# Bayesian Analysis of Overdispersed Binomial Data using Mixture Link Regression

Andrew M. Raim[1*], Marissa N. Gargano[1], Nagaraj K. Neerchal[2] & Jorge G. Morel[2]

[1]Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, 20233, U.S.A.

[2]Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, 21250, U.S.A.

**Abstract**

Overdispersion is commonly encountered in the analysis of categorical and count data. When it occurs, standard regression models may not adequately explain variability observed in the data. Finite mixture distributions arise in sampling a heterogeneous population, and data drawn from such a population will exhibit extra variability relative to any single subpopulation. The Mixture Link binomial distribution was recently developed to account for such heterogeneity in a generalized linear model setting. This model is completely likelihood-based, and maintains a link between the regression function and the overall mixture mean by assuming a certain random effects structure on the set representing enforcement of the link. This paper first presents an illustrative example in a heterogeneous population, comparing binomial regression with a binomial finite mixture of regressions and Mixture Link regression. We then compare the three models in a Bayesian setting using a classical dataset studying chromosome aberrations in atomic bomb survivors. The benefits of acknowledging the extra variation are seen through improved residual plots and widened prediction intervals. When regression on the overall mean is of interest and the heterogeneity is considered a nuisance, Mixture Link may be preferred over a finite mixture of regressions because only one regression function must be specified.

**Key Words:** Finite Mixture; GLM; Random Effects; Prediction Interval.

## 1. Introduction

In a binomial regression setting, we observe $T \in \{0, 1, \ldots, m\}$ successes for some event of interest, out of $m$ prescribed trials, and a covariate $\boldsymbol{x} \in \mathbb{R}^d$ which is associated with the probability of a success for a trial. In this "grouped" binomial setting, a single covariate value is observed for the $m$ trials. A typical but often overly simple assumption is that $T$ follows a binomial distribution $f(t \mid m, p) = \binom{m}{t} p^t (1-p)^{m-t}$, notated by $T \sim \text{Bin}(m, p)$, with $p = G(\boldsymbol{x}^T \boldsymbol{\beta})$. The function $G : \mathbb{R} \to [0, 1]$ is called an inverse link in the Generalized Linear Model (GLM) literature (McCulloch et al., 2008). We make use of both logit and probit links in this paper. $\boldsymbol{\beta} \in \mathbb{R}^d$ is a vector of coefficients for the linear regression function $\boldsymbol{x}^T \boldsymbol{\beta}$. In a typical data analysis, we observe $\{(T_i, m_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ and assume

$$T_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, p_i), \quad \text{where } p_i = G(\boldsymbol{x}_i^T \boldsymbol{\beta}). \tag{1}$$

A problem commonly encountered with this model is overdispersion, which can generally be said to occur when a given statistical model can not capture the

*Email: andrew.raim@gmail.com.

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

variability observed in the data. Specifically under model (1), $\mathrm{E}(T_i) = p_i$ and $\mathrm{Var}(T_i) = m_i p_i(1 - p_i)$, so that the variance is directly a function of the success probability. In analysis of real data, this can be a serious restriction.

Morel and Neerchal (2012) give an overview of some established approaches to handle overdispersion in analysis of binomial and categorical data. One approach is to generalize the simple binomial likelihood by assuming the presence of latent random variables. beta-binomial (Otake and Prentice, 1984), zero-inflated binomial (Hall, 2000), and random-clumped binomial (Morel and Nagaraj, 1993) models are all obtained in this way. The class of Generalized Linear Mixed Models is obtained by placing random effects into the regression function (McCulloch et al., 2008). Quasi-likelihood methods extend the likelihood in ways that do not yield a formal likelihood (i.e. random variables cannot be drawn from them), but allow the regression to be studied. A simple quasi-likelihood is obtained from placing a dispersion multiplier to the variance (Agresti, 2002, Section 4.7). Generalized Estimating Equations (GEE) represents a more sophisticated quasi-likelihood method; in the setting of nested data, an analyst assumes a "working" correlation structure for observations within a subject (Hardin and Hilbe, 2012).

In this paper, we focus on the Mixture Link binomial model proposed in Raim and Neerchal (2013) and Raim (2014). Mixture Link supposes that there are $J$ latent subpopulations with potentially different regression functions, just as in a finite mixture of regressions (Frühwirth-Schnatter, 2006, Section 9.4). However, unlike the finite mixture of regressions, Mixture Link treats the multiple regression functions as a nuisance and seeks only to model a single regression function for the overall probability of success of the entire (mixed) population. In this sense, Mixture Link leaves less opportunity for model misspecification and allows for a more parsimonious model. Whether the $J$ individual regression functions are of interest or are a nuisance is problem specific. Mixture Link is fully likelihood-based, and can be considered an alternative to beta-binomial and random-clumped binomial for modeling overdispersion in the basic binomial regression setting. However, Mixture Link has proven to be far more computationally challenging. For example, evaluation of the Mixture Link density requires numerical evaluation of multiple (univariate) integrals. Also, the likelihood appears to be not differentiable in some parts of the parameter space.

This paper explores Bayesian analysis for the Mixture Link model. The model does not appear to permit conjugate priors for closed form Gibbs sampling, so we make use of a basic Markov Chain Monte Carlo (MCMC) sampler. Bayesian computation helps to alleviate some of the challenges faced in frequentist analysis of Mixture Link. The results are compared to standard binomial regression and the finite mixture of regressions, both in a Bayesian setting, using a classical dataset studying chromosome aberrations in atomic bomb survivors. The benefits of capturing overdispersion relative to the simple binomial regression are readily seen through improved model fit and appropriately widened prediction intervals. Note that

The rest of the paper proceeds as follows. Section 2 recalls the Mixture Link and finite mixture of regression models, and gives an illustrative example of binomial regression in a mixed population. Section 3 presents details on Bayesian computation for the three models: binomial regression, finite mixture of binomials, and Mixture Link. Analysis of the chromosome aberration dataset is compared among the three models in Section 4. Finally, Section 5 gives concluding remarks.

## 2. Regression on the Mean of a Finite Mixture

We will first recall the binomial finite mixture of regressions model. Suppose there are $J$ possible regression functions in our population of interest, $\boldsymbol{x}^T\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{x}^T\boldsymbol{\beta}^{(J)}$, so that the probability of a success is determined by $G(\boldsymbol{x}^T\boldsymbol{\beta}^{(j)})$ in proportion $\pi_j$ of the population, for $j = 1, \ldots, J$. Let $Z_i$, for $i = 1, \ldots, n$, be discrete random variables which take on value $j$ with probability $\pi_j$, notated as $Z_i \overset{\text{ind}}{\sim} \text{Discrete}(1, \ldots, J; \boldsymbol{\pi})$ with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$. The finite mixture of regressions model can be written as

$$T_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, G(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(Z_i)})).$$

When the $Z_i$ are not observed, the likelihood is

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} \pi_j \binom{m_i}{t_i} \left[G(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(j)})\right]^{t_i} \left[1 - G(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(j)})\right]^{m_i - t_i} \right\},$$

with $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(J)}, \pi_1, \ldots, \pi_{J-1})$ typically taken to be unknown parameters in a data analysis. Note that $\pi_J = 1 - \sum_{j=1}^{J-1} \pi_j$, and is therefore redundant. The overall success probability of a single trial is

$$\text{E}\left(T_i/m_i \mid \boldsymbol{x}_i\right) = \sum_{j=1}^{J} \pi_j G(\boldsymbol{x}^T\boldsymbol{\beta}^{(j)}).$$

In the rest of the paper, we will use "BinMix" as a shorthand for the the binomial finite mixture of regressions model, and "BinMixJx" to refer to BinMix with $J = x$.

**Mixture Link Distribution.** The Mixture Link distribution starts with a similar assumption of a $J$ component finite mixture,

$$T_i \overset{\text{ind}}{\sim} f(t \mid m_i, \boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j \binom{m_i}{t_i} \mu_{ij}^{t_i} (1 - \mu_{ij})^{m_i - t_i},$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ is an element in the probability simplex $\mathcal{S}^J$ in $\mathbb{R}^J$, and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iJ})$ is an element in the unit cube $[0, 1]^J$. The mixture success probability for a single trial $\text{E}(T_i/m_i) = \boldsymbol{\mu}_i^T\boldsymbol{\pi}$ is linked to a regression $\boldsymbol{x}_i^T\boldsymbol{\beta}$ through an inverse link function $G$, as in the traditional GLM framework.[1] The quantity $\boldsymbol{\mu}_i^T\boldsymbol{\pi}$ is a composite parameter which does not appear explicitly in the likelihood, so special machinery is needed to enforce the link. Consider the set

$$A(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in [0, 1]^J : \boldsymbol{\mu}^T\boldsymbol{\pi} = G(\boldsymbol{x}_i^T\boldsymbol{\beta})\},$$

which is exactly the set of $\boldsymbol{\mu}_i$ which honors the link for a given $G(\boldsymbol{x}_i^T\boldsymbol{\beta})$ and $\boldsymbol{\pi}$. Mixture Link is formulated by decomposing $A(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\pi})$ into its convex hull,

$$A(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\pi}) = \left\{ \sum_{\ell=1}^{k_i} \lambda_\ell \boldsymbol{v}_\ell^{(i)} : \boldsymbol{\lambda} \in \mathcal{S}^{k_i} \right\} = \left\{ \boldsymbol{V}^{(i)}\boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathcal{S}^{k_i} \right\},$$

[1] We currently assume a linear regression function for simplicity.

where $\boldsymbol{V}^{(i)}$ is a $J \times k_i$ matrix with the vertices $\boldsymbol{v}_1^{(i)}, \ldots, \boldsymbol{v}_{k_i}^{(i)}$ as columns. By drawing $\boldsymbol{\lambda}^{(i)} \overset{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\boldsymbol{\alpha})$, $\boldsymbol{\mu}_i = \boldsymbol{V}^{(i)}\boldsymbol{\lambda}^{(i)}$ may be regarded as a random effect drawn from $A(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\pi})$. We therefore write Mixture Link as a hierarchy

$$
\begin{aligned}
&T_i \mid \boldsymbol{\mu}_i, \boldsymbol{\pi} \overset{\text{ind}}{\sim} \text{BinMix}(m_i, \boldsymbol{\mu}_i, \boldsymbol{\pi}), \\
&\boldsymbol{\mu}_i = \boldsymbol{V}^{(i)}\boldsymbol{\lambda}^{(i)}, \\
&\boldsymbol{V}^{(i)} = (\boldsymbol{v}_1^{(i)} \cdots \boldsymbol{v}_{k_i}^{(i)}) \text{ are vertices of } A(\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\pi}), \\
&\boldsymbol{\lambda}^{(i)} \overset{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\kappa, \ldots, \kappa).
\end{aligned}
$$

In the last stage, a Symmetric Dirichlet assumption constrains all dimensions to share the same $\kappa$ parameter. The density for a typical observation can then be written as

$$
f(t \mid m, \boldsymbol{\theta}) = \binom{m}{t} \sum_{j=1}^{J} \pi_j \int w^t (1-w)^{m-t} \cdot f_{\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}}(w)\, dw, \tag{2}
$$

where $f_{\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}}$ is the density of $\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}$. The notation $\boldsymbol{v}_{j\cdot}^T$ denotes the $j$th row of the matrix of vertices $\boldsymbol{V}$. The density is parameterized by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\pi}, \kappa)$, with $\boldsymbol{\pi} \in \mathcal{S}^J$ and $\kappa > 0$. In the case of no regression, when $G(\boldsymbol{x}_i^T \boldsymbol{\beta}) = p$ for $i = 1, \ldots, n$, we may take $\boldsymbol{\theta} = (p, \boldsymbol{\pi}, \kappa)$ with $p \in [0, 1]$.

**Beta Approximation to the Density.** A simple closed form for the density $f_{\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}}$ is not generally available (Provost and Cheong, 2000), which makes (2) difficult to compute efficiently. Raim (2014) observes through an empirical study that (2) can be well-approximated by replacing $f_{\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}}$ with a moment-matched beta distribution, for $j = 1, \ldots, J$, which has been shifted and scaled to have the same support. The moment-matched density is readily computed in statistical software packages. It is obtained as the density of $B_j^* = (u_j - \ell_j)B_j + \ell_j$, where $(\ell_j, u_j)$ is the support of $f_{\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}}$, $B_j \sim \text{Beta}(a_j, b_j)$, and

$$
a_j = (\bar{v}_{j\cdot} - \ell_j)^2 \left( \frac{k \boldsymbol{v}_{j\cdot}^T \boldsymbol{v}_{j\cdot} - (k\bar{v}_{j\cdot})^2}{k^2(1 + k\kappa)} \right)^{-2} \frac{u_j - \bar{v}_{j\cdot}}{u_j - \ell_j} - \frac{\bar{v}_{j\cdot} - \ell_j}{u_j - \ell_j}, \qquad b_j = a_j \left( \frac{u_j - \bar{v}_{j\cdot}}{\bar{v}_{j\cdot} - \ell_j} \right).
$$

Using the approximation, the density for a typical observation can be written as

$$
f(t \mid m, \boldsymbol{\theta}) = \binom{m}{t} \sum_{j=1}^{J} \pi_j \int w^t (1-w)^{m-t} \cdot f_{B_j^*}(w)\, dw, \tag{3}
$$

which is much simpler to compute numerically than (2). The approximated Mixture Link distribution can now be written as

$$
\begin{aligned}
&T_i \mid \boldsymbol{\mu}_i, \boldsymbol{\pi} \overset{\text{ind}}{\sim} \text{BinMix}(m_i, \boldsymbol{\mu}_i, \boldsymbol{\pi}), \\
&\mu_{ij} = (u_{ij} - \ell_{ij})B_{ij} + \ell_{ij}, \quad j = 1, \ldots, J \\
&B_{ij} \sim \text{Beta}(a_{ij}, b_{ij}),
\end{aligned}
$$

with $a_{ij}, b_{ij}, \ell_{ij}, u_{ij}$ computed from $\boldsymbol{V}^{(i)}$ for the $i$th observation in the same way $a_j, b_j, \ell_j, u_j$ are computed from $\boldsymbol{V}$, as described previously.

In this paper, we will make use of the beta approximated Mixture Link distribution only. Suppose random variable $T$ is drawn from Mixture Link based on $m$ trials with mixed probability of success linked to $\boldsymbol{x}^T\boldsymbol{\beta}$ through inverse link function $G$. We will write $T \sim \text{MixLink}_J(m, G(\boldsymbol{x}^T\boldsymbol{\beta}), \boldsymbol{\pi}, \kappa)$, and use the shorthand "MixLink" or "MixLinkJx" to refer to this model.

**Illustrative Example.** The following example illustrates the use of BinMix and Mixture Link in the presence of a mixed population. Consider drawing

$$T_i \overset{\text{ind}}{\sim} \begin{cases} \text{Bin}[50, \mu_1(x_i)] & \text{w.p. } \pi_1 = 0.1, \\ \text{Bin}[50, \mu_2(x_i)] & \text{w.p. } \pi_2 = 0.9, \end{cases}$$

for i = 1,..., 200, where

$$\mu_1(x) = G(1 + x), \quad \mu_2(x) = G(0 + 0.1x), \quad \mu(x) = \pi_1\mu_1(x) + \pi_2\mu_2(x).$$

The logit link is assumed throughout this example, so that $G(x) = 1/(1 + e^{-x})$ represents the CDF of the logistic distribution. The covariate $x_i$ is drawn randomly from a normal distribution with mean 0.5 and variance 4. Figure 1a shows the generated data along with the subpopulation mean functions $\mu_1(x)$, $\mu_2(x)$, and the mixed mean function $\mu(x)$. Fitting logistic regression to this data using maximum likelihood results in the estimates shown in Table 1a.[2] The resulting mean function is plotted in Figure 1b and appears to be a good estimate of the true mean.

However, there is a lack-of-fit with the logistic regression model. To see this we consider the randomized quantile residuals proposed by Dunn and Smyth (1996), which can be used with non-standard models such as finite mixtures. Randomized quantile residuals are based on the CDF transformation, and are expected to behave as a draw from N(0, 1) under an adequately fitting model. For independently drawn $y_i$, the residuals are computed as $e_i = \Phi^{-1}\{u_i\}$, where

$$u_i \overset{\text{ind}}{\sim} \text{Uniform}(a_i, b_i), \quad a_i = \lim_{\varepsilon \downarrow 0} F(y_i - \varepsilon \mid \hat{\boldsymbol{\theta}}), \quad \text{and} \quad b_i = F(y_i \mid \hat{\boldsymbol{\theta}});$$

$F(y_i \mid \boldsymbol{\theta})$ represents the CDF of $y_i$ under the proposed model and $\Phi^{-1}(\cdot)$ represents the quantile function of N(0, 1).

Figures 2a and 2d plot the quantile residuals from logistic regression applied to the example data. There are a number of observations with large residual values which are not well-explained by the model. These presumably correspond to observations generated from $\mu_1(x)$, since the estimated mean is not far from $\mu_2(x)$.

Next we consider fitting a finite mixture of $J = 2$ logistic regressions. Table 1b gives the resulting estimates and Figure 1c displays the estimated mean function. With the correct model, we can now recognize that there are the two groups with distinct regressions influencing their responses. Figures 2b and 2e show the corresponding quantile residuals; the model fit now appears to be adequate, as we would anticipate with the true model at hand.

Finally, we consider fitting the MixLink with $J = 2$ mixture components. Table 1c gives the estimates and Figure 1d shows the estimated mean function. Again we have captured the overall mean function, although the fit could potentially be more accurate with a more expressive regression function. Figures 2c and 2f show the corresponding quantile residuals. The MixLink fit is not as good as the (correct) finite mixture model. For example, the Q-Q plot indicates that some of the

[2]Note that all models are fit by numerical maximum likelihood in this example.

**Table 1**: Estimates for example dataset using three models.

(a) Logistic regression.

| | Estimate | SE | z-value | p-value |
|---|---|---|---|---|
| $\beta_0$ | 0.0817 | 0.0205 | 3.9890 | $< 0.0001$ |
| $\beta_1$ | 0.1191 | 0.0101 | 11.8010 | $< 0.0001$ |
| LogLik: | -724.77 | AIC: 1453.54 | | BIC: 1460.13 |

(b) BinMixJ2.

| | Estimate | SE | t-value | p-value |
|---|---|---|---|---|
| $\beta_1^{(1)}$ | 0.0134 | 0.0220 | 0.6076 | 0.5441 |
| $\beta_2^{(1)}$ | 0.0901 | 0.0104 | 8.6419 | $< 0.0001$ |
| $\beta_1^{(2)}$ | 1.1817 | 0.1468 | 8.0510 | $< 0.0001$ |
| $\beta_2^{(2)}$ | 0.9862 | 0.0990 | 9.9596 | $< 0.0001$ |
| $\pi$ | 0.9227 | 0.0231 | 40.0027 | 0.0010 |
| LogLik: | -582.95 | AIC: 1177.89 | | BIC: 1197.68 |

(c) MixLinkJ2.

| | Estimate | SE | t-value | p-value |
|---|---|---|---|---|
| $\beta_0$ | 0.0140 | 0.0202 | 0.6914 | 0.4901 |
| $\beta_1$ | 0.0821 | 0.0104 | 7.8600 | $< 0.0001$ |
| $\pi$ | 0.9274 | 0.0160 | 57.9670 | $< 0.0001$ |
| $\kappa$ | 0.4727 | 0.1157 | 4.0854 | $< 0.0001$ |
| LogLik: | -597.59 | AIC: 1205.18 | | BIC: 1221.67 |

outcomes are expected to be smaller under the fitted model. However, the fit is substantially better than the logistic regression model. In this example, MixLink is able to capture the overall mean $\mu(x)$ while accounting for much of the variability due to two distinct underlying subpopulations.

## 3. Bayesian Computation

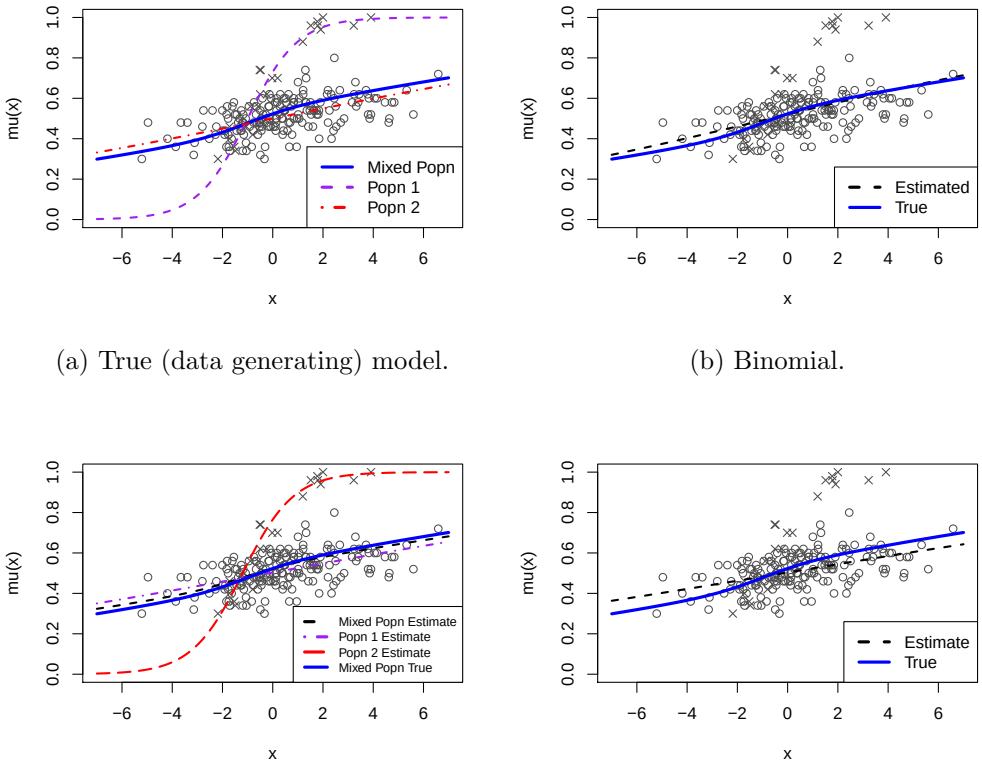
In this section, we present Bayesian algorithms to fit the three models under consideration: binomial regression, mixture of binomial regressions, and Mixture Link. A probit link is assumed throughout this section so that the inverse link $G$ is taken to be $\Phi$, the CDF of $\mathrm{N}(0,1)$. This assumption admits a closed form Gibbs sampler for Binomial and BinMix using augmented data approach of Albert and Chib (1993). It appears that a closed form Gibbs sampler can not be obtained with Mixture Link, but we may proceed with an off-the-shelf MCMC sampling algorithm.

**Binomial Regression.** Suppose the observed data follow a probit regression model,

$$y_i \overset{\text{ind}}{\sim} \mathrm{Bin}(m_i, \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})), \quad i = 1, \ldots, n. \tag{4}$$

Let the augmented model be

$$\begin{aligned} y_i &= \sum_{\ell=1}^{m_i} I(w_{i\ell} \geq 0), \quad i = 1, \ldots, n, \\ w_{i\ell} &\overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1), \quad i = 1, \ldots, n \text{ and } \ell = 1, \ldots, m_i, \\ \boldsymbol{\beta} &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\beta}}). \end{aligned}$$




(a) True (data generating) model.

(b) Binomial.

(c) BinMixJ2.

(d) MixLinkJ2.

**Figure 1**: Estimated mean for three models fitted to example dataset. Crosses (`x`) represent draws using mean function $\mu_1$ and circles (`o`) represent draws using $\mu_2$.

Here, each binomial observation $y_i$ is composed of iid continuous trials $w_{i1}, \ldots, w_{im}$; trials which are positive are observed as successes while the rest are observed as failures. A Gibbs sampler consists of the following steps:

1. Sample $\boldsymbol{\beta}$ from $\mathrm{N}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\Delta}^T\boldsymbol{w}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = [\boldsymbol{\Delta}^T\boldsymbol{\Delta} + \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}]^{-1}$,

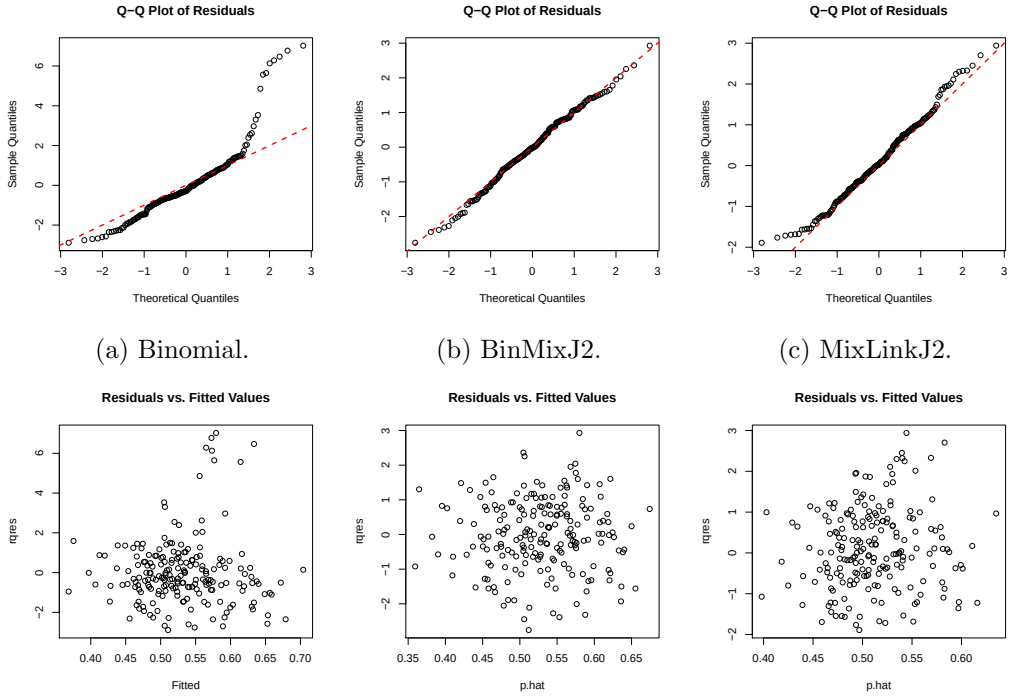
$$\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n), \quad \boldsymbol{w}_i = (w_{i1}, \ldots, w_{i,m_i}), \quad \text{and} \quad \boldsymbol{\Delta} = \begin{pmatrix} \mathbf{1}_{m_1} \otimes \boldsymbol{x}_1^T \\ \vdots \\ \mathbf{1}_{m_n} \otimes \boldsymbol{x}_n^T \end{pmatrix}.$$

2. For $i = 1, \ldots, n$,

   (a) Sample $w_{i1}, \ldots, w_{i,y_i}$ independently from $\mathrm{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1)$ truncated to $(0, \infty)$.
   (b) Sample $w_{i,y_i+1}, \ldots, w_{i,m_i}$ independently from $\mathrm{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1)$ truncated to $(-\infty, 0)$.

**Finite Mixture of Binomial Regressions.** The augmented data approach can be extended to the finite mixture

$$y_i \overset{\text{ind}}{\sim} \sum_{j=1}^{J} \pi_j \mathrm{Bin}(m_i, \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(j)})), \quad i = 1, \ldots, n. \tag{5}$$




(a) Binomial. (b) BinMixJ2. (c) MixLinkJ2.

(d) Binomial. (e) BinMixJ2. (f) MixLinkJ2.

**Figure 2**: Residual plots for example dataset.

Let the augmented model be

$$
\begin{aligned}
y_i &= \sum_{\ell=1}^{m_i} I(w_{i\ell} \geq 0), \quad \ell = 1, \ldots, n, \\
w_{i\ell} &\overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(z_i)}, 1), \quad i = 1, \ldots, n \text{ and } \ell = 1, \ldots, m_i, \\
\boldsymbol{\beta}^{(j)} &\overset{\text{ind}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\beta}}), \quad j = 1, \ldots, J, \\
z_i &\overset{\text{ind}}{\sim} \text{Discrete}(1, \ldots, J; \boldsymbol{\pi}), \quad i = 1, \ldots, n, \\
\boldsymbol{\pi} &\sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_J).
\end{aligned}
$$

In addition to the continuous trials $w_{ij}$ assumed in the binomial regression model, we also make use of latent subpopulation labels $z_i$ which have been discussed in Section 2. A Gibbs sampler consists of the following steps:

1. Sample $\boldsymbol{\beta}^{(j)}$ from $\mathrm{N}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(j)}} \boldsymbol{\Delta}_j^T \boldsymbol{w}_j, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(j)}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(j)}} = [\boldsymbol{\Delta}_j^T \boldsymbol{\Delta}_j + \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}]^{-1}$,

$$
\boldsymbol{w}_j = (\boldsymbol{w}_{a_1}, \ldots, \boldsymbol{w}_{a_{n_j}}), \quad \text{where } \{a_1, \ldots, a_{n_j}\} \overset{\text{def}}{=} \{i : z_i = j\}, \quad \text{and}
$$

$$
\boldsymbol{\Delta} = \begin{pmatrix} \mathbf{1}_{m_{a_1}} \otimes \boldsymbol{x}_1^T \\ \vdots \\ \mathbf{1}_{m_{a_{n_j}}} \otimes \boldsymbol{x}_n^T \end{pmatrix}.
$$

2. For $i = 1, \ldots, n$,
   (a) Sample $w_{i1}, \ldots, w_{i,y_i}$ independently from $\mathrm{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(z_i)}, 1)$ truncated to $(0, \infty)$.
   (b) Sample $w_{i,y_i+1}, \ldots, w_{i,m_i}$ independently from $\mathrm{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(z_i)}, 1)$ truncated to $(-\infty, 0)$.
3. Sample $\boldsymbol{\pi}$ from $\text{Dirichlet}(\sum_{i=1}^n I(z_i = 1) + \gamma_1, \ldots, \sum_{i=1}^n I(z_i = J) + \gamma_J)$.

**Mixture Link Regression.** We make use of a basic Random Walk Metropolis sampler, adapted from code in the `LearnBayes` package in `R` (R Core Team, 2015).[3] To do this, we assume the model

$$\begin{aligned} y_i &\overset{\text{ind}}{\sim} \text{MixLink}_J(m_i, \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}), \boldsymbol{\pi}, \kappa), \\ \boldsymbol{\beta} &\sim \text{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\beta}}), \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_J), \\ \kappa &\sim \text{Gamma}(a_\kappa, b_\kappa), \end{aligned}$$

where the the Gamma distribution is parameterized such that $\text{E}(\kappa) = a_\kappa b_\kappa$.

## 4. Bayesian Analysis of Chromosome Aberration Data

Awa et al. (1971) and Sofuni et al. (1978) study the effects of radiation exposure on chromosome aberrations in survivors of the atomic bombs that were used in Hiroshima and Nagasaki. A subset of the data is presented in Morel and Neerchal (2012) as an example of binomial regression with extra variation. This dataset contains data on $n = 648$ subjects in Hiroshima. For the $i$th subject, a chromosome analysis has been carried out on $m_i$ circulating lymphocytes to determine the count $t_i$ out of $m_i$ containing chromosome aberrations. As potential covariates, two types of radiation exposure have been measured: neutron radiation and gamma radiation. We denote $\text{NeuRad}_i$ and $\text{GamRad}_i$ as the respective radiation doses which have been standardized to have mean 0 and standard deviation 1.

Raim and Neerchal (2013) and Raim (2014) previously used this dataset to illustrate Mixture Link. Those works used numerical maximum likelihood for the analysis, and simply took the standardized sum of the two radiation doses as the covariate. In the present work, we looked at several other possible regression functions, which are shown in Table 2. The variable `PC1` is based on the first principal component of a design matrix with unstandardized neutron and gamma radiation doses as columns. The two principal components are characterized by the linear combinations $\boldsymbol{c}_1 = (0.707, 0.707)$ and $\boldsymbol{c}_2 = (-0.707, 0.707)$ with eigenvalues 1.9102 and 0.0898 respectively. Recall that $1/\sqrt{2} \approx 0.707$, so that $\boldsymbol{c}_1$ is a normalized version of the vector $(1, 1)$. Therefore `PC1` is simply the sum of the unstandardized doses, and it captures $1.9102/(1.9102 + 0.0898) = 95.51\%$ of the variation within this design matrix. Using AIC as a guide for model selection, we select Model 5 which includes an intercept, `PC1`, and a quadratic term `PC1^2`.

With the regression function fixed, we next proceeded to fit seven Bayesian models: Binomial, BinMix with $J = 2, 3, 4$, and MixLink with $J = 2, 3, 4$. Here are some details regarding the computations:

- For Binomial, we used 50,000 draws from the Gibbs sampler; the first 10,000 were discarded (burn-in) and one out of each remaining 50 was kept as the final MCMC sample (thinning). The prior variance for $\boldsymbol{\beta}$ was taken to be $\boldsymbol{\Omega}_{\boldsymbol{\beta}} = 1000 \cdot \boldsymbol{I}$.

- For BinMix, we took 100,000 draws from the Gibbs sampler, discarding the first 20,000 and keeping one out of each remaining 100. The prior variance for $\boldsymbol{\beta}$ was taken to be $\boldsymbol{\Omega}_{\boldsymbol{\beta}} = 1000 \cdot \boldsymbol{I}$ and the prior parameter for $\boldsymbol{\pi}$ was taken to be $\gamma = (1, \ldots, 1)$.

[3] http://cran.r-project.org/web/packages/LearnBayes

**Table 2**: Comparison of seven probit regression models fitted using maximum likelihood.

| | Model | LogLik | AIC |
|---|---|---|---|
| 1 | -1.6550 + 0.3685 GamRad + 0.2480 NeuRad - 0.0812 GamRad^2 - 0.0232 NeuRad^2 - 0.0399 GamRad*NeuRad | -1800.79 | 3613.58 |
| 2 | -1.6582 + 0.4678 GamRad + 0.1327 NeuRad - 0.1372 GamRad^2 | -1806.33 | 3620.66 |
| 3 | -1.7504 + 0.2539 GamRad + 0.1255 NeuRad | -1965.60 | 3937.20 |
| 4 | -1.7504 + 0.3725 TotalRad | -1966.95 | 3937.89 |
| 5 | -1.6591 + 0.4340 PC1 - 0.0704 PC1^2 | -1803.85 | 3613.69 |
| 6 | -1.7504 + 0.2657 PC1 | -1973.36 | 3950.72 |
| 7 | -1.6626 + 0.5922 TotalRad - 0.1330 TotalRad^2 | -1809.57 | 3625.15 |

- For MixLink, we took 200,000 draws from the Metropolis sampler, discarding the first 50,000 and keeping one out of the remaining 300. The hyperparameters were taken to be $\boldsymbol{\Omega_\beta} = 1000 \cdot \boldsymbol{I}$, $\gamma = (1, \ldots, 1)$, $a_\kappa = 1$ and $b_\kappa = 1/10$. The sampler produced draws $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \vartheta_3) \in \mathbb{R}^{d+J}$, where $\boldsymbol{\vartheta}_1 \in \mathbb{R}^d$, $\boldsymbol{\vartheta}_2 \in \mathbb{R}^{J-1}$, and $\vartheta_3 \in \mathbb{R}$. These were transformed to the appropriate parameter space by applying the transformations $\boldsymbol{\beta} = \boldsymbol{\vartheta}_1$, $\boldsymbol{\pi} = \text{mlogit}^{-1}(\boldsymbol{\vartheta}_2)$, and $\kappa = e^{\vartheta_3}$. Notice that the multinomial logit function $\text{mlogit}(\boldsymbol{\pi}) = (\log(\pi_1/\pi_J), \ldots, \log(\pi_{J-1}/\pi_J))$ is a bijection from $\mathbb{S}^J$ to $\mathbb{R}^{J-1}$. A starting value for MCMC was found by Laplace approximation using code that we extended from the `LearnBayes` package.

A comparison of Deviance Information Criteria (DIC) from the seven fitted models is shown in Table 3. We proceed focusing on the models Binomial, BinMix with $J = 3$, and MixLink with $J = 2$; further improvements to DIC are possible by continuing to increase $J$, but they appear to be diminishing.

To provide diagnostic checks on MCMC convergence, we examined trace plots, autocorrelation function (ACF) plots, and histograms for each parameter of the three selected models; most of the plots are not shown due to space restrictions. Diagnostics for the Binomial MCMC showed good mixing, low autocorrelation, and normal marginals. For BinMixJ3 and MixLinkJ2, examples of the most worrisome diagnostic plots are shown in Figure 3. Non-negligible autocorrelation is present in the saved draws for several parameters. There are signs of slight departure from normality in several parameters in BinMixJ3. The trace plot for $\kappa$ in MixLinkJ2 appears to show slower mixing than other parameters. These issues do not appear to be serious, but more draws and a higher thinning rate might further improve the diagnostics.

Table 4 summarizes the posterior draws for each of the three models. For Binomial, Table 4a shows very similar results as in Table 2 where maximum likelihood was used. MixLinkJ2, shown in Table 4c, gives similar coefficients for the regression; the standard errors are larger than in Binomial, but the credible intervals are a bit narrower. A summary plot for BinMixJ3 is shown in Table 4b for completeness.

Consequences of ignoring overdispersion are more clear when moving beyond estimates of the parameters. To see this, we compute randomized quantile residuals based on the posterior distribution and prediction intervals based on the posterior predictive distribution. Residuals are computed from the posterior as

**Table 3**: Comparison of Bayesian models for Chromosome Aberration dataset. "Elapsed" reports the time required for MCMC computation on a Linux workstation with an Intel Core i7-2600 quad core CPU operating at 3.40GHz. "Accept" reports the proportion of accepted Metropolis proposals.

| Model | J | DIC | Elapsed | Accept |
|---|---|---|---|---|
| Binomial | - | 3613.459 | 0h 16m | --- |
| BinMix | 2 | 3114.298 | 2h 44m | --- |
| BinMix | 3 | 2887.584 | 2h 48m | --- |
| BinMix | 4 | 2866.922 | 3h 16m | --- |
| MixLink | 2 | 2870.340 | 6h 07m | 0.1459 |
| MixLink | 3 | 2863.231 | 5h 59m | 0.1626 |
| MixLink | 4 | 2853.754 | 8h 27m | 0.1601 |

$e_i = \frac{1}{R}\sum_{r=1}^{R} \Phi^{-1}\{u_i^{(r)}\}$, where

$$u_i^{(r)} \overset{\text{ind}}{\sim} \text{Uniform}(a_i^{(r)}, b_i^{(r)}), \quad a_i^{(r)} = \lim_{\varepsilon \downarrow 0} F(y_i - \varepsilon \mid \hat{\boldsymbol{\theta}}^{(r)}), \quad b_i^{(r)} = F(y_i \mid \hat{\boldsymbol{\theta}}^{(r)}).$$

Here, $R$ is the number of draws obtained from the posterior. For each $i = 1, \ldots, n$, the posterior predictive distribution for the $i$th observation is computed by drawing

$$y_i^{(r)} \overset{\text{ind}}{\sim} \text{MixLink}_J(m_i, \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(r)}), \boldsymbol{\pi}^{(r)}, \kappa^{(r)}),$$

for $r = 1, \ldots, R$. A prediction $\hat{y}_i$ can be obtained from the mean of $y_i^{(1)}, \ldots, y_i^{(R)}$, and a 95% prediction interval can be obtained by taking the 0.025 and 0.975 quantiles.

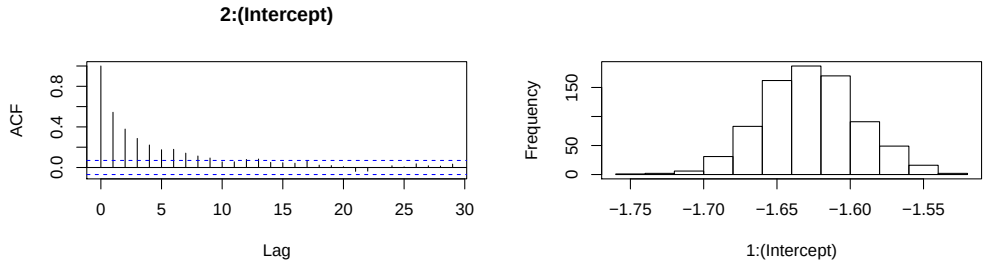
Figure 4 shows Q-Q plots of the residuals and plots of the residuals versus the predicted proportions $\hat{y}_i/m_i$. There is a clear lack-of-fit in Binomial, indicated by the presence of residuals well outside of the $(-3, 3)$ range anticipated under the standard normal assumption. The fit of BinMixJ3 and MixLinkJ2 appears to be comparable; both are significantly improved over Binomial. However, there is a clear pattern in all models' residuals showing that smaller $\hat{y}_i/m_i$ tend to have larger residuals. This may be an indication that we are missing an important covariate which was not available in the Morel and Neerchal (2012) version of the data.

Figure 5 plots predictions and prediction intervals against the variable `PC1`. The predictions themselves are not dramatically different between the three models, but prediction intervals for Binomial are too narrow. These intervals do not reflect the large variability that is especially apparent for larger values of `PC1`. Prediction intervals for BinMixJ3 and MixLinkJ2 are appropriately wider.

It is interesting that $J = 3$ mixture components appear to be needed for BinMix to arrive at a fit comparable to MixLinkJ2. This was not the case in our example from Section 2, where BinMix was the true data generating model. Now that we are in a real data analysis setting, the parsimony of MixLink over BinMix may be an advantage, as $J$ regression models must be specified for BinMix but only one must be specified for MixLink.

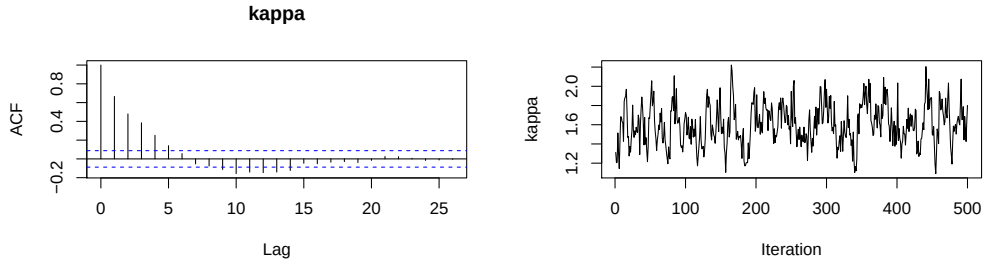
## 5. Conclusions

In this paper, we compared binomial regression, binomial finite mixture of regressions, and the recently proposed Mixture Link binomial regression model. First we presented an illustrative example using data simulated from the finite mixture. Ignoring the mixture led to variation which could not be expressed by the model;




(a) BinMixJ3: ACF plot for $\beta_1^{(3)}$.

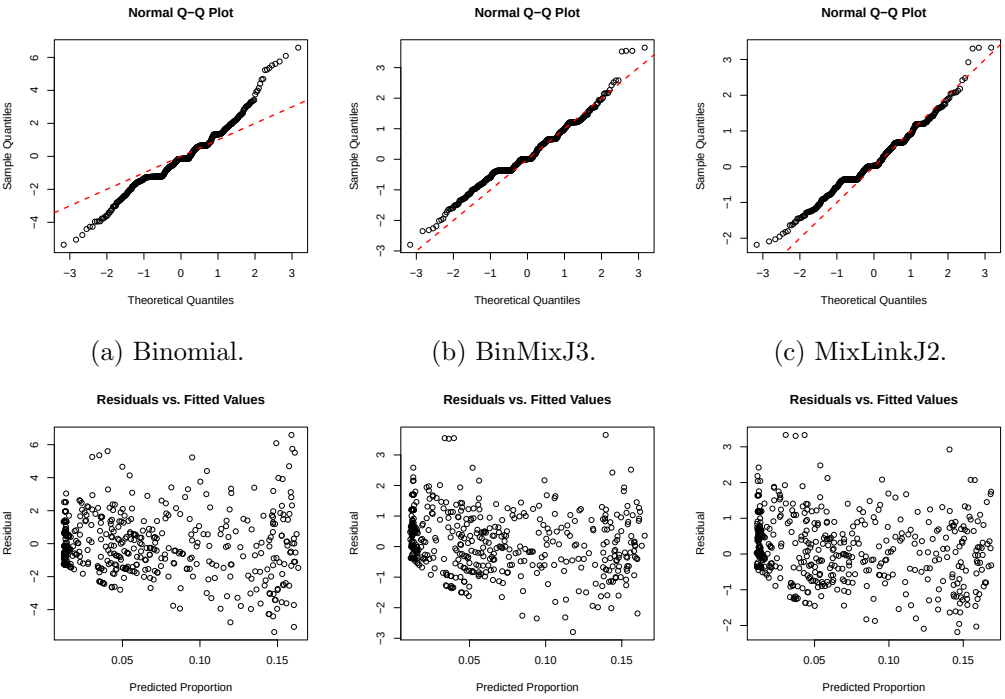
(b) BinMixJ3: Histogram for $\beta_1^{(2)}$.




(c) MixLinkJ2: ACF plot for $\kappa$.

(d) MixLinkJ2: Trace plot for $\kappa$.

**Figure 3**: Selected diagnostic plots from MCMC sampling in chromosome aberration analysis.




(a) Binomial.

(b) BinMixJ3.

(c) MixLinkJ2.

(d) Binomial.

(e) BinMixJ3.

(f) MixLinkJ2.

**Figure 4**: Posterior quantile residuals from chromosome aberration data.

**Table 4**: Summary of posterior draws for Bayesian models. The columns represent the posterior mean, standard deviation, and 2.5%, 50%, and 97.5% percentiles of the posterior draws which were not discarded due to burning or thinning.

(a) Binomial.

| | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| Intercept | -1.659011 | 0.011255 | -1.681000 | -1.658888 | -1.637236 |
| PC1 | 0.433437 | 0.010920 | 0.413804 | 0.433285 | 0.455346 |
| PC1^2 | -0.070334 | 0.003781 | -0.077804 | -0.070254 | -0.063513 |

(b) BinMixJ3.

| | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| 1:Intercept | -1.627399 | 0.032854 | -1.687566 | -1.628707 | -1.561228 |
| 1:PC1 | 0.427218 | 0.020600 | 0.389729 | 0.427630 | 0.467896 |
| 1:PC1^2 | -0.065063 | 0.007204 | -0.078454 | -0.064791 | -0.051752 |
| 2:Intercept | -2.087556 | 0.059700 | -2.216716 | -2.084888 | -1.980454 |
| 2:PC1 | 0.423070 | 0.053652 | 0.324533 | 0.419214 | 0.537476 |
| 2:PC1^2 | -0.097401 | 0.015693 | -0.130326 | -0.097630 | -0.066147 |
| 3:Intercept | -1.205755 | 0.038939 | -1.280563 | -1.207473 | -1.127679 |
| 3:PC1 | 0.574393 | 0.037490 | 0.500261 | 0.573859 | 0.650442 |
| 3:PC1^2 | -0.097022 | 0.013552 | -0.124498 | -0.096995 | -0.069645 |
| Pi1 | 0.533819 | 0.046071 | 0.440920 | 0.533025 | 0.620678 |
| Pi2 | 0.329537 | 0.047543 | 0.245021 | 0.328920 | 0.425162 |
| Pi3 | 0.136644 | 0.030366 | 0.081797 | 0.134373 | 0.199584 |

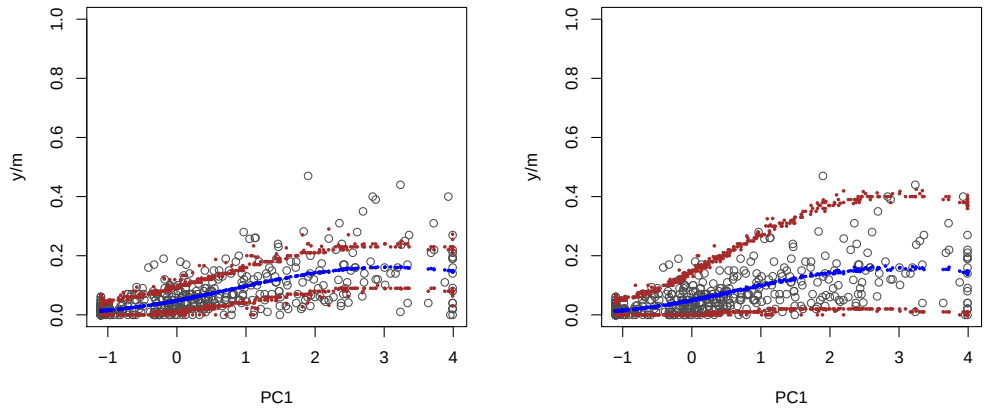
(c) MixLinkJ2.

| | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| Intercept | -1.650165 | 0.021010 | -1.689014 | -1.651180 | -1.610766 |
| PC1 | 0.448817 | 0.018791 | 0.415617 | 0.448000 | 0.488959 |
| PC1^2 | -0.075307 | 0.007872 | -0.090360 | -0.075550 | -0.060539 |
| Pi1 | 0.334374 | 0.017324 | 0.302826 | 0.332676 | 0.368576 |
| Pi2 | 0.665626 | 0.017324 | 0.631424 | 0.667324 | 0.697174 |
| kappa | 1.601981 | 0.215080 | 1.204079 | 1.585946 | 2.031749 |

resulting lack-of-fit was highlighted using quantile residuals, whose computation involves the CDF of the proposed model. Mixture Link was able to capture much of the variation in the example data. We also presented a Bayesian analysis of a classical chromosome aberration dataset. Prediction intervals were conveniently computed in the Bayesian setting. Binomial regression was unable to express the apparent uncertainty in predictions, but Mixture Link was able to provide suitably wider intervals. Interestingly, Mixture Link with $J = 2$ mixture components fit comparably to a finite mixture of $J = 3$ binomial regressions; we presume this is because Mixture Link is more parsimonious and offers less opportunity to misspecify regression functions.

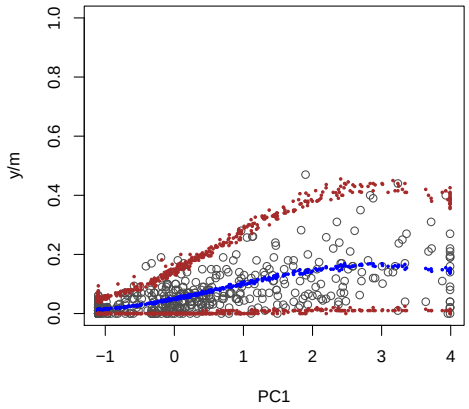
The Bayesian framework offers a certain convenience for the Mixture Link distribution while some difficulties remain in using frequentist approaches such as maximum likelihood. With code for the density function available (Raim et al., 2015), it was a relatively simple matter to program the likelihood and plug it into an MCMC sampler. It is worth noting that the diagnostics for Mixture Link with $J = 3$ appeared much worse than the $J = 2$ model. Namely, the $\boldsymbol{\pi}$ parameters showed evidence of poor mixing, high autocorrelation, and non-normality. The diagnostics might improve by running longer MCMC chains. However, it may be worthwhile to investigate other MCMC samplers for improved efficiency.

Note that Raim and Neerchal (2013) found a beta-binomial regression model




(a) Binomial.

(b) BinMixJ3.




(c) MixLinkJ2.

**Figure 5**: Predictions and 95% intervals from posterior predictive distribution. The blue curve in the center marks predictions, and the surrounding brown curves represent upper and lower (pointwise) intervals.

to fit equally as well as Mixture Link to the chromosome aberration dataset using numerical maximum likelihood. The beta-binomial model was permitted to have a second regression on its dispersion parameter, giving it an advantage. Likelihood-based models such as beta-binomial and random-clumped binomial currently have an advantange over Mixture Link because computations are relatively much simpler. However, each of these models addresses a different potential cause for overdispersion. Despite its increased complexity, Mixture Link may be a more appropriate alternative in some problems.


## Acknowledgements

Thanks to Tommy Wright and Kimberly Sellers, both in the Center for Statistical Research & Methodology at the U.S. Census Bureau, for reviewing this manuscript.

## References


A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2nd edition, 2002.

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

A. Awa, T. Honda, T. Sofuni, S. Neriishi, M. Yoshida, and T. Matsui. Chromosome-aberration frequency in cultured blood-cells in relation to radiation dose of A-bomb survivor. *The Lancet*, 298(7730):903–905, 1971.

P. K. Dunn and G. K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

D. B. Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039, 2000.

J. W. Hardin and J. M. Hilbe. *Generalized Estimating Equations*. Chapman and Hall/CRC, 2nd edition, 2012.

C. E. McCulloch, S. R. Searle, and J. M. Neuhaus. *Generalized, Linear, and Mixed Models*, volume 2. Wiley-Interscience, 2nd edition, 2008.

J. G. Morel and N. K. Nagaraj. A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80(2):363–371, 1993.

J. G. Morel and N. K. Neerchal. *Overdispersion Models in SAS*. SAS Institute, 2012.

M. Otake and R. L. Prentice. The analysis of chromosomally aberrant cells based on beta-binomial distribution. *Radiation Research*, 98(3):456–470, 1984.

S. B. Provost and Y.-H. Cheong. On the distribution of linear combinations of the components of a dirichlet random vector. *Canadian Journal of Statistics*, 28(2): 417–425, 2000.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

A. M. Raim. *Computational Methods in Finite Mixtures using Approximate Information and Regression Linked to the Mixture Mean*. Ph.D. Thesis, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 2014.

A. M. Raim and N. K. Neerchal. Modeling overdispersion in binomial data with regression linked to a finite mixture probability of success. In *JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association*, pages 2760–2774, 2013.

A. M. Raim, N. K. Neerchal, and J. G. Morel. Modeling overdispersion in R. Technical Report HPCF–2015–1, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2015.

T. Sofuni, T. Honda, M. Itoh, S. Neriishi, and M. Otake. Relationship between the radiation dose and chromosome aberrations in atomic bomb survivors of Hiroshima and Nagasaki. *Journal of Radiation Research*, 19(2):126–140, 1978.