNOULLI} \left(\frac{\omega_i \theta_{l,i}}{\theta_i} \right) \quad (11)$$

and

$$y_{ij}|(n_{ij} = 0) \sim \text{BERNOULLI} \left(\frac{\omega_i(1 - \theta_{l,i})}{1 - \theta_i} \right), \quad (12)$$

where $\theta_i = \omega_i \theta_{l,i} + (1 - \omega_i) \theta_{h,i}$. Applying the Bayes theorem, we also get the following conditional distributions, which are useful for calculating the conditional expectations of the complete-data sufficient statistics,

$$x_{ij}|(y_{ij} = 1, n_{ij}) \sim \text{MULTINOMIAL} \left(1, \left(\frac{M_{i,0} \gamma_0}{\omega_i}, \dots, \frac{M_{i,m} \gamma_m}{\omega_i} \right) \right), \quad (13)$$

$$x_{ij}|(y_{ij} = 0, n_{ij}) \sim \text{MULTINOMIAL} \left(1, \left(\frac{(1 - M_{i,0}) \gamma_0}{1 - \omega_i}, \dots, \frac{(1 - M_{i,m}) \gamma_m}{1 - \omega_i} \right) \right), \quad (14)$$

$$z_{ij}|(y_{ij} = 1, n_{ij} = 1) \sim \text{MULTINOMIAL} \left(1, \left(\frac{L_{i,0} \alpha_0}{\theta_{l,i}}, \dots, \frac{L_{i,m} \alpha_m}{\theta_{l,i}} \right) \right), \quad (15)$$

$$z_{ij}|(y_{ij} = 1, n_{ij} = 0) \sim \text{MULTINOMIAL} \left(1, \left(\frac{(1 - L_{i,0}) \alpha_0}{1 - \theta_{l,i}}, \dots, \frac{(1 - L_{i,m}) \alpha_m}{1 - \theta_{l,i}} \right) \right), \quad (16)$$

$$z_{ij}|(y_{ij} = 0, n_{ij} = 1) \sim \text{MULTINOMIAL} \left(1, \left(\frac{H_{i,0} \beta_0}{\theta_{h,i}}, \dots, \frac{H_{i,m} \beta_m}{\theta_{h,i}} \right) \right), \quad (17)$$

$$z_{ij}|(y_{ij} = 0, n_{ij} = 0) \sim \text{MULTINOMIAL} \left(1, \left(\frac{(1 - H_{i,0}) \beta_0}{1 - \theta_{h,i}}, \dots, \frac{(1 - H_{i,m}) \beta_m}{1 - \theta_{h,i}} \right) \right). \quad (18)$$

For the complete-data sufficient statistics for α 's, we have

$$\begin{aligned} \mathbb{E}(S_y | Y_{\text{Obs}}, \alpha, \beta, \gamma) &= \sum_{i=1}^m [n_i \mathbb{E}(y_{i1} | n_{i1} = 1) + (N_i - n_i) \mathbb{E}(y_{i1} | n_{i1} = 0)] \\ &\stackrel{(11,12)}{=} \sum_{i=1}^m \left[\frac{n_i \omega_i \theta_{l,i}}{\theta_i} + \frac{(N_i - n_i) \omega_i (1 - \theta_{l,i})}{1 - \theta_i} \right] \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}(S_{yz}(k) | Y_{\text{Obs}}, \alpha, \beta, \gamma) &= \sum_{i=1}^m [n_i \mathbb{E}(y_{i1} z_{i1}(k) | n_{i1} = 1) + (N_i - n_i) \mathbb{E}(y_{i1} z_{i1}(k) | n_{i1} = 0)] \\ &= \sum_{i=1}^m [n_i \text{Prob}(y_{i1} = 1 | n_{i1} = 1) \mathbb{E}(z_{i1}(k) | n_{i1} = 1, y_{i1} = 1) \end{aligned}$$

$$\stackrel{(15,16)}{=} \alpha_k \sum_{i=1}^m \left[\frac{n_i \omega_i L_{i,k}}{\theta_i} + \frac{(N_i - n_i) \omega_i (1 - L_{i,k})}{1 - \theta_i} \right] + (N_i - n_i) \text{Prob}(y_{i1} = 1 | n_{i1} = 0) E(z_{i1}(k) | n_{i1} = 0, y_{i1} = 1)$$

for $k = 0, \dots, m$. Similarly, we have for the complete-data sufficient statistics for β 's

$$E(S_{1-y} | Y_{\text{Obs}}, \alpha, \beta, \gamma) = \sum_{i=1}^m \left[\frac{n_i (1 - \omega_i) \theta_{h,i}}{\theta_i} + \frac{(N_i - n_i) (1 - \omega_i) (1 - \theta_{h,i})}{1 - \theta_i} \right]$$

and

$$E(S_{(1-y)z}(k) | Y_{\text{Obs}}, \alpha, \beta, \gamma) = \beta_k \sum_{i=1}^m \left[\frac{n_i (1 - \omega_i) H_{i,k}}{\theta_i} + \frac{(N_i - n_i) (1 - \omega_i) (1 - H_{i,k})}{1 - \theta_i} \right].$$

For the complete-data sufficient statistics for γ 's, we have

$$\begin{aligned} & E(S_x(k) | Y_{\text{Obs}}, \alpha, \beta, \gamma) \\ = & \sum_{i=1}^m [n_i E(x_{i1}(k) | n_{i1} = 1) + (N_i - n_i) E(x_{i1}(k) | n_{i1} = 0)] \\ = & \sum_{i=1}^m \left\{ n_i [\text{Prob}(y_{i1} = 1 | n_{i1} = 1) E(x_{i1}(k) | n_{i1} = 1, y_{i1} = 1) \right. \\ & \quad \left. + \text{Prob}(y_{i1} = 0 | n_{i1} = 1) E(x_{i1}(k) | n_{i1} = 1, y_{i1} = 0)] \right. \\ & \quad \left. + (N_i - n_i) [\text{Prob}(y_{i1} = 1 | n_{i1} = 0) E(x_{i1}(k) | n_{i1} = 0, y_{i1} = 1) \right. \\ & \quad \left. + \text{Prob}(y_{i1} = 0 | n_{i1} = 0) E(x_{i1}(k) | n_{i1} = 0, y_{i1} = 0)] \right\} \\ \stackrel{(11,12)}{=} & \sum_{i=1}^m \left\{ n_i \left[\frac{\omega_i \theta_{l,i}}{\theta_i} E(x_{i1}(k) | n_{i1} = 1, y_{i1} = 1) \right. \right. \\ & \quad \left. \left. + \left(1 - \frac{\omega_i \theta_{l,i}}{\theta_i}\right) E(x_{i1}(k) | n_{i1} = 1, y_{i1} = 0) \right] \right. \\ & \quad \left. + (N_i - n_i) \left[\frac{\omega_i (1 - \theta_{l,i})}{1 - \theta_i} E(x_{i1}(k) | n_{i1} = 0, y_{i1} = 1) \right. \right. \\ & \quad \left. \left. + \left(1 - \frac{\omega_i (1 - \theta_{l,i})}{1 - \theta_i}\right) E(x_{i1}(k) | n_{i1} = 0, y_{i1} = 0) \right] \right\} \\ \stackrel{(13,14)}{=} & \sum_{i=1}^m \left\{ n_i \left[\frac{\omega_i \theta_{l,i}}{\theta_i} \frac{M_{i,k} \gamma_k}{\omega_i} + \left(1 - \frac{\omega_i \theta_{l,i}}{\theta_i}\right) \frac{(1 - M_{i,k}) \gamma_k}{1 - \omega_i} \right] \right. \\ & \quad \left. + (N_i - n_i) \left[\frac{\omega_i (1 - \theta_{l,i})}{(1 - \theta_i)} \frac{M_{i,k} \gamma_k}{\omega_i} + \left(1 - \frac{\omega_i (1 - \theta_{l,i})}{(1 - \theta_i)}\right) \frac{(1 - M_{i,k}) \gamma_k}{1 - \omega_i} \right] \right\} \\ = & \gamma_k \sum_{i=1}^m \left[\frac{n_i (\theta_{l,i} M_{ik} + \theta_{h,i} (1 - M_{ik}))}{\theta_i} + \frac{(N_i - n_i) ((1 - \theta_{l,i}) M_{ik} + (1 - \theta_{h,i}) (1 - M_{ik}))}{1 - \theta_i} \right]. \end{aligned}$$

These expectations provide the needed expressions for the E-step of the EM algorithm. The M-step of the EM algorithm is simply obtained by replacing the complete-data sufficient statistics with their conditional expectations given the observed data and the current estimates of the parameters. Denote by θ the vector of the unknown parameters, then the E-step and M-step of the t -th iteration of the EM algorithm can be summarized as follows.

The EM algorithm:

E-step: *Compute the conditional expected sufficient statistics given the observed data and the current estimate $\theta^{(t-1)}$:*

$$\begin{aligned} S_x^{(t)}(k) &= E(S_x(k)|Y_{obs}, \theta^{(t-1)}), \\ S_{yz}^{(t)} &= E(S_{yz}(k)|Y_{obs}, \theta^{(t-1)}), \\ S_y^{(t)} &= E(S_y|Y_{obs}, \theta^{(t-1)}), \\ S_{(1-y)z}^{(t)} &= E(S_{(1-y)z}(k)|Y_{obs}, \theta^{(t-1)}), \\ S_{1-y}^{(t)} &= E(S_{1-y}|Y_{obs}, \theta^{(t-1)}). \end{aligned}$$

M-step: *Update the estimates of the parameters:*

$$\gamma_k^{(t)} = \frac{S_x^{(t)}(k)}{\sum_{i=1}^m N_i}, \quad \alpha_k^{(t)} = \frac{S_{yz}^{(t)}(k)}{S_y^{(t)}}, \quad \text{and} \quad \beta_k^{(t)} = \frac{S_{(1-y)z}^{(t)}(k)}{S_{1-y}^{(t)}} \quad (k = 0, \dots, m).$$

6 Bayesian estimation using the DA algorithm

For Bayesian estimation, we use the conjugate prior distribution of the form

$$\pi(\alpha, \beta, \gamma) = \pi(\alpha)\pi(\beta)\pi(\gamma) = \text{DIRICHLET}(\alpha; \kappa_\alpha) \times \text{DIRICHLET}(\beta; \kappa_\beta) \times \text{DIRICHLET}(\gamma; \kappa_\gamma), \quad (19)$$

where the three κ 's are vectors of constants. If κ 's are vectors of 0.5, the prior is non-informative with respect to the associated complete data from multinomial distributions (Box and Tiao, 1973).

For the Poisson approximation to the single risk group model (1) with the death rates increasing convex (2) and (3), Gelman (1996) showed that the model with the flat prior

distribution for the death rates has a tendency to fit quadratic curves and that when m increases, *i.e.*, the scale of the time intervals becomes smaller and smaller, the posterior with the flat prior would eventually ignore the data. For more discussion on the related issues, see, for example, Berger and Bernardo (1992). For the use of the Dirichlet mixtures for the prior distribution, see, for example, Good (1967, 1976).

For simplicity, we use the prior (19) with $\kappa_\gamma(j) = 1/2$, $\kappa_\alpha(j) = 1/2$, and $\kappa_\beta(j) = 1/2$ for all $j = 0, \dots, m$. The derivation of the EM algorithm provides the posterior-predictive distributions for the Imputation (I) step of the DA algorithm. Given the imputed complete data Y_{com} in (10), we have

$$\begin{aligned} \pi(\alpha, \beta, \gamma | Y_{\text{com}}) &= \text{DIRICHLET}(\alpha; \kappa_\alpha + \sum_{i=1}^m \sum_{j=1}^{N_i} y_{ij} z_{ij}) \\ &\quad \times \text{DIRICHLET}(\beta; \kappa_\beta + \sum_{i=1}^m \sum_{j=1}^{N_i} (1 - y_{ij}) z_{ij}) \\ &\quad \times \text{DIRICHLET}(\gamma; \kappa_\gamma + \sum_{i=1}^m \sum_{j=1}^{N_i} x_{ij}), \end{aligned} \tag{20}$$

which leads to a simple Posterior (P) step of the DA algorithm. To summarize, we have the following DA algorithm.

The DA algorithm:

I-step: *Given the current draw of (α, β, γ) and observed data, draw missing components of Y_{com} from the corresponding conditional distribution, which is given in Section 5.*

P-step: *Given the current draw of Y_{com} , draw (α, β, γ) from the complete-data posterior distribution (20).*

7 Reanalyzing the data

We consider the use of the constrained binomial models with two risk groups, that is, the mixture increasing convex Bernoulli models for analyzing the data in Table 1. To distinguish this model from the Single-Risk Group Model discussed in Section 2, we call this model the Two-Risk Group Model (TRGM).

The maximum likelihood estimates of the mortality rates of the mixed, low-risk, and high-risk groups are displayed in Figures 2 (a), (b), and (c), respectively. Figure 2 (d) displays the maximum likelihood estimates of the weights ω_t , the probability that the an insured person at age t belongs to the low-risk group. The corresponding results of the Bayesian estimation are displayed in Figure 3: (a) are the observed death rates (in dot), posterior median (in solid line), and five posterior draws of the mortality rate of the mixed group; and (b), (c), and (d) are similar to the display of (a), but for the mortality rates of the two different risk groups and the mixing probabilities. We see that the point-wise posterior medians provides a mortality rate curve that is smoother than the maximum likelihood estimates. We note that since the curve ω_t does not appear to be over-forced to a straight line, the resulted posterior of the ω_t from the use of the Dirichlet prior with $\kappa_\gamma(0) = \dots = \kappa_\gamma(m) = 1/2$ does ignore the data.

Figure 4 provides residual plots for model checking. Figure 4 (b) displays the standardized residuals (4) from the maximum likelihood estimation. From Figure 4 (b), we see that the model appears to over-fit the data in the sense of underestimating the uncertainty. Nevertheless, it is interesting to see that with the parameters that are three times as many as the sufficient statistics $\{n_i/N_i : i = 1, \dots, m\}$, the smoothing has certainly taken place due to the monotonicity and convexity constraints. Although Figure 4 (b) shows that TRGM appears to be adequate for the data, it is difficult to capture the uncertainty about the constrained parameters. Thus, the Bayesian estimation is more appropriate. Figure 4 (d) displays the standardized residuals (5) from the Bayesian estimation. Comparing Figure 4 (d) to Figure 1 (d), we see that TRGM is better than SRGM.

Figure 5 displays the Bayesian estimates of the first-order differences, $\omega_t - \omega_{t-1}$, of the sampling proportions ω_t of the low-risk groups. Although the associated uncertainties are large, it can be seen that the large drops of the proportions of high-risk groups occurred around ages from 45 to 50. From Figure 3 (d), we see that the estimated low-risk group is mainly represented by the insured people at later ages. However, as we shall discuss in the next section, care must be taken in giving interpretations of the results on the parameters for the latent variables — the risk-group indicators for insured people.

8 Discussion

It is noticed that there exists the lack-of-fit of the SRGM for the mortality rate data (Broffitt, 1988). An improved model, *i.e.*, TRGM, is built by taking it into account the possibility of different sampling proportions from different risk groups. This class of constrained Bernoulli-mixture models can be also useful by themselves. It is also straightforward to extend this class of models to handle more than two risk groups.

The formulation of TRGM is based on the sounded underlying scientific considerations via both data augmentation and parameter expansion. This methodology is not new. The well known *factor analysis* provides such an example. As with factor analysis, there can be a (practical) nonidentifiability problem, and thus care must be taken in interpreting the results obtained from the observed data. With the monotonicity and convexity constraints, the non-identifiability in TRGM is not as obvious as that in factor analysis model. For example, a shift of a proportion from the high risk-group to the low-risk group without destroying the imposed monotonicity for the sampling proportions can not be identified from the observed data. Again, with *confirmative* factor analysis, *confirmative* versions of TRGM can be considered to provide *conditional* and thereby sharper inferences given additional prior knowledge.

We provided the EM algorithm and the DA algorithm for maximum likelihood estimation and Bayesian estimation, respectively. Although the algorithms are simple to implement, they can converge very slowly, especially when the hyper-parameters κ_α , κ_β , and κ_γ in the prior distribution (19) are fixed at small values. To accelerate the EM algorithm, the PX-EM algorithm (Liu, Rubin, and Wu, 1998) can be considered. Stochastic versions of PX-EM (*e.g.*, Meng and van Dyk, 1999; Liu and Wu, 1999; and Liu, 2003) can also be used to accelerate the DA algorithm for Bayesian estimation.

References

- Berger, J. O. and Bernardo, J. M. (1992). Ordered group reference priors with application to the multinomial problem, *Biometrika*, 79, 25-37.
- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, New York: John Wiley.
- Broffitt, J. D. (1988). Increasing and increasing convex Bayesian graduation, *Transactions of the Society of Actuaries*, 40, 115-148.
- Carlin, B. P. (1992). A simple Monte Carlo approach to Bayesian graduation, *Transactions of the Society of Actuaries*, 44, 55-76.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 39, 1-38.
- Efron, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation. *Journal of the American Statistical Association*, 73, 113-121.
- Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81, 709-721.
- Gelman, A. (1996). Bayesian model-building by pure thought: some principles and examples. *Statistica Sinica*, 6, 215-232.
- Gelman, A., Meng, X-L, and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion), *Statistica Sinica*, 6, 733-807.
- Good, I. J. (1967). A Bayesian significance test for multinomial distributions (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 29, 399-431.
- Good, I. J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, *The Annals of Statistics*, 4, 1159-1189.

- Lambert, D. and Roeder, K. (1995). Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association*, 90, 1225-1236
- Liu, C. (2000). Estimation of discrete distributions with a class of simplex constraints and its applications, *Journal of the American Statistical Association*, 95, 109-120.
- Liu, C. (2003). Alternating subspace-spanning resampling to accelerate Markov Chain Monte Carlo simulation, *Journal of the American Statistical Association*, 98, 110-117.
- Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm, *Biometrika*, 85, 755-770.
- Liu, J. S. and Wu, Y. (1999). Parameter expansion for data augmentation, *Journal of the American Statistical Association*, 94, 1264-1274.
- Meng, X. L., and van Dyk, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation, *Biometrika*, 86, 301-320.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151-1172.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association*, 82, 528-550.

i	t_i	N_i	n_i	i	t_i	N_i	n_i
1	35	1171.5	3	16	50	1516.0	4
2	36	2126.5	1	17	51	1371.5	7
3	37	2743.5	3	18	52	1343.0	4
4	38	2766.0	2	19	53	1304.0	4
5	39	2463.0	2	20	54	1232.5	11
6	40	2368.0	4	21	55	1204.5	11
7	41	2310.0	4	22	56	1113.5	13
8	42	2306.5	7	23	57	1048.0	12
9	43	2059.5	5	24	58	1155.0	12
10	44	1917.0	2	25	59	1018.5	19
11	45	1931.0	8	26	60	945.0	12
12	46	1746.5	13	27	61	853.0	16
13	47	1580.0	8	28	62	750.0	12
14	48	1580.0	2	29	63	693.0	6
15	49	1467.5	7	30	64	594.0	10

Table 1: Mortality rate data from Broffitt (1988), where i — index, t_i — age, N_i — number of insured, n_i — number of insured who died.

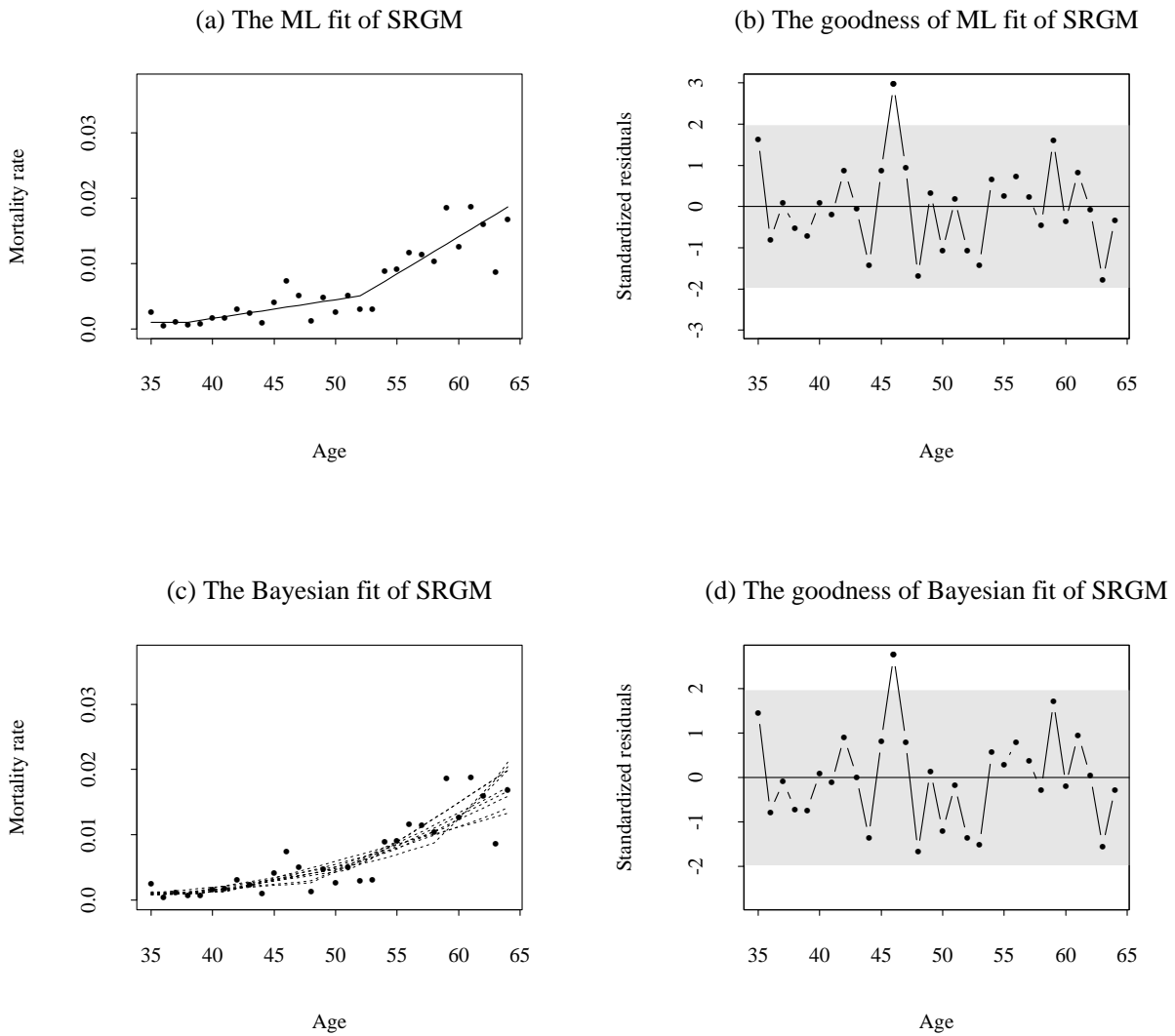


Figure 1: The maximum likelihood estimate of the increasing convex binomial model with a single risk group for the mortality rate data: (a) are the observed death rates (in dots) and the maximum likelihood (ML) estimates of the mortality rate curve (in line); (b) the standardized residuals from the ML fit with the grey area indicating the point-wise 95% confidence intervals; (c) nine posterior draws; and (d) the standardized residuals and the point-wise 95% posterior-predictive intervals.

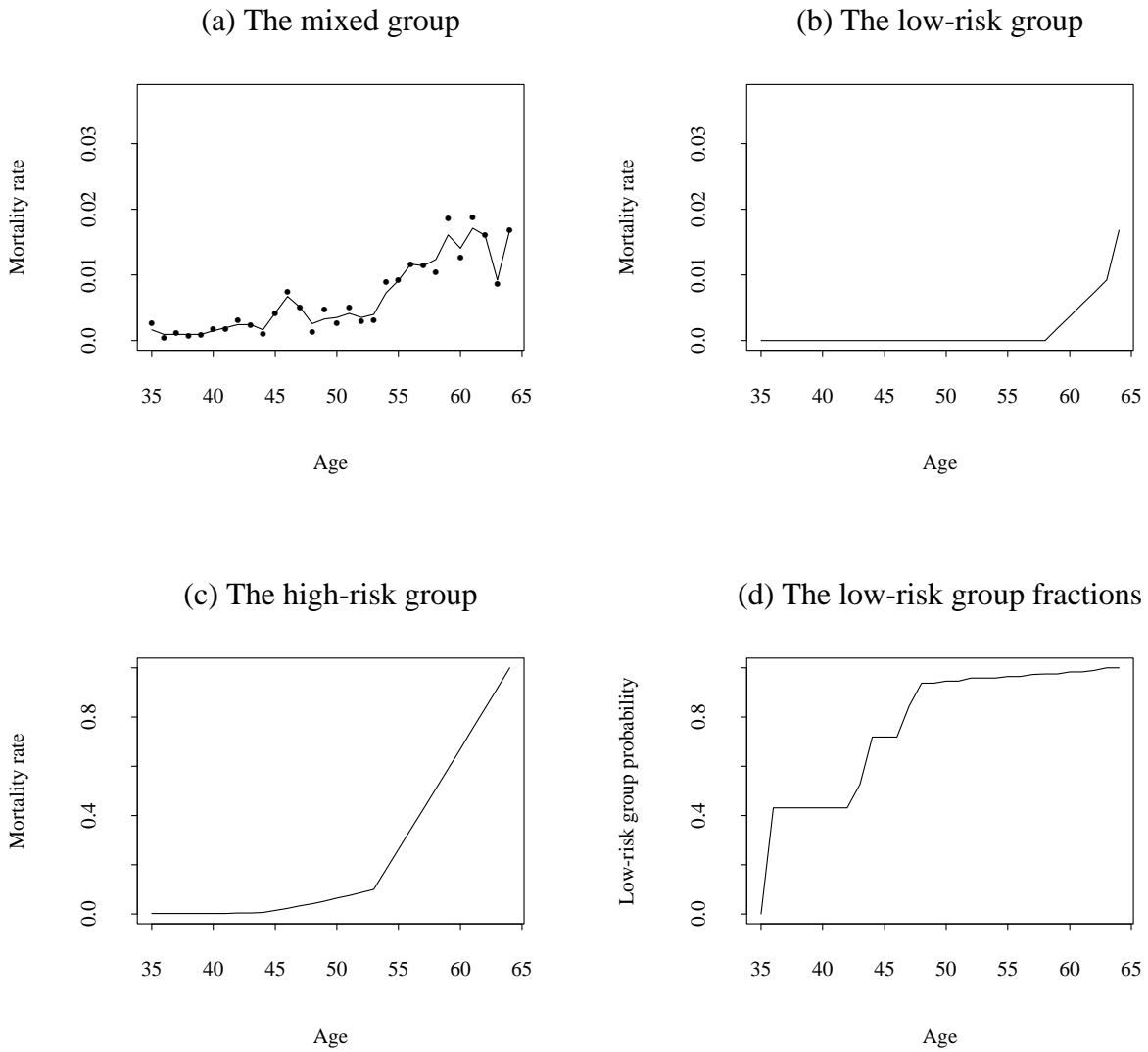


Figure 2: The maximum likelihood estimate of the binomial mixture model for the mortality rate data: (a) are the observed death rates (in dots) and the maximum likelihood estimates of the mortality rate curve (in line); (b) and (c) are the maximum likelihood estimates of the mortality rate curves for the low-risk and high-risk groups, respectively; and (d) the estimated fractions of the low-risk group.

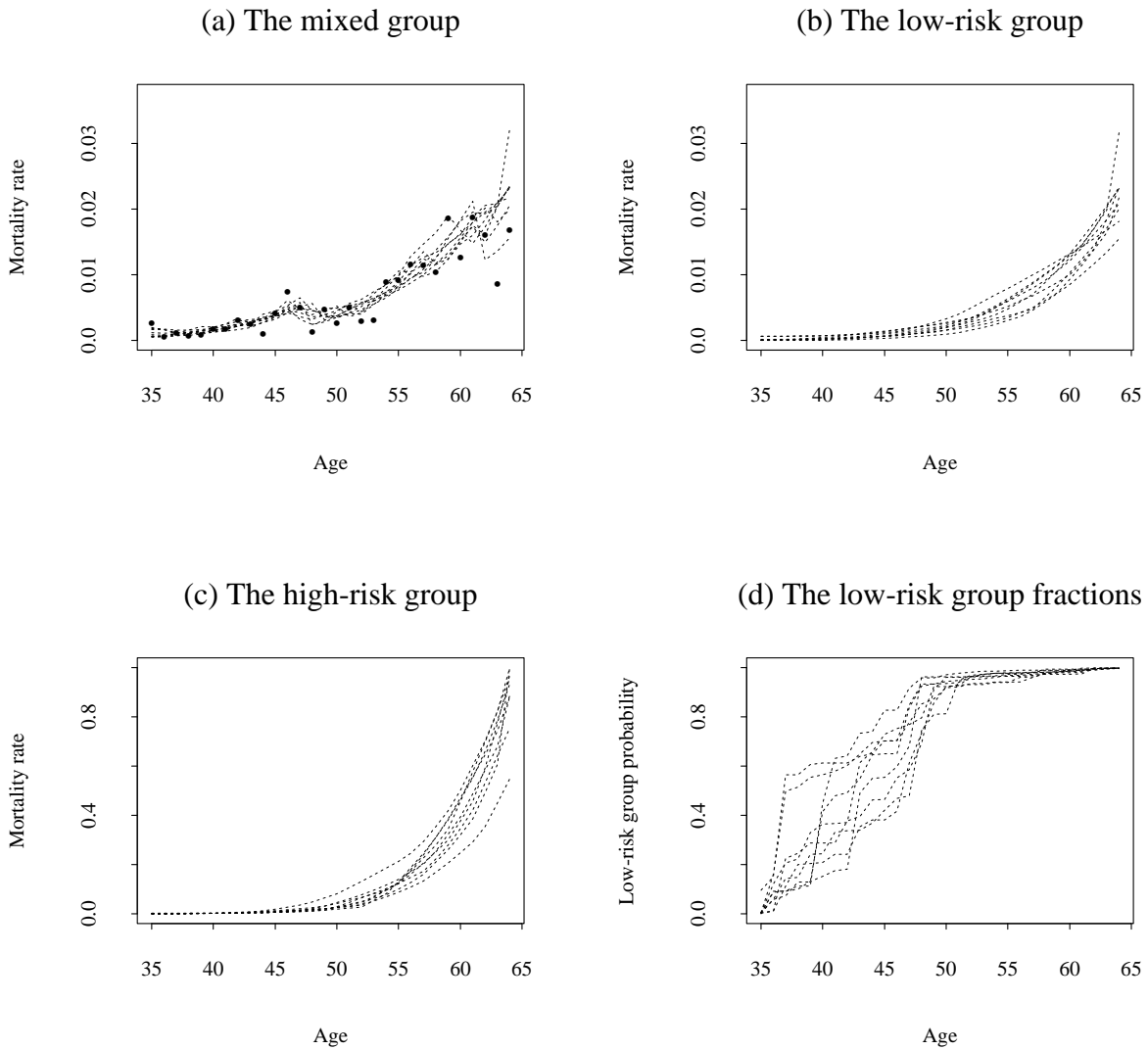


Figure 3: The Bayesian estimate of the binomial mixture model for the mortality rate data: (a) are the observed death rates (in dot) and nine posterior draws of the mortality rate of the mixed group; and (b), (c), and (d) are similar to the display of (a), but for the mortality rates of the two different risk groups and the fractions of the low-risk group.

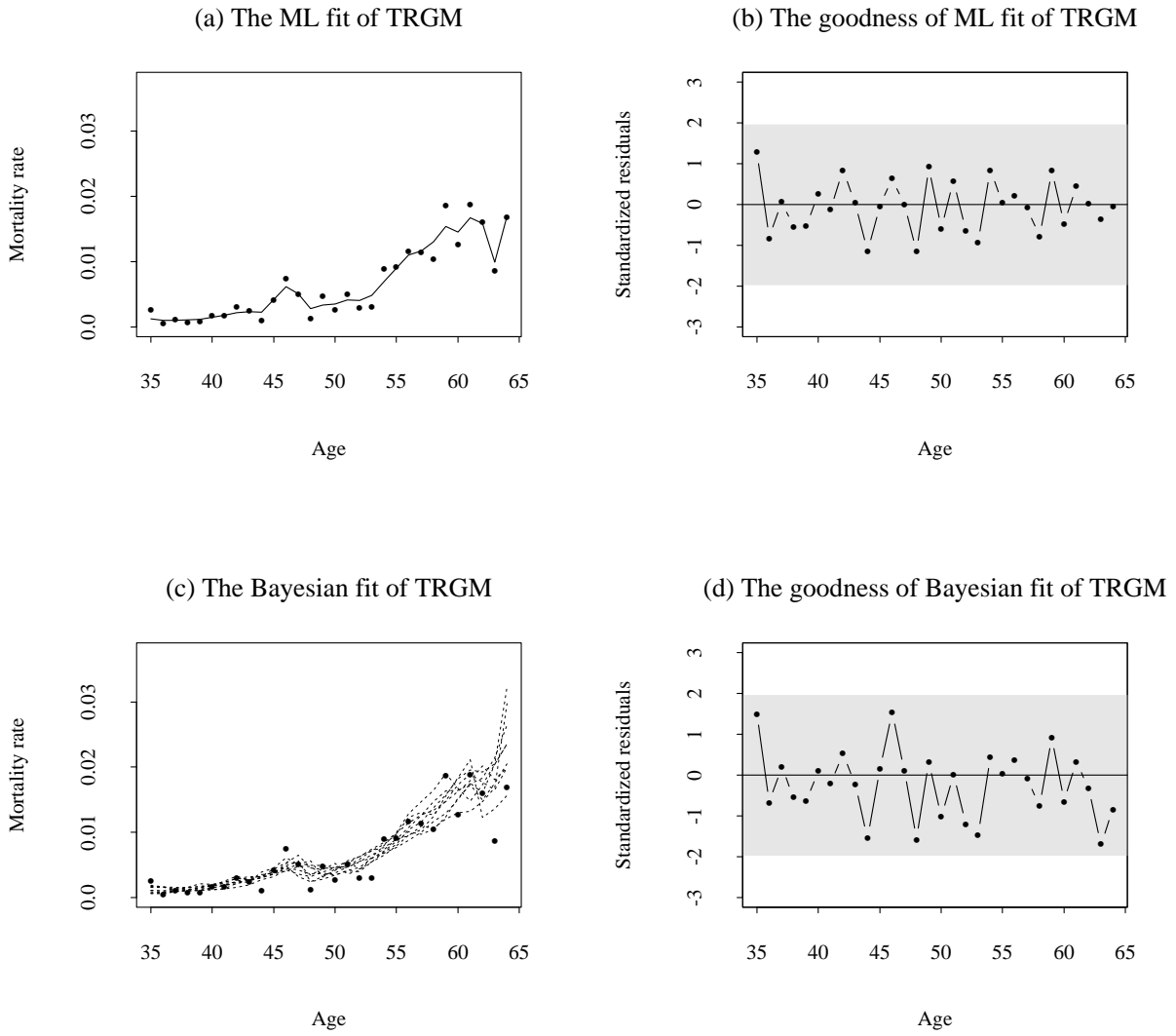


Figure 4: This is similar to Figure 1 but for the constrained binomial model with two risk groups.

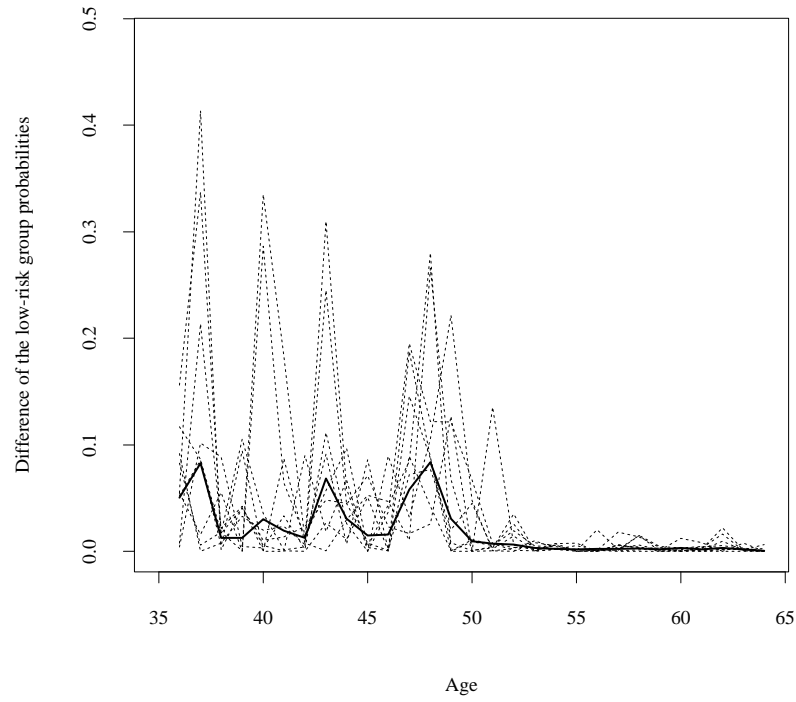


Figure 5: Nine posterior draws of the differences of the low-risk group probabilities ω_t in the constrained binomial model with two risk groups. The solid line is the point-wise posterior median.