7-13-2006

A FLEXIBLE GENERAL CLASS OF MARGINAL AND CONDITIONAL RANDOM INTERCEPT MODELS FOR BINARY OUTCOMES USING MIXTURES OF NORMALS

Brian Caffo
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, bcaffo@jhsph.edu

Ming-Wen An
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, man@jhsph.edu

Charles A. Rohde
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, crohde@jhsph.edu

A Flexible General Class of Marginal and Conditional Random Intercept Models for Binary Outcomes Using Mixtures of Normals

Brian Caffo, Ming-Wen An and Charles Rohde
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

July 13, 2006

Abstract

Random intercept models for binary data are useful tools for addressing between-subject heterogeneity. Unlike linear models, the non-linearity of link functions used for binary data force a distinction between marginal and conditional interpretations. This distinction is blurred in probit models with a normally distributed random intercept because the resulting model implies a probit marginal link as well. That is, this model is closed in the sense that the distribution associated with the marginal and conditional link functions and the random effect distribution are all of the same family. In this manuscript we explore another family of random intercept models with this property. In particular, we consider instances when the distributions associated with the conditional and marginal link functions and the random effect distribution are mixtures of normals. We show that this flexible family of models is related to several others presented in the literature. Moreover, we also show that this family of models offers considerable computational benefits. A diverse series of examples illustrates the wide applicability of the approach.

Keywords: Probit-normal, logit-normal, marginalized multilevel models

1 Introduction

Random intercept models for binary data are useful tools for addressing between subject heterogeneity. Typically, random intercept models are implemented by adding a normally distributed random effect into the linear predictor of a generalized linear model (or GLM, see Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), giving rise to the generalized linear mixed model (or GLMM, see Breslow and Clayton, 1993). Because of the nonlinearity of the link functions for binary GLMMs, such models force a distinction between parameter interpretations conditional on the random effect and marginal interpretations averaged over the random effect.

Random intercept models for binary outcomes with a probit link function and normally distributed random intercept (probit-normal models) have the interesting property that the marginal link function is the inverse of a normal cumulative distribution function (CDF). In this case, we say the model is "closed" in the sense that the distributions associated with the marginal and conditional link functions and the random effect distribution are all of the same family.

In this manuscript we explore a general family of closed random intercept models. In particular, we consider instances when the distribution associated with the conditional link function and the random effect distribution are mixtures of normals. Simple properties of mixture of normals then imply that the distribution function associated with the marginal link function is also a mixture of normals. We emphasize both the *conceptual* and *practical* benefits of this class of models. Notably, we explore models that yield conditional and marginal interpretations of parameters.

To summarize results, the principal conceptual benefit of the proposed model is that it contains a wide class of common models for binary data as either special or limiting cases. Furthermore, we highlight three interesting practical advantages of these models:

i marginal link functions can be approximated easily given fitted values for conditional models, without additional Monte Carlo or numerical integration,

ii marginalized multilevel models can be efficiently fit without the need for inverting a numerical integral,

iii simple and elegant Gibbs samplers can be applied for Bayesian modeling for arbitrary link functions.

The manuscript is laid out as follows. In Section 2 we present the notation and the model. In Section 3 we connect the mixture of normals model with several variants of random effect models in the literature. In Section 4 we illustrate with a diverse collection of useful applications of the mixture of normals approximation. Finally, in Section 5, we provide a summary and discussion of future work.

2 Random intercept model for binary outcomes

2.1 Notation

Consider the data given in Table 3, which arose from a teratology experiment (Weil, 1970), and was subsequently analyzed in Liang and Hanfelt (1994) and Heagerty and Zeger (1996). The objective is to compare the survival of rat pups in 16 control litters with that of the pups in the 16 treated litters. The treatment was a chemical agent administered to the mothers of each treated litter. We use this data set and experiment to motivate the model.

Assume that $\{Y_{ij}\}$ are repeated binary responses for subject/cluster $i = 1, \ldots, I$ and response $j = 1, \ldots, J_i$. Therefore, in the Teratology data set, Y_{ij} represents mortality or not (1 versus 0 respectively) for pup j from litter i. Let $\mathbf{x}_{ij}$ be a vector of covariates associated with Y_{ij}. For the Teratology data $\mathbf{x}_{ij} = (1, x_{ij1})^t$, containing an intercept term and a treatment indicator, respectively.

Let F_w^{-1} be a link function (see McCullagh and Nelder, 1989) that relates the probability of a success to a function of the covariates. As is typical for binary data, we assume that F_w (the inverse link function) is a distribution function, referred to as the "link distribution". We assume that

$$\Pr(Y_{ij} = 1 \mid U_i = u_i) = F_w\{\Delta_{ij} - u_i\}, \tag{1}$$

where the $\{U_i\}$ are cluster-specific random effects, used to model correlation and heterogeneity arising from unmeasured covariates specific to a cluster. The $\{U_i\}$ are assumed to be independent and identically distributed random variables, having distribution function F_u. Throughout we assume that the $\{Y_{ij}\}$ are conditionally independent given the $\{U_i\}$.

Users familiar with GLMMs will note two departures from common notation. First, the "transfer function", Δ_{ij}, is typically omitted and replaced with a linear combination of the covariates and slope parameters, such as

$$\Delta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta}^c. \tag{2}$$

This departure is adopted to consider a broader class of marginal and conditional models, which we describe in detail. Secondly, the random effect is subtracted in (1) rather than added, a convention that will be discussed below.

2.2 Conditional models

A conditional model specifies Δ_{ij} as in (2). The superscript c on the slope effects is used to denote that the effects are conditional, having an interpretation on the conditional link function's scale.

Defining the Δ_{ij} as such implies a marginal model. Specifically

$$\Pr(Y_{ij} = 1) = F_q\{\Delta_{ij}\}, \tag{3}$$

where F_q is the distribution of the sum of independent random variables having distribution functions F_u and F_w. To prove this fact, let $\{W_{ij}\}$ be iid draws from F_w, then note that

$$\begin{aligned}
\Pr(Y_{ij} = 1) &= E_{U_i}\left[\Pr(Y_{ij} = 1 \mid U_i = u_i)\right] \\
&= \int F_w\{\Delta_{ij} - u_i\} dF_u(u_i) \\
&= \int \Pr(W_{ij} \leq \Delta_{ij} - u_i \mid U_i = u_i) dF_u(u_i) \\
&= E_{U_i}\left[\Pr\left(W_{ij} + u_i \leq \Delta_{ij} \mid U_i = u_i\right)\right] \\
&= \Pr\left(W_{ij} + u_i \leq \Delta_{ij}\right) \\
&= F_q\{\Delta_{ij}\}.
\end{aligned}$$

From this proof, we hope that the reason for the somewhat unusual convention of subtracting the random intercept is now clear.

We summarize the basic properties of the conditional model as

Conditional model	$\Pr(Y_{ij} = 1 \mid U_i = u_i) = F_w(\Delta_{ij} - u_i)$
Transfer function	$\Delta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta}^c$
Random effect distribution	$\Pr(U_i \leq u_i) = F_u(u_i)$
Implied marginal model	$\Pr(Y_{ij} = 1) = F_q(\mathbf{x}_{ij}^t \boldsymbol{\beta}^c)$.

As an example, consider again the Teratology data set. Assume that F_w is the standard normal distribution, F_u is a normal distribution with 0 mean and variance $\sigma_{u,1}^2$, and Δ_{ij} is defined as in Equation 2. This model then corresponds to a probit-normal GLMM. By the standard properties of the normal distribution, the distribution of the sum of a standard normal (F_w) and a normal with mean 0 and variance $\sigma_{u,1}^2$ (F_u) results in F_q being a normal distribution with 0 mean and variance $1 + \sigma_{u,1}^2$. Thus, using Equation 3, we have the well known result (see Zeger et al., 1988, for example) that the induced marginal model is

$$\Pr(Y_{ij} = 1) = F_q(\Delta_{ij}) = F_q(\beta_0^c + x_{ij1}\beta_1^c) = \Phi\left\{\frac{\beta_0^c + x_{ij1}\beta_1^c}{(1 + \sigma_{u,1}^2)^{1/2}}\right\},$$

where Φ denotes the standard normal distribution function. Hence, the marginal link is also a probit, with the marginal effects being scaled versions of the conditional effects, $\boldsymbol{\beta}^c/(1 + \sigma_{u_1}^2)^{1/2}$.

2.3 Marginal Models

Consider again the Teratology probit-normal example from the previous section - i.e. F_w is a standard normal and F_u is a normal with mean 0 and variance $\sigma_{u,1}^2$. Had we defined

$$\Delta_{ij} = (\beta_0^m + x_{ij1}\beta_1^m)(1 + \sigma_{u,1}^2)^{1/2},$$

then the marginal probability of success would satisfy

$$\Pr(Y_{ij} = 1) = F_q(\Delta_{ij}) = F_q\left\{(\beta_0^m + x_{ij1}\beta_1^m)(1 + \sigma_{u,1}^2)^{1/2}\right\} = \Phi(\beta_0^m + x_{ij1}\beta_1^m).$$

Therefore, the estimated slope parameters would have a marginal probit interpretation without rescaling; hence the superscript m. That is, appropriately defining Δ_{ij} results in parameters with marginal interpretations.

In fact, Heagerty and Zeger (2000) showed that this technique can be applied more generally. Specifically, consider defining

$$\Delta_{ij} = F_q^{-1}\{F_w(\mathbf{x}_{ij}^t \boldsymbol{\beta}^m)\}. \tag{4}$$

Under this definition for Δ_{ij} and using (3), the marginal probability of success satisfies

$$\Pr(Y_{ij} = 1) = F_q(\Delta_{ij}) = F_q\left[F_q^{-1}\{F_w(\mathbf{x}_{ij}^t \boldsymbol{\beta}^m)\}\right] = F_w(\mathbf{x}_{ij}^t \boldsymbol{\beta}^m)$$

That is, under an appropriate modification of Δ_{ij}, the slope parameters can be given a marginal interpretation with F_w as the link distribution. We summarize the marginal model with

Conditional model	$\Pr(Y_{ij} = 1 \mid U_i = u_i) = F_w(\Delta_{ij} - u_i)$
Transfer function	$\Delta_{ij} = F_q^{-1}\{F_w(\mathbf{x}_{ij}^t \boldsymbol{\beta}^m)\}$
Random effect distribution	$\Pr(U_i \leq u_i) = F_u(u_i)$
Implied marginal model	$\Pr(Y_{ij} = 1) = F_w(\mathbf{x}_{ij}^t \boldsymbol{\beta}^m)$.

Marginalized multilevel models defined as such offer several advantages over competing methods. Unlike generalized estimating equations (GEE, see Liang and Zeger, 1986), they enjoy the benefits of a completely specified model, which includes the ability to plot profile likelihoods, the availability of likelihood ratio tests and Bayesian analysis and the relaxation on assumptions for missing data. Also, these models are more parsimonious and extensible than other marginal likelihood based models (see Lang and Agresti, 1994).

2.4 Mixtures of normals

The distinction between the conditional and marginal approaches is especially interesting for the probit-normal model, because of the fact that the probit-normal model is closed - the conditional, random effect and marginal link distributions all belong to the same family. In this manuscript we present another closed random intercept model for binary data that is considerably more flexible than the probit-normal model. In particular, when F_w and F_u are mixtures of normal distributions, then so is F_q.

To prove this, consider a model of the form

$$F_w(w) = \sum_{l=1}^{L_w} \pi_{w,l} \Phi\left(\frac{w - \mu_{w,l}}{\sigma_{w,l}}\right) \quad \text{and} \quad F_u(u) = \sum_{l=1}^{L_u} \pi_{u,l} \Phi\left(\frac{u - \mu_{u,l}}{\sigma_{u,l}}\right),$$

where, the $\{\pi_{w,l}\}$ and $\{\pi_{u,l}\}$ are each assumed to be greater than 0 and sum to one. Using simple properties of mixtures of normals and Equation 3, we have that

$$F_q(q) = \sum_{l=1}^{L_w} \sum_{l'=1}^{L_u} \pi_{w,l}\pi_{u,l'} \Phi\left(\frac{q - \mu_{w,l} - \mu_{u,l'}}{(\sigma_{w,l}^2 + \sigma_{u,l'}^2)^{1/2}} \right). \tag{5}$$

That is, under this model, the random effect, conditional and marginal link distributions are all mixtures of normals. We summarize the model as

$$\begin{array}{ll} \text{Conditional model} & \Pr(Y_{ij} = 1 \mid U_i = u_i) = \sum_{l=1}^{L_w} \pi_{w,l} \Phi\left(\frac{\Delta_{ij} - u_i - \mu_{w,l}}{\sigma_{w,l}} \right) \\ \text{Transfer function} & \Delta_{ij} \text{ defined by either (2) or (4)} \\ \text{Random effect distribution} & \Pr(U_i \leq u_i) = \sum_{l=1}^{L_u} \pi_{u,l} \Phi\left(\frac{u_i - \mu_{u,l}}{\sigma_{u,l}} \right) \\ \text{Implied marginal model} & \Pr(Y_{ij} = 1) = \sum_{l=1}^{L_w} \sum_{l'=1}^{L_u} \pi_{w,l}\pi_{u,l'} \Phi\left(\frac{\Delta_{ij} - \mu_{w,l} - \mu_{u,l'}}{(\sigma_{w,l}^2 + \sigma_{u,l'}^2)^{1/2}} \right). \end{array} \tag{6}$$

To summarize, the model of interest in this manuscript combines the conditional and marginal approaches, while adding the constraint that the conditional link and random effect distributions are both mixtures of normals.

For completeness, we add that the log-likelihood for (6) is

$$\sum_{i=1}^{I} \log \int_{u_i} \prod_{j=1}^{J_i} F_w(\Delta_{ij} - u_i)^{y_{ij}} \{1 - F_w(\Delta_{ij} - u_i)\}^{1 - y_{ij}} dF_u(u_i), \tag{7}$$

an equation that holds regardless of whether F_w and F_u are mixtures of normals.

To illustrate a potential use, consider the specific instance summarized by the following

$$\begin{array}{ll} \text{Conditional model} & \pi_{w,1}\Phi\left\{(\Delta_{ij} - u_i)/\sigma_{w,1}\right\} + \pi_{w,2}\Phi\left\{(\Delta_{ij} - u_i)/\sigma_{w,2}\right\} \\ \text{Transfer function} & \Delta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta}^c \\ \text{Random effect distribution} & \pi_{u,1}\Phi\left\{(u_i - \mu_{u,1})/\sigma_{u,1}\right\} + \pi_{u,2}\Phi\left\{(u_i - \mu_{u,2})/\sigma_{u,2}\right\} \\ \text{Implied marginal model} & \sum_{l=1}^{2} \sum_{l'=1}^{2} \pi_{w,l}\pi_{u,l'} \Phi\left\{(\mathbf{x}_{ij}^t \boldsymbol{\beta}^c - \mu_{u,l'})/(\sigma_{w,l}^2 + \sigma_{u,l'}^2)^{1/2}\right\}. \end{array}$$

One could specify $\pi_{w,1}$ $\sigma_{w,1}$ and $\sigma_{w,2}$ to approximate the logistic distribution, which produces a model that retains the computational benefits of this mixture approach (discussed later) while (approximately) retaining the convenient interpretation of the logit. Estimating $\pi_{u,1}$, $\mu_{u,1}$, $\mu_{u,2}$, $\sigma_{u,1}$ and $\sigma_{u,2}$ leads to a more flexible random effect distribution than the univariate normal.

Of course, Model 6 is excessively rich with all of the mixture probabilities, means and variances left unspecified; estimating both the conditional link distribution and the random effect distribution is a hopeless cause for most binary data sets. However, by specifying components of one or both of the free mixture distributions, one can achieve a variety of important models. In what follows we explore these ideas.

3 Literature review

In this section we argue the principal conceptual benefit of the modeling framework (6). That is, the proposed model contains several important random intercept models for binary data as special or limiting cases.

GLM and GLMMs

Clearly if F_u is degenerate at 0 and $\Delta_{ij} = \mathbf{x}_{ij}^t\beta$, then the model yields a GLM for binary data. Extending this setting so that F_u is not degenerate and $L_u = 1$ and $\mu_{u,1} = 0$ yields a GLMM for binary data with a normally distributed random intercept (see Breslow and Clayton, 1993; Agresti et al., 2000).

To be technical, only those GLM and GLMMs for binary data whose conditional link distribution, F_w, is a mixture of normals are special cases of the model we have suggested. However, all of the common link functions (logit, complementary log-log) can be obtained as limiting cases. In Appendix C we provide an algorithm to solve for $\pi_{w,l}$, $\sigma_{w,l}$ and $\mu_{w,l}$ that yields very accurate approximations for a finite number of mixture components.

As an example, consider a mixture of normals as an approximation of the logistic distribution. The results using the algorithm in Appendix C with 150 quadrature points and $\{\mu_{w,j}\} = \{0\}$ yields the values given in Table 1. Figure 1 shows how accurate the approximation is, by depicting the exact logistic quantiles by a mixture of normals approximation. The mixture of normals approximation, with 5 mixture components, is nearly exact to logits of ± 10. By comparison, the plot also shows the standard normal and T quantiles, both of which are also used as approximations to the logit (see Caffo and Griswold, 2005). The linearity of the probit approximations breaks down at logits of around ± 3, while the T approximation around ± 5. Furthermore, we note that the mixture of normals approximation applies generally, to links other than the logistic, and can be made more accurate by simply adding more mixture components.

Approximating the logistic distribution with a single normal distribution or mixture of normals has a rich history (see Demidenko, 2004, and the references therein). Perhaps most relevant, Monahan and Stefanski (1992) used weighted Gaussian distributions to explore the logistic-normal integral.

Latent variable models

Representing the link function by a latent variable was considered in the proof of Equation 3. In Section 4.3 we consider a much more ambitious latent variable representation of Model 6, using latent variables to represent the normal mixture distributions as well. The general latent variable approach to binary data was considered in Albert and Chib (1993), who also introduced a Gibbs sampler that motivates the one presented in Section 4.3. Relevant extensions to multivariate settings were considered in Chib and Greenberg (1998); however they focused on probit links and more general covariance structures than the random intercept models considered here.

Marginalized multilevel models

Consider again the instance where $L_u = 1$, $\mu_{u,1} = 0$ (the random intercept is normally distributed). As described in Section 2.3, Heagerty and Zeger (2000) defined the Δ_{ij} to be non-linear (see Equation 4), so that the slope parameters have linear interpretations on the marginal link's scale. These marginalized multilevel models for binary data are a special case of the models presented (by appropriately defining Δ_{ij}). Moreover, later we demonstrate that using mixtures of normals for the link distribution can greatly facilitate computing for these models.

A potentially negative aspect of this model is that, because Δ_{ij} is defined non-linearly, the conditional model is non-linear. The degree to which this is true depends on how close to linear $F_q^{-1}F_w$ is. However, this may be of no concern whatsoever if only marginal interpretations are required (though see Lee and Nelder, 2004).

A clear generalization of the marginalized model would replace F_w in Equation 4 with any other desired link distribution, thus, allowing the conditional and marginal link functions to be different. This idea was explored in Griswold (2005) and extended to ordered multinomial data in Caffo and Griswold (2005). Again, this approach easily fits into the current framework by appropriately redefining Δ_{ij}.

Estimating the link function

There has been a relatively small amount of research using mixtures of normals to estimate the link distribution. Geweke and Keane (1997) used a mixture of normals as a link function for dichotomous choice models. They presented an MCMC algorithm for fitting the model, including estimating the mixture components. In related work, Erkanli et al. (1993) used mixtures of normals to estimate the link function for ordinal data models and also presented an MCMC algorithm for estimating the mixture components. These approaches are conceptually related to the proposed model by forcing the random effect distribution to be degenerate at 0 and estimating $\{\pi_{w,l}\}$, $\{\mu_{w,l}\}$ and $\{\sigma_{w,l}\}$.

Estimating the random effect distribution

In contrast, using mixtures to estimate the random effect distribution has received much more attention. Perhaps most relevant, Magder and Zeger (1996) used mixtures of normals as the random effect distribution and estimated the mixture parameters with an MCMC algorithm. This corresponds to estimating the $\{\mu_{u,l}\}$, $\{\sigma_{u,l}\}$ and $\{\pi_{u,l}\}$. Aitkin (1999) and Follmann and Lambert (1989) used discrete mixtures to non-parametrically estimate the random effect distribution using maximum likelihood. Such models are obtained under the current framework as the $\{\sigma_{u,j}\}$ tend to 0 and $\{\mu_{u,j}\}$ and $\{\pi_{u,j}\}$ are estimated.

Summary

It is our goal that this literature review demonstrates that many of the primary models for binary data are closely related to the mixture of normals model (6). The fact that the model can synthesize so many other approaches is its main conceptual benefit. We now present a battery of examples that illustrates the practical utility of using mixtures of normals to approximate link functions.

4 Examples

In addition to synthesizing many common models, Model 6 offers many practical benefits as well. In this section we explore a subset of these practical considerations illustrated through four data sets. We explore two marginal and one conditional modeling settings, where computations are significantly simplified by using mixtures of normals. Moreover, we consider a case where mixture modeling of the random effect offers additional protection against model misspecification.

We consider four well studied data sets for illustration:

1. The Teratology data, introduced in Section 2.1.
2. The Approval Rating data set given in Table 2. This 2×2 contingency table cross-classifies approval ratings of the British Prime Minister collected at two occasions. Here, Y_{ij} represents approval (1) or not (0) for individual i on occasion j, where $j = 1, 2$ for the two sampling occasions. The covariate vector, $\mathbf{x}_{ij} = (1, x_{ij})^t$, contains an intercept term and an indicator function representing occasion, taking the value 1 when $j = 2$.
3. The Crossover data, given in Table 4, concerns a well-studied crossover study from Jones and Kenward (1987). Here, Y_{ij} represents an abnormal (1) or a normal (0) response for subject i during period j for $j = 1, 2$. The objective is to study the response in relation to the treatment and period. Thus, $\mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2})^t$ contains an intercept term, a treatment indicator and a period indicator, taking the value 1 for the second period.

4. The Item Response data, given in Table 5, concerns subjects' response to three scenarios (given in the table) on abortion stratified by gender. We let Y_{ij} be the response of subject i on question j, where a response of 1 is supportive of legalized abortion (and 0 is not). The covariate vector, $\mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2}, x_{ij3})^t$, contains an intercept term, an indicator for male gender, an indicator for Scenario 1, and an indicator for Scenario 2, respectively. We use the Item Response data to illustrate an instance where a mixture random effect distribution is warranted.

To focus this discussion, we assume that the principal parameter of interest for each data set is: the (marginal or conditional) log odds-ratio comparing treated to controls in the Teratology data, the log odds-ratio comparing time 2 to time 1 for the Approval Rating data, the log odds-ratio comparing treated to controls in the Crossover data and the log odds-ratio comparing males to females in the Item response data. Therefore, in each case the regressor corresponding to the effect of interest is x_{ij1}.

4.1 Post-hoc calculation of marginal effects

Given results from a conditional random effect model, an obvious question asks, "What is the corresponding marginal effects and link distribution?". Such a question is especially relevant in situations such as in interpreting published results, where only effect estimates (and not the original data) are available. Model 6 allows one to approximate the necessary calculations easily.

Consider the conditional logit model

$$\text{logit}\,\{\Pr(Y_{ij} = 1 \mid U_i = u_i)\} = \mathbf{x}_{ij}^t \boldsymbol{\beta}^c - u_i \;\text{ and }\; U_i \sim \mathrm{N}(0, \sigma_{u,1}^2). \tag{8}$$

If we are willing to accept the approximation that F_w is the 5 component mixtures of normals, then Model 8 is simply a special case of Model 6. Hence, we have that

$$\Pr(Y_{ij} = 1) = F_q(\mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}}^c) = \sum_{l=1}^{L_w} \pi_{w,l} \Phi\left(\frac{\hat{\mathbf{x}}_{ij}^t \hat{\boldsymbol{\beta}}^c - \mu_{w,l}}{(\sigma_{w,l}^2 + \sigma_{u,1}^2)^{1/2}} \right), \tag{9}$$

where $\{\pi_{w,l}\}$, $\{\mu_{w,l}\}$ and $\{\sigma_{w,l}\}$ are from Table 1.

Below we use this approximation to obtain marginal logit interpretations from conditional logit models. However, before doing so, we emphasize the benefits of Equation 9 over Monte Carlo and numerical integration, which can also give very accurate approximations of marginal effects. For example, unlike numerical integration or Monte Carlo approximations, the approximation (9) can be performed quickly and easily. In addition, obtaining delta method estimates of standard errors is also easy. Furthermore, the method applies to any conditional link function, provided the relevant mixture components are known. Finally, and perhaps most importantly, we note that this method leads to an accurate and simple approximation to the marginal link distribution, F_q, whereas quadrature or Monte Carlo approximations only yield F_q for specific values of the covariates.

Table 7 gives estimated marginal logit effects for the four data sets calculated using (9). To illustrate the calculations, consider the Teratology dataset. The fitted values (SE) from Model 8 using the SAS procedure NLMIXED are $\hat{\beta}_0^c = 2.63\ (0.48)$, $\hat{\beta}_1^c = -1.08\ (0.63)$ and $\hat{\sigma}_{u,1} = 1.35\ (0.33)$. Plugging the estimated parameters into (9) yields a marginal probability of death of 0.76 for the treated and 0.88 for the untreated. Then, the marginal log odds ratio of death (SE) comparing the treated to the control litters is $\mathrm{logit}(0.76) - \mathrm{logit}(0.88) = -0.86\ (0.51)$ (see Appendix D for details about obtaining standard errors).

Table 7 applies these techniques to the three other data sets as well, each time taking the conditional estimates output by SAS (Table 6). Because of the additional covariates in the Crossover and Item Response data sets, the estimated marginal logit effects are reported within strata.

4.2 Easier marginalized multilevel models

The previous section addressed the issue of obtaining marginal effects from conditional results, which is useful when interpreting published results without access to the underlying data. However, when the data are available and marginal interpretations are desired, direct fitting is preferable. This section illustrates how the mixture of normals modeling framework can ease the calculations required to directly obtain marginal estimates.

We consider the marginal Model 6 where Δ_{ij} is given by Equation 4. Furthermore, assume that the $U_i \sim N(0, \sigma^2_{u,1})$ and F_w is the 5 component mixtures of normals approximation to the logistic distribution function.

The benefit of using the F_w as a mixture of normals rather than the exact logistic distribution is that there is a closed form for F_q (see Equation 4); also its quantiles can easily be calculated using Newton's method. Hence, representing the logistic distribution as such eliminates the difficult task of numerically approximating the convolution integral defining F_q and its inverse. It should be emphasized that while defining F_w as a mixture of normals eases the calculation of F_q and hence Δ_{ij}, calculation of the likelihood (7) still requires numerical integration, for which we employed Gauss/Hermite quadrature.

We implemented this model for the four data sets. We highlight the use of profile likelihoods - the functions obtained by maximizing the likelihood for each value of the parameter of interest. See Royall (1997) for more information regarding the benefits and interpretation of profile likelihoods.

The results of the model fits are given in Table 8. For example, for the Teratology data, -0.86 (the estimate for β^m_1) estimates the change in the marginal log-odds of death comparing a treated pup to a control. For each of the data sets, Figure 2 shows the profile likelihood with 1/8 and 1/16 reference lines see (see Royall, 1997) for the parameter of interest (β^m_1) and the variance component ($\sigma_{u,1}$) for each of the four models.

4.3 Bayesian analysis

In this section, we illustrate how specific instances of Model (6) are particularly well suited for Bayesian analysis via MCMC. We note that similar methods utilizing latent variables have been proposed to simulate from the posterior distributions of parameters for binary and multinomial responses (see Albert and Chib, 1993; McCulloch and Rossi, 1994; Chib et al., 1998; Imai and van Dyk, 2005). In addition, close variants of the sampling schemes can be used for the Monte Carlo EM algorithm (see Chib et al., 1998; Natarajan et al., 2000).

We apply these methods to binary responses with random effects, using the mixture of normals link approximation (similar to Geweke and Keane, 1997; Erkanli et al., 1993).

Consider the latent variable representation of Model (6) given by

1. $\{D_{u,i}\}$ are iid discrete random variables with support $1, \ldots, L_u$ so that $\Pr(D_{u,i} = l) = \pi_{u,l}$,
2. the $\{U_i\}$ given that the $\{D_{u,i} = d_{u,i}\}$ are independent $N(\mu_{d_{u,i}}, \sigma^2_{u,d_{u,i}})$,
3. the $\{D_{w,ij}\}$ are discrete iid random variables with support $1, \ldots, L_w$ so that $\Pr(D_{w,ij} = l) = \pi_{w,l}$,
4. the $\{M_{ij}\}$ given that the $\{D_{w,ij} = d_{w,ij}\}$ and $\{U_i = u_i\}$ are independent Normals with mean $\mu_{w,d_{w,ij}} + u_i - \Delta_{ij}$ and variance $\sigma^2_{w,d_{w,ij}}$,
5. the $\{Y_{ij}\}$ are 1 iff $M_{ij} \leq 0$ and 0 otherwise,
6. each $\Delta_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta}^c$,

To summarize the model, items 1 and 2 yield the mixture model for F_u, items 3-5 yield the conditional model for the y_{ij} and item 6 forces a conditional interpretation for the $\boldsymbol{\beta}^c$. To prove that 3-5 induces the mixture of normals model for the y_{ij}, consider

$$\begin{aligned}
\Pr(Y_{ij} = 1 \mid U_i = u_i) &= \Pr(M_{ij} \leq 0 \mid U_i = u_i) \\
&= \sum_{l=1}^{L_w} \Pr(M_{ij} \leq 0 \mid U_i = u_i, D_{w,ij} = l) \Pr(D_{w,ij} = l) \\
&= \sum_{l=1}^{L_w} \Phi\left\{\frac{\Delta_{ij} - u_i}{\sigma_{w,l}}\right\} \pi_{w,l}.
\end{aligned}$$

We complete the Bayesian model by specifying that $\boldsymbol{\beta}^c \sim$ Normal$(\boldsymbol{\mu}_{\beta^c}, \boldsymbol{\Sigma})$, $\sigma^2_{u,l} \sim$ IG(ν, τ). In the examples where the random effect mixture distribution had more than one component, the $\{\mu_{u,l}\}$ were independent normals with mean η and variance θ and $\{\pi_{u,l}\}$ were Dirichlet with shape parameters $\boldsymbol{\alpha}$. We note a small complication is that the mean of the random effect distribution is aliased with an intercept parameter. Therefore, throughout this section we assume that the intercept term is excluded and instead the random effect mean is estimated. A second complication could potentially arise when the random effect mixture distribution has more than one component, because of the non-identifiability of the parameters due to permutation invariance. In the examples we considered, however, imposing identifiability constraints (see Jasra et al., 2005) did not impact results.

The benefit of this model specification is that all of the full conditionals are common distributions and an elegant Gibbs sampler, which does not employ any Metropolis/Hastings steps, is available for exploring the posterior. We emphasize that the algorithm can be used for any link function whose associated distribution function can be represented as a mixture of normals. Moreover, this approach accommodates general modeling of the random effect distribution.

The full conditionals are as follows:

1. the full conditional for $D_{u,i}$ is discrete so that the probability $D_{u,i}$ takes value l is
$$\frac{\sigma_{u,l}^{-1}\exp\left\{-(u_i-\mu_{u,l})^2/2\sigma_{u,l}^2\right\}\pi_{u,k}}{\sum_k \sigma_{u,k}^{-1}\exp\left\{-(u_i-\mu_{u,k})^2/2\sigma_{u,k}^2\right\}\pi_{u,k}};$$

2. the full conditional for U_i is normal with mean
$$\left(\sum_j \sigma_{w,d_{w,ij}}^{-2}+\sigma_{u,d_{u,i}}^{-2}\right)^{-1}\left(\sum_j \frac{m_{ij}-\mu_{w,d_{w,ij}}+\Delta_{ij}}{\sigma_{w,d_{w,ij}}^2}+\frac{\mu_{d_{u,i}}}{\sigma_{u,d_{u,i}}^2}\right)$$
and variance
$$\left(\sum_j \sigma_{w,d_{w,ij}}^{-2}+\sigma_{u,d_{u,i}}^{-2}\right)^{-1};$$

3. the full conditional for $D_{w,ij}$ is discrete so that the probability $D_{w,ij}$ takes value l is
$$\frac{\sigma_{w,l}^{-1}\exp\{-(m_{ij}-\mu_{w,l}-u_i+\Delta_{ij})^2/2\sigma_{w,l}^2\}\pi_{w,l}}{\sum_k \sigma_{w,k}^{-1}\exp\{-(m_{ij}-\mu_{w,k}-u_i+\Delta_{ij})^2/2\sigma_{w,k}^2\}\pi_{w,k}};$$

4. the full conditional for M_{ij} is truncated normal with mean $\mu_{w,d_{w,ij}}+u_i-\Delta_{ij}$ and variance $\sigma_{w,d_{w,ij}}^2$ with $M_{ij}\leq 0$ when $y_{ij}=1$ and $M_{ij}>0$ when $y_{ij}=0$; that is, the distribution function is
$$\frac{\Phi\left\{(m_{ij}-\mu_{w,d_{w,ij}}-u_i+\Delta_{ij})/\sigma_{w,d_{w,ij}}\right\}}{\Phi\left\{(-\mu_{w,d_{w,ij}}-u_i+\Delta_{ij})/\sigma_{w,d_{w,ij}}\right\}}I(m_{ij}\leq 0)$$
when $y_{ij}=1$ and
$$\frac{\Phi\left\{(w_{ij}-\mu_{w,d_{w,ij}}-u_i+\Delta_{ij})/\sigma_{w,d_{w,ij}}\right\}-\Phi\left\{(-\mu_{w,d_{w,ij}}-u_i+\Delta_{ij})/\sigma_{w,d_{w,ij}}\right\}}{1-\Phi\left\{(-\mu_{w,d_{w,ij}}-u_i+\Delta_{ij})/\sigma_{w,d_{w,ij}}\right\}}I(m_{ij}\geq 0)$$
when $y_{ij}=0$;

5. the full conditional for $\boldsymbol{\beta}^c$ is multivariate normal with mean

$$\left(\boldsymbol{\Sigma}^{-1} + \mathbf{X}^t \mathbf{W}^{-1} \mathbf{X}\right)^{-1} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\beta^c} + \mathbf{X}^t \mathbf{W}^{-1} \boldsymbol{\eta}\right)$$

   where $\mathbf{X}$ is the design matrix, $\mathbf{W}$ is a diagonal matrix of the $\sigma^2_{w,d_{w,ij}}$ and $\boldsymbol{\eta}$ is a vector with elements $\mu_{w,d_{w,ij}} + u_i - m_{ij}$ and variance

$$\left(\boldsymbol{\Sigma}^{-1} + \mathbf{X}^t \mathbf{W}^{-1} \mathbf{X}\right)^{-1};$$

6. the full conditional for $\sigma^2_{u,l}$ is inverted gamma with shape parameter

$$\nu + \sum_i I(d_{u,i} = l)/2$$

   and rate parameter

$$\tau + \sum_i I(d_{u,i} = l)(u_i - \mu_{u,l})^2/2,$$

7. the full conditional for $\mu_{u,l}$ is normal with mean

$$\left(\sum_i I(d_{u,i} = l)\frac{1}{\sigma^2_{u,d_{u,i}}} + \frac{1}{\theta}\right)^{-1} \left(\sum_i I(d_{u,i} = l)\frac{u_i}{\sigma^2_{u,d_{u,i}}} + \frac{\eta}{\theta}\right)$$

   and variance

$$\left(\sum_i I(d_{u,i} = l)\frac{1}{\sigma^2_{u,d_{u,i}}} + \frac{1}{\theta}\right)^{-1},$$

8. the full conditional for the $\{\pi_{u,l}\}$ is Dirichlet with shape parameter

$$\boldsymbol{\alpha} + \sum \{I(d_{u,i} = 1), \ldots, I(d_{u,i} = L_u)\}^t.$$

We apply the Gibbs sampler to the four datasets employing diffuse priors with a single normal random intercept. Throughout we assume that F_w is the five component mixture of normals approximation to the logistic distribution. Figures 3 shows the estimated posterior distributions the the parameter of interest after $20,000$ simulations for each of the data sets employing 1, 2 and 3 mixture components for the random effect distribution. For each of the examples we set $\nu = 10^{-6}$, $\tau = 10^{-4}$, $\boldsymbol{\alpha} = (1, \ldots, 1)^t$, $\mu_{\boldsymbol{\beta}^c} = (0, \ldots, 0)^t$, Σ as a diagonal matrix with entries 10. Though the results are not reported, the impact of hyperparameter specification was investigated by varying the diffuseness of the priors.

The benefit of allowing for a small number of discrete mixture components for the random effect distribution is to protect against the impact of misspecification (see Agresti et al., 2004). This is particularly interesting for the Item Response data, since a three level random effect distribution makes practical sense in this situation. Specifically, it is likely that three populations, one opposed to abortion under any circumstance, one in favor of abortion rights regardless of the circumstance, and a more heterogeneous group dominate the random effect distribution.

Regardless, for the parameter of interest for these four data sets, misspecification of the random effect distribution does not appear to be impacting results. The estimated posterior densities appear to be the same regardless of the number of mixture components implemented (Figure 3).

5 Discussion

In this manuscript, we discussed the conceptual and computational benefits of using mixtures of normals as the conditional link distribution and random effect distribution for random intercept models for binary outcomes. The principal conceptual benefits are that this representation unifies many of the existing models for analyzing binary data. This includes models for estimating random effect distributions and link functions.

In addition, the mixture of normals representation makes the connection between the conditional and marginal link functions explicit. Like the probit-normal model, these mixture models represent a closed class with the conditional link, marginal link and random intercept distributions being all from the same family.

We also demonstrated some of the computational benefits of approximating links with mixtures of normals. First, it was demonstrated how they allow for simple post-hoc calculations of marginal effects from conditional estimates. Second, it was shown how they can greatly ease the computation of the “transfer function” for Heagerty and Zeger’s marginalized models. Finally, for a specific class of Bayesian models, the mixture of normals approximation leads to common distributions for all of the full conditionals, rendering the coding of a Gibbs sampler almost trivial.

The use of mixtures of normals could be exploited for further generalizations of the random intercept model. In particular, the extension to multivariate random effects, using mixtures of multivariate normals, is plausible. Furthermore, this mixture approach is potentially very useful for jointly modeling discrete and continuous outcomes. Finally, further work may also explore how the mixture approach facilitates description of the "bridge" random effect distribution as introduced by Wang and Louis (2003) and Wang and Louis (2004).

In closing we note that we have put all of the relevant code to reproduce all of the results, and the derivations of the Bayesian full conditionals at

```
http://www.biostat.jhsph.edu/~bcaffo/downloads.htm
```

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, second edition.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis*, 47(3):639–653.

Agresti, A. A., Booth, J., Hobert, J., and Caffo, B. S. (2000). Random effects modeling of categorical response data. *Sociological Methodology*, 30:27–80.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128.

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–678.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Caffo, B. and Griswold, M. (2005). A user-friendly tutorial on link-probit-normal models. Technical report, Johns Hopkins University, Department of Biostatistics.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit normals. *Biometrika*, 85(2):347–361.

Chib, S., Greenberg, E., and Chen, Y. (1998). MCMC methods for fitting and comparing multinomial response models. *Economics Working Paper Archive Econ WPA: Econometrics*. http://econwpa.wustl.edu:80/eps/em/papers/9802/9802001.pdf.

Demidenko, E. (2004). *Mixed Models Theory and Applications*. Wiley.

Erkanli, A., Stangl, D., and Mueller, P. (1993). A bayesian analysis of ordinal data using mixtures. *American Statistical Association Proceedings of the Section on Bayesian Statistical Science*, pages 51–56.

Follmann, D. A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84:295–300.

Geweke, J. and Keane, M. (1997). Mixture of normals probit model. Technical Report 237, Federal Reserve Bank of Minneapolis.

Griswold, M. (2005). *Complex Distributions, Hmmmm... Hiearchical Mixtures of Marginalized Multilevel Models*. PhD thesis, Johns Hopkins University.

Heagerty, P. J. and Zeger, S. L. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, 91:1024–1036.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1):1–26.

Imai, K. and van Dyk, D. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of econometrics*, 124:311–334.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–61.

Jones, B. and Kenward, M. (1987). Modelling binary data from a three ponit cross-over trial. *Statistics in Medicine*, 6:555–564.

Lang, J. B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, 89:625–632.

Lange, K. (1999). *Numerical Analysis for Statisticians*. Springer-Verlag.

Lee, Y. and Nelder, J. (2004). Conditional and marginal models: Another view. *Statistical Science*, 19(2):219–238.

Liang, K. and Hanfelt, J. (1994). On the use of the quasi-likelihood method in teratolgy experiments. *Biometrics*, 50:872–880.

Liang, K. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Magder, L. and Zeger, S. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association*, 91:1141–1151.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

McCulloch, R. and Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64:207–240.

Monahan, J. and Stefanski, L. (1992). Normal scale mixture approximations to $f^*(z)$ and computation of the logistic-normal integral. In Balakrishnan, editor, *Handbook of the Logistic Distribution*, pages 529–540. Marcel Dekker.

Natarajan, R., McCulloch, and Kiefer, N. (2000). A Monte Carlo EM method for estimating multinomial probit models. *Computational Statistics and Data Analysis*, 34:33–50.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall.

Wang, Z. and Louis, T. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*, 90(4):765–775.

Wang, Z. and Louis, T. (2004). Marginalized binary mixed-effects with covariate-dependent random effects and likelihood inference. *Biometrics*, 60(4):884–891.

Weil, C. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenisis or carcinogenisis. *Food and cosmetics toxicology*, 8:177–182.

Zeger, S., Liang, K., and Albert, P. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44:1049–1060.

A Tables

π_1	π_2	π_3	π_4	π_5
0.126840496	0.543170220	0.261711982	0.066181589	0.002066853
σ_1	σ_2	σ_3	σ_4	σ_5
2.8420536	1.8257138	1.1943048	1.0757749	0.5631853
μ_1	μ_2	μ_3	μ_4	μ_5
0	0	0	0	0

Table 1: Mixing probabilities, standard deviations and means of the mixture components for a mixture-of-normals approximation to the logistic distribution.

First	Second Survey	
Survey	Approve	Disapprove
Approve	794	150
Disapprove	86	570

Table 2: Prime minister approval rating. Source Agresti (2002).

	(number survived,number dead)							
Control	(13, 0)	(12, 0)	(9, 0)	(9, 0)	(8, 0)	(8, 0)	(12, 1)	(11, 1)
	(9, 1)	(9, 1)	(8, 1)	(11, 2)	(4, 1)	(5, 2)	(7, 3)	(7, 3)
Treatment	(12, 0)	(11, 0)	(10, 0)	(9, 0)	(10, 1)	(9, 1)	(9, 1)	(8, 1)
	(8, 1)	(4, 1)	(7, 2)	(4, 3)	(5, 5)	(3, 3)	(3, 7)	(0, 7)

Table 3: Teratology data. Numbers are (number survived, number dead) in each litter by treatment arm. For example, in the first control litter, all thirteen pups survived. Source Weil (1970).

Response		Treatment sequence	
Period 1	Period 2	Drug-Placebo	Placebo-Drug
Normal	Normal	22	18
Abnormal	Normal	0	4
Normal	Abnormal	6	2
Abnormal	Abnormal	6	9

Table 4: Crossover data, frequency of responses by treatment regimen. Source Jones and Kenward (1987).

	Sequence of Responses							
Gender	111	110	011	010	101	100	001	000
male	342	26	6	21	11	32	19	356
female	440	25	14	18	14	47	22	457

Table 5: Response to questions on abortion stratified by gender from Agresti (2002). A response of “1” was in favor of legalized abortion in a specific scenario while a response of “0” was not. The scenarios are i if the family has a very low income ii the woman is not married and does not want to marry the man iii for any reason.

	Parameter				
Data Set	$\hat{\sigma}_{u,1}$	$\hat{\beta}_0^c$	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_3^c$
Teratology	1.35 (0.33)	2.63 (0.48)	−1.08 (0.63)		
Approval	5.16 (0.35)	1.24 (0.19)	−0.56 (0.14)		
Crossover	4.94 (1.91)	2.22 (1.17)	1.86 (0.93)	−1.04 (0.82)	
Item Response	8.75 (0.54)	−0.61 (0.34)	−0.013 (0.49)	0.83 (0.16)	0.29 (0.16)

Table 6: Conditional Estimates (standard errors) for multilevel models from Section 4.2.

Data Set	Marginal Estimate ($\hat{\beta}_1^m$)		
Teratology	−0.86 (0.51)		
Approval	−0.16 (0.04)		
	Period 1	Period 2	
Crossover	0.59 (0.31)	0.58 (0.29)	
	Question 1	Question 2	Question 3
Item Response	−0.002 (0.03)	−0.002 (0.03)	−0.002 (0.03)

Table 7: Marginal logit estimates (standard errors) for the examples from Section 4.1.

	Parameter				
Data Set	$\hat{\sigma}_{u,1}$	$\hat{\beta}_0^m$	$\hat{\beta}_1^m$	$\hat{\beta}_2^m$	$\hat{\beta}_3^m$
Teratology	1.35 (0.33)	2.03 (0.39)	−0.87 (0.51)		
Approval	5.16 (0.35)	0.36 (0.05)	−0.16 (0.04)		
Crossover	4.94 (1.91)	0.68 (0.28)	0.58 (0.23)	−0.32 (0.23)	
Item Response	8.71 (0.54)	−0.048 (0.054)	0.004 (0.074)	0.150 (0.028)	0.053 (0.028)

Table 8: Estimates (standard errors) for marginalized multilevel models from Section 4.2.

B Figures

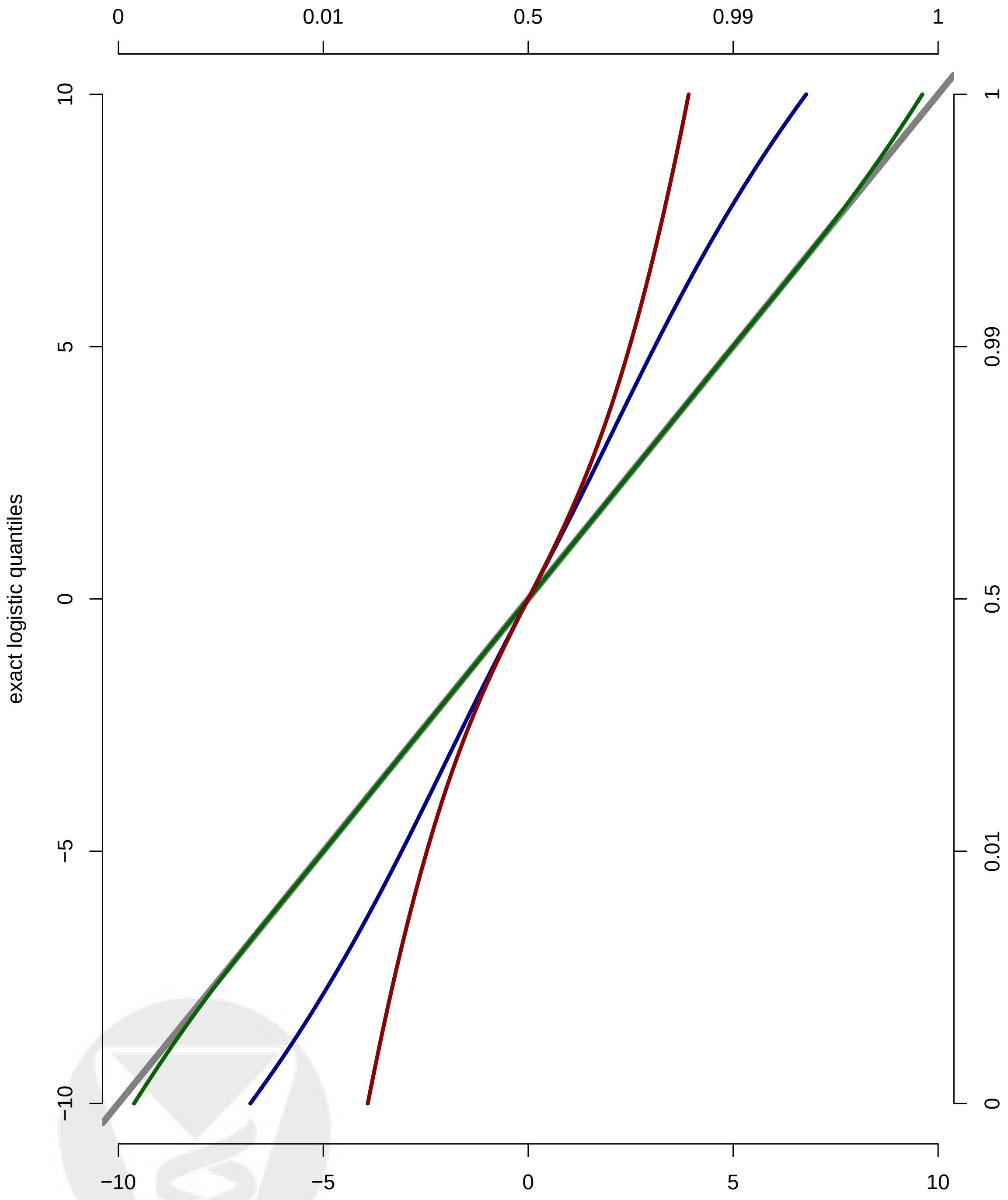

Figure 1: Quantile-quantile plot of the logistic distribution (vertical axis) by three approximations: the mixture of normals (green), the probit (red), the T (blue). A reference identity line is depicted in grey. The corresponding probability scale is given on the right and upper axes.

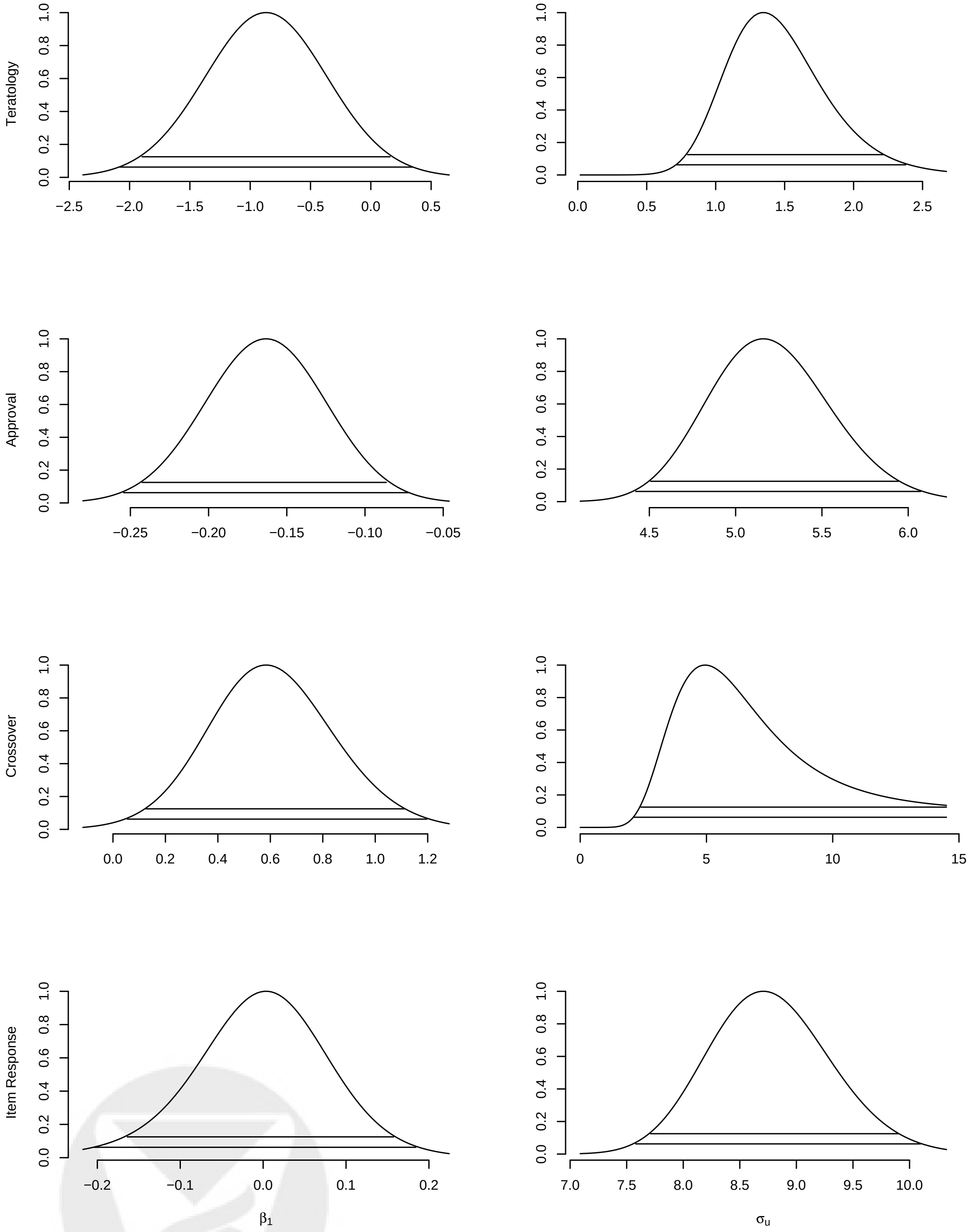

Figure 2: Profile likelihood plots with 1/8 and 1/16 reference lines, see (Royall, 1997) for β_1^m and $\sigma_{u,1}$ for the marginalized multilevel model from 4.2. The rows from top to bottom correspond to the Teratology, Approval, Crossover and Item Response data sets respectively.

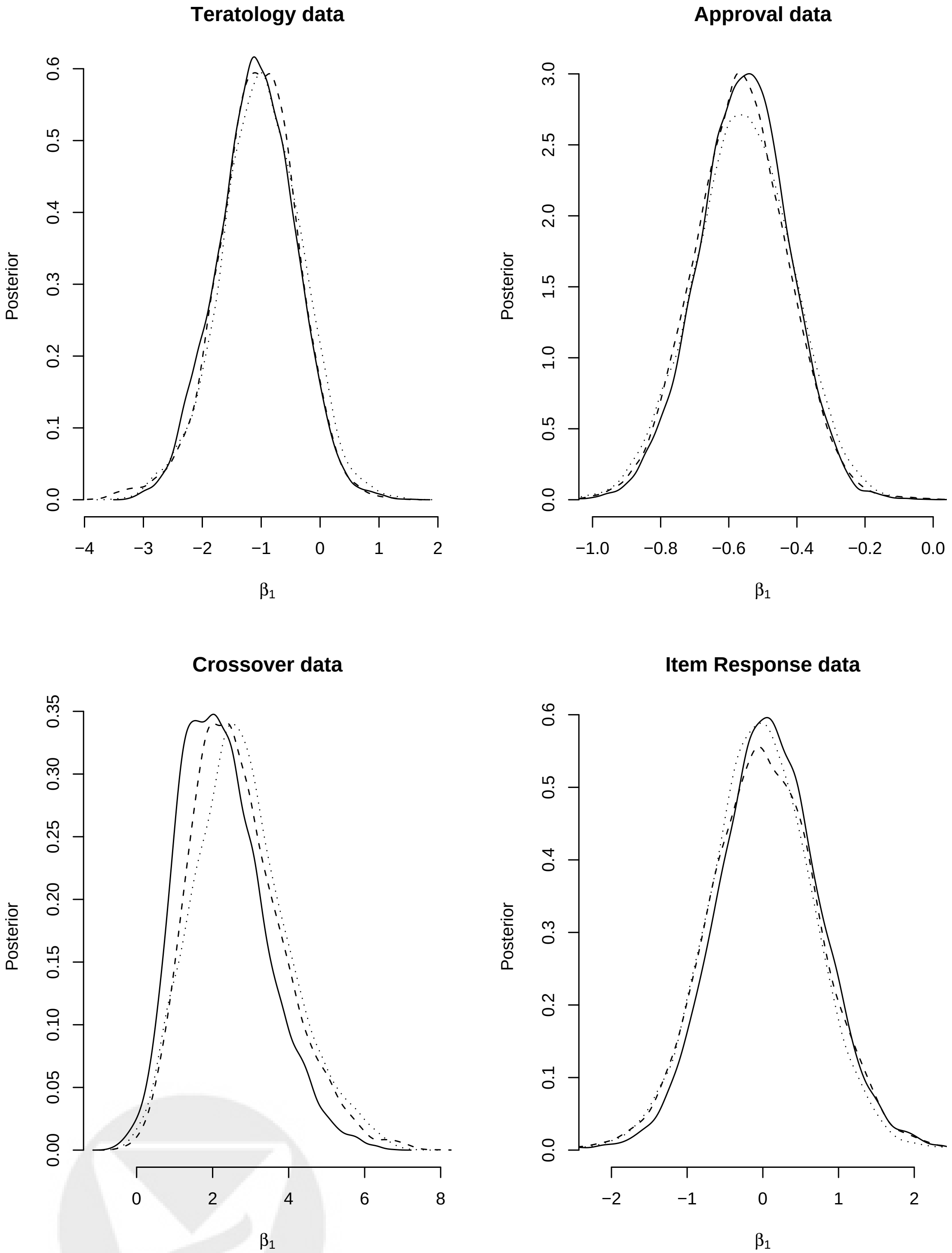

Figure 3: Estimated posterior densities using for the examples from Section 4.3 using one (solid), two (dashed) and three (dotted) component mixtures for the random effect distributions.

C Approximating link functions with mixtures of normals

In this section we give an estimation procedure for approximating a distribution with a mixture of normals. For a given number of mixture elements, we chose to minimize the Kullback/Liebler distance between the mixture approximation and the true density. That is, if g is the density associated with the link function of interest and f is the mixture approximation, we minimize $E_g[\log\{f(X)/g(X)\}]$. The algorithm was obtained as the limit of the standard EM algorithm for estimating normal mixture components as the number of observed data points goes to infinity.

Let $\pi_j^{(t)}$, $\sigma_j^{(t)}$ and $\mu_j^{(t)}$ be the current estimates,

$$\begin{aligned}
P_j^{(t)}(x) &= \frac{\pi_j^{(t)}\phi\{(x-\mu_j^{(t)})/\sigma_j^{(t)}\}/\sigma_j^{(t)}}{\sum_l \pi_l^{(t)}\phi\{(x-\mu_l^{(t)})/\sigma_l^{(t)}\}/\sigma_l^{(t)}} \\
\pi_j^{(t+1)} &= E_g\left[P_j^{(t)}(X)\right] \\
\mu_j^{(t+1)} &= E_g\left[XP_j^{(t)}(X)\right]/\pi_j^{(t+1)} \\
\sigma_j^{(t+1)} &= \left\{E_g\left[X^2P_j^{(t)}(X)\right]/\pi_j^{(t+1)} - \left(\mu_j^{(t+1)}\right)^2\right\}^{1/2}.
\end{aligned}$$

The expected values generally need to be evaluated numerically. In this manuscript we use Gauss/Hermite quadrature (see Lange, 1999).

D Obtaining standard error estimates of marginal parameters using the Multivariate Delta Method

In this section, we detail how to obtain the standard error estimate for $\hat\beta_1^m$ when there is one binary covariate. Note that $\hat\beta_1^m$ is a function of $\hat\beta_0^c$ and $\hat\beta_1^c$:

$$\hat\beta_1^m = g\begin{pmatrix}\beta_0^c \\ \beta_1^c\end{pmatrix} = \log\left\{\frac{F_q(\hat\beta_0^c+\hat\beta_1^c)}{1-F_q(\hat\beta_0^c+\hat\beta_1^c)}\right\} - \log\left\{\frac{F_q(\hat\beta_0^c)}{1-F_q(\hat\beta_0^c)}\right\},$$

with gradient

$$\nabla g^t = \begin{pmatrix} \frac{f_q(\hat\beta_0^c+\hat\beta_1^c)}{F_q(\hat\beta_0^c+\hat\beta_1^c)[1-F_q(\hat\beta_0^c+\hat\beta_1^c)]} - \frac{f_q(\hat\beta_0^c)}{F_q(\beta_0^c)[1-F_q(\beta_0^c)]} \\ \frac{f_q(\hat\beta_0^c+\hat\beta_1^c)}{F_q(\hat\beta_0^c+\hat\beta_1^c)[1-F_q(\hat\beta_0^c+\hat\beta_1^c)]} \end{pmatrix}$$

Since $(\hat{\beta}_0^c, \hat{\beta}_1^c)^t$ is normally distributed with covariance matrix Σ_β, we can apply the multivariate Delta Method to obtain a standard error estimate of β_1^m:

$$\text{SE}(\hat{\beta}_1^m) = \nabla g \; \Sigma_\beta \; \nabla g^t.$$