A Unified Approach for Analyzing Exchangeable Binary Data with Applications to Clinical and Developmental Toxicity Studies

Xin Dang\textsuperscript{a}, Hanxiang Peng\textsuperscript{a,\*}, Xueqin Wang\textsuperscript{b}
\textsuperscript{a}Department of Mathematics
University of Mississippi
University, MS 38677
\textsuperscript{\*} Corresponding author
Email: mmpeng@olemiss.edu
\textsuperscript{b}Department of Epidemiology and Public Health
Yale University School of Medicine
New Haven, CT 06520-8034,

March 26, 2008
Abstract

In this article, we present an approach to analyze data that are exchangeable (i.e. realizations of binomial mixtures). Our approach unifies the existing approaches and is practicable and computationally easy. Resulted from completely monotonic functions, we give a class of parsimonious binomial mixtures, including the incomplete Gamma-, Normal-, Poisson-, and Beta-Binomial, generalizing the Beta-Binomial. We show that the class is closed under convex linear combination, product, and two types of composites. We point out that the folded logistic link by Bowman and George(1995) is incorrect. The moments and Markov property are described. With such numerous mixtures, we can perform statistical inference on correlated binary data, in particular, over-dispersed data. We propose a regression procedure which generalizes the logistic regression. We give a stepwise forward procedure about how a possible best model from the class can be achieved. A small simulation is run to validate the inclusion of the Binomial. The proposed procedure is applied to the 2, 4, 5-T and the E2 data and compared favorably with the existing procedures.

Key words and phrases: Beta-binomial, Binomial mixture, Complete monotonicity, Exchangeability, Link, Logistic regression, Over-dispersion.
1 Introduction

Binomial distribution is widely used in modeling binary response data in many different areas of science. One key assumption in Binomial distribution is that the binary responses are assumed to be independent. The independence, however, may not be valid and non-independence usually leads to a variance that is greater than the nominal variance. This is known as over-dispersion (or extra-binomial variation). The perception of exchangeability, intensely studied over the past century, is meant to capture the notion of symmetry in a collection of random variables and is often used as an alternative to independence.

In the familial correlated data such as a developmental toxicity experiment in animal studies, the fetuses from a litter will respond more similarly to a stimulus than from different litters. Consider a litter of \(m\) fetuses with binary responses \(B_1, \ldots, B_m\) where \(B_i = 1\) or \(0\) denoting death or no death, tumor or no tumor. It cannot be assumed that these binary responses in a litter are independent. One way to simplify the analysis is to assume that they are exchangeable. In the risk analysis in Finance such as CreditRisk\(^+\) (Credit Suisse Financial Products, 1997), the default probability of a company is assumed to depend on a set of common economic factors; given these common factors, defaults of the individual obligors are conditionally independent. Consider a portfolio of \(m\) obligors with concentration on the binary outcomes \(B_1, \ldots, B_m\) of default and non-default of the obligors. One way in practice to formalize the notion of a homogeneous group is to assume that these binary outcomes are exchangeable. For a full description, see e.g. Frey and McNeil(2003). Specifically \(B_1, \ldots, B_m\) are exchangeable if for every \(\{0, 1\}\)-valued variables \(b_1, \ldots, b_m\), one has

\[
P(B_1 = b_1, \ldots, B_m = b_m) = P(B_{\pi_1} = b_1, \ldots, B_{\pi_m} = b_m),
\]

for every permutation \(\pi_1 \ldots \pi_m\) of \(1 \ldots m\). Let \(Y = B_1 + \ldots + B_m\) be the total number of “successes”. Under the exchangeability, the distribution of \(Y\) is given by

\[P(Y = y) = \binom{m}{y} \sum_{k=0}^{m-y} (-1)^k \binom{m-y}{k} \lambda_{y+k}, \quad y = 0, 1, \ldots, m, (1.1)\]

where \(\lambda_0 = 1, \lambda_k = P(B_1 = 1, \ldots, B_k = 1), 1 \leq k \leq m\) are the marginal probabilities. See e.g. Kendall(1967), George & Bowman(1995), or Chow & Teicher(1998). We write \(Y \sim EB(\lambda)(\lambda = \{\lambda_k\})\) and refer it to as the Exchangeable Binomial Distribution(EB).

Applications are not limited to the aforementioned financial and familial data. A wide variety can be found in other fields of science. For example, in social science, see e.g. Conaway(1990); in clinical trials, see Fanaroff et al(1994); and in botany, see Example 1.3, Collett(2003). A theoretical application is in the study of latent variable models, see Frey & McNeil(2003).
The literature on correlated binary data is extensive, see e.g. Collett(2003) or Joe(1997). The statistical inference based on the conditional analysis does not use any information about the distribution of the latent effects since it is lost in the conditioning. The EM algorithm method by Mislevy(1985) is conceptually simple, but the computations may be formidable because each E-step in the computation may require a numerical integration. The empirical Bayes approach by Stiratelli, Laird, and Ware(1984) may also be computationally difficult. The GEE approach (e.g. Liang, Qaqish and Zeger, 1992) only uses the first two moments while higher order of correlation is approximated by a “working matrix”. In this article, we present an approach to analyze correlated binary data by exploiting the notions of exchangeability and complete monotonicity. Our approach uses all the information and has mathematical simplicity and computational ease and unifies the existing approaches.

By virtue of the celebrated de Finnetti representation theorem, a sequence of exchangeable binary random variables is a mixture of binomial distributions, while the mixture can be characterized by a complete monotonic sequence. Using the relationship between exchangeability and complete monotonicity, Bowman and George(1995) initiated an approach on correlated binary data and gave the folded logistic link, though the folded logistic link is not a legitimate link as pointed out in Section 2, where we also give a modified version which is a legitimate link. Xu and Prorok(2003) showed that the MLEs of the population parameters have no closed form in general and pointed out that the MLE’s can only be calculated by numerical methods. They applied their results to a double-blind randomized clinical trial data in comparing two antibiotics, cefaclor and amoxicillin and performed several simulation studies. Kuk(2004) proposed the power link and established its complete monotonicity. They also applied their approach in clinical and developmental toxicity studies and achieved superior results to other exiting approaches such as GEE, Beta-Binomial, etc. It is to our thinking that further study on this approach is worthwhile, for example, providing enough completely monotone links that is of practicable use in performing real statistical inference; giving a procedure about finding a possible best model. In this article, we further their idea and attempt to present a unified theory and give a wide variety of complete monotonic links. We introduce a rich class of parsimonious binomial mixtures (completely monotonic links) including the incomplete Beta-, Gamma-, Normal-, and Poisson-Binomial. These mixtures are called incomplete because their probability mass functions contain incomplete special functions. For example, the incomplete Beta-Binomial include the incomplete Beta function. Since the Beta function is a special case of the incomplete Beta function, the popularly used Beta-Binomial is a special case of the incomplete Beta-Binomial. Further, the incomplete Beta-Binomial includes the Binomial as a special case, unlike the Beta-Binomial which does not include the Binomial as a special case in its parameter domain. Indeed all the
other introduced incomplete mixtures contain the Binomial as a special case, while existing binomial mixtures usually do not; for example, the random effects model for binary data by Conaway(1990), the logistic regression by nonparametric mixing by Follmann and Lambert(1989), and the mixtures resulted from random variables (see section 2). This inclusion of the Binomial is useful, for instance, regression based on these incomplete mixtures will boil down to the logistic regression if data is resulted from independent binary responses. We further demonstrate that the class is closed under convex linear combination, product and two types of composites. These properties enormously enlarge the class. The probability mass functions of the mixtures are quite simple both mathematically and computationally. The likelihoods can be computed without numerical integration and parameter estimation can be obtained by maximum likelihood using the R or SAS packages. In addition, we present the Markov property of the stochastic binomial process, moments and correlations of the mixtures. We give a stepwise procedure about how a possible best model can be arrived from such numerous mixtures. A simulation is conducted to validate the inclusion of the Binomial. The proposed procedure is applied to the 2, 4, 5-T and the E2 data. Comparisons are made with the Binomial, the Correlated Binomial, the Beta-Binomial, GEE’s, Williams’ logistic regression(1982), etc. Our results indicate that the proposed procedure improves these models and, in particular, the Gamma-Binomials outperformed the other models considered in regression on the 2, 4, 5-T data.

The proposed parsimonious mixtures unify the existing approaches. In addition to the aforesaid approaches, additional examples include the usual Binomial, the Correlated Binomial(Kupper and Haseman, 1978), the Additive Model(Altham, 1978), and the generalized logistic regression by nonparametric mixing(Follmann and Lambert, 1989). The model by Conaway(1990), interestingly, turns out to be a special case of our incomplete Gamma Power Binomial. Brooks(1997) explored finite mixture models for proportions. These models can be recovered from the linear combinations of the proposed class. From this standpoint, our work generalizes Brooks’ result to conclude that the products and two types of composites of the CM links(see section 2) are also closed, see section 4. Some additional related results are as follows. Stephanescu and Turnbull(2003) used the EM algorithm to model exchangeable binary data. Wood(1992) investigated binomial mixtures and finite exchangeability. Meester and MacKey(1994) introduced a parametric model for cluster correlated categorical data. Prentice explored the binary regression using extended Beta-Binomial in 1986 and the correlated binary regression in 1988. An early work on extra-binomial variation was Williams’ procedure in 1982.

The rest of the article is organized as follows. The parsimonious mixtures are introduced in section 2, followed by the methods about obtaining such mixtures and the Markov prop-
Section 3 gives various mixtures. Section 4 studies obtaining mixtures from existing ones. The moments and correlations are also given. Section 5 is devoted to regression. The relation with the logistic regression is discussed. We give a small simulation and apply our results to two real data in section 6. Technical details are collected in the Appendix.

2 CM-Binomials and Markov Property

In this section, we first introduce the parsimonious distributions, followed by the methods about obtaining such distributions. The Markov property is given in the end.

The marginal probabilities \( \lambda = \{ \lambda_i : i = 0, 1, ..., m \} (\lambda_0 = 1) \) in (1.1) form a completely monotone (CM) sequence as pointed out by George and Bowman (1995), namely,

\[
(-1)^k \Delta^k \lambda_i \geq 0, \quad i = 0, 1, ..., m, k + i \leq m, \quad (2.2)
\]

where \( \Delta \) is the difference operator defined by \( \Delta a_i = a_{i+1} - a_i \) and \( \Delta^2 = \Delta(\Delta) \) with \( \Delta^0 = I \) the identity operator for a sequence \( \{a_i\} \). A formula which is useful for verifying CM is

\[
(-1)^k \Delta^k a_r = \sum_{i=0}^{k} \binom{k}{i} (-1)^i a_{r+i}, \quad r = 0, 1, ..., k = 0, 1, 2, ...
\]

(2.3)

See Feller (1971, pp. 221) for a proof. Conversely, for \( \lambda \) satisfying (2.2), define

\[
f_{eb}(y|\lambda) = \frac{m!}{y!(m-y)!} \sum_{k=0}^{m-y} (-1)^k \binom{m-y}{k} \lambda_{y+k}, \quad y = 0, 1, 2, ..., m.
\]

(2.4)

Then \( f_{eb} \) well defines a probability distribution. This definition is based on finitely many complete monotone numbers \( \{\lambda_1, ..., \lambda_m\} \), while (1.1) is based on finitely many exchangeable events \( B_1, ..., B_m \). However, it follows from Kendall (1967) that the two definitions are equivalent. It is noteworthy to mention that slightly negative correlation of \( B_1, ..., B_m \) is allowed, while an infinite sequence \( B_1, B_2, ... \) only allows nonnegative correlation, see Kingman (1987).

The mean, the variance and the second order correlation are

\[
E(Y) = m\lambda_1, \quad \text{Var}(Y) = m(\lambda_1 - \lambda_2) + m^2(\lambda_2 - \lambda_1^2), \quad \phi = (\lambda_2 - \lambda_1^2)/\lambda_1(1 - \lambda_1).
\]

(2.5)

In terms of \( \phi \), we can rewrite \( \text{Var}(Y) = \sigma_{\text{ind}}^2 \{1 + (m-1)\phi\} \), where \( \sigma_{\text{ind}}^2 = m\lambda_1(1 - \lambda_1) \) is the variance of the Binomial. Since \( \phi \geq 0 \), it follows \( \text{Var}(Y) \geq \sigma_{\text{ind}}^2 \), manifesting that the EB and its parsimonious models can be used to study overdispersed data. Williams’ model is in fact an approximation to the EB model because his model variance is just this \( \text{Var}(Y) \).

Note that \( \text{EB}(\lambda) \) has \( m \) parameters and is the saturated model with parameter space \( \Lambda = \{ \lambda \in \mathbb{R}^m : \lambda \text{ satisfies } (2.2) \} \). Bowman and George (1995) gave the maximum likelihood
estimators of the parameters. Stefanescu and Turnbull (2003) proposed a procedure based on the EM algorithm to estimate the parameters. By mapping a lower $d$-dimensional subset $\Theta \subset \mathbb{R}^d$ into $\Lambda$ where $d \leq m$, we obtain a parsimonious model. Consider such a map from $\Theta$ into $\Lambda$ defined by $\lambda = h(\theta)$, where $\theta \in \Theta$ is the parameter space of the parsimonious model. Write $h = (h_1, ..., h_m)^\top$ so that $\lambda_j = h_j(\theta)$, $j = 1, ..., m$. Substitution of these representations in (1.1) results in a parsimonious model, which can be expressed as

$$f(y; \theta) = \left(\frac{m}{y}\right)^{m-y} \sum_{k=0}^{m-y} (-1)^k \binom{m-y}{k} h_{y+k}(\theta), \quad \theta \in \Theta, y = 0, 1, ..., m,$$  

where $h_0(\theta) = 1$ for every $\theta \in \Theta$. In order to ensure that the above expression is a legitimate probability mass function, these $h_1(\theta), ..., h_m(\theta)$ have to be complete monotone (2.2):

$$(-1)^k \Delta^k h_i(\theta) \geq 0, \quad \theta \in \Theta, \quad i = 0, 1, ..., m, i + k \leq m.$$  

Such $h$ is called a completely monotone link. Indeed we have the following theorem and the proof is given in the Appendix.

**Theorem 1** Suppose that $h_j(\theta), j = 1, ..., m$ satisfy (2.7) with $h_0(\theta) = 1$ for $\theta \in \Theta$. Then (2.6) is a probability mass function (pmf) and is thus a submodel of the full model (1.1).

Denote the above submodel by $\text{EB}(\theta, m; h)$ (or simply $\text{EB}(h)$). By the above theorem, every CM link $h$ gives a parsimonious model. We now discuss methods about obtaining CM links. One trivial CM link is $h_t = \theta^t$ for $\theta \in [0, 1]$, the independence link, corresponding to the independent binary responses (Binomial model). A direct verification can use (2.3). A sufficient condition is that $h_j(\theta) = \tilde{h}(j; \theta), j = 0, 1, ...$ with $\tilde{h}(0; \theta) = 1$ for some complete monotone (Widder, 1941) function $\tilde{h}(t; \theta)$ defined on $t \in [0, \infty)$ in the sense that $\tilde{h}$ has all order of derivatives $\tilde{h}^{(k)}(t; \theta)$ w.r.t. $t$ satisfying

$$(-1)^k \Delta^k \tilde{h}(t; \theta) \geq 0, \quad k = 0, 1, 2, ..., \quad \theta \in \Theta,$$  

where $t \in [0, \infty)$ (henceforth this is assumed except otherwise explicitly stated). George and Bowman (1995) also mentioned this sufficient condition. However, their proposed folded logistic link

$$h_t(\theta) = \frac{2}{1 + (t + 1)\theta}, \quad \theta \in [0, \infty),$$  

is not completely monotone on $\theta \geq 0$ so that it is not a legitimate link. More specifically it is not CM on $\theta > 1$ and it is CM on $0 \leq \theta \leq 1$. Here is a counter example. Let $\theta = 2$. Then it is easy to check that the folded logistic link does not satisfy (2.7) at $t = 6$ and $k = 4$-th differece. Indeed we have given a proof in the Appendix. A modified version which is
given below and referred as the *piecewise folded logistic (Piecewise Flogit) link* is completely monotone.

\[
h_t(\theta) = \begin{cases} 
  \frac{2}{1 + (1 + t)\theta}, & 0 \leq \theta \leq 1, \\
  \frac{1}{1 + t/2}, & \theta > 1.
\end{cases}
\] (2.10)

This link is relatively simple and our application with it in the real data sets shows it is a quite good link.

**Remark 1** For \( \bar{h}(\theta) \) satisfying (2.8), let \( h_t = \bar{h}_{t_0+t}/\bar{h}_{t_0} \) for some \( t_0 \geq 0 \). Then \( \{h_t : t \geq 0\} \) is CM with \( h_0 = 1 \). We say a completely monotone sequence is normalized at \( t_0 \).

It is well known that if a function is completely monotone on \([0, \infty)\), then it is a Laplace transform. Specifically we have the following theorem quoted from Feller(1971).

**Theorem 2** A function \( \bar{h} \) on \([0, \infty)\) is completely monotone with \( \bar{h}(0) = 1 \) if and only if it is a Laplace transform of probability distribution \( H \) on \([0, \infty)\), i.e.,

\[
\bar{h}(t) = \int_0^\infty \exp(-tx) \, dH(x), \quad t \in [0, \infty).
\] (2.11)

The distribution \( H \) is uniquely determined by \( \bar{h} \).

We may obtain CM links from the tabulated Laplace transforms and in particular from the moment generating functions (mgf’s). Specifically, suppose that the mgf \( M_H(t) \) of distribution \( H \) exists for all \( t \leq t_0 \) for some \( t_0 \geq 0 \), then

\[
\bar{h}(t) = M_H(-t), \quad t \in [0, \infty).
\] (2.12)

Using the relation between the Laplace transforms and characteristic functions (chf’s) we can also get CM links. Suppose that the chf of \( H \) is \( \varphi_H \). Then one easily shows that

\[
\bar{h}(t) = \varphi_H(it), \quad t \in [0, \infty), \quad i = \sqrt{-1}.
\] (2.13)

A substitution \( x = -\log p \) in (2.11) yields a useful representation

\[
\bar{h}(t) = \int_0^1 p^t \, dG(p), \quad t \in [0, \infty),
\] (2.14)

where \( G = H \circ \log \) is the induced probability measure. Thus we may get CM links from the moments of distribution \( G \). Suppose that \( G \) has \( j^{th} \) moment \( \mathbb{E}_G(p^j) \) for \( j = 0, 1, 2, ... \). Then

\[
\bar{h}(j) = \mathbb{E}_G(p^j), \quad j = 0, 1, 2, ...
\] (2.15)

If the mixing distribution \( G \) is parametric \( G = G_{\theta} \) with \( \theta \in \Theta \), then the resulting CM link \( \bar{h}(t; \theta) \) is parametric. Additional parameters in a distribution allows more modeling.
flexibility. Analogous to the method of *tolerance distribution* in obtaining link functions in the generalized linear models, a CM link containing additional parameters can be obtained from putting additional parameters in the limits of the above integration, i.e.,

\[
\bar{h}(t; \theta) = \int_{\theta_1}^{\theta_2} p' dG_{\varphi}(p), \quad \theta = (\theta_1, \theta_2, \varphi) \in [0, 1] \times \Theta.
\]  

(2.16)

Or equivalently,

\[
\bar{h}(t; \theta) = \int_{a}^{b} \exp(-tx) dH_{\varphi}(x), \quad \theta = (a, b, \varphi) \in [0, \infty) \times \Theta.
\]

(2.17)

It follows from Remark 1 that the following is CM with \(h_0(\theta_1, \theta_2, \varphi) = 1\),

\[
h_t(\theta_1, \theta_2, \varphi) = \frac{\bar{h}(t + t_0; \theta_1, \theta_2, \varphi)}{\bar{h}(t_0; \theta_1, \theta_2, \varphi)}, \quad \theta \in [0, 1] \times \Theta
\]

(2.18)

for \(t_0 \geq 0\). It should be noted out that the CM link is not defined when \(\theta_1 = \theta_2\). Nevertheless it is readily verified that

\[
\lim_{\theta_2 \to \theta_1} h_t(\theta_1, \theta_2, \varphi) = \theta_1 t_1.
\]

(2.19)

This corresponds to the *independence link* (Binomial model). Consequently we can extend the definition of \(h_t(\theta_1, \theta_2, \varphi)\) to admit the equality by defining it to be the limit, i.e.,

\[
h_t(\theta_1, \theta_2, \varphi) = \theta_1^t, \quad \theta_1 = \theta_2.
\]

(2.20)

Thus \(h_t(\theta)\) is well defined for all \(\theta \in [0, 1] \times \Theta\). Hereafter we shall adopt this definition without explicitly referring to it. It is interesting to observe that even though the positiveness of (2.16) or (2.17) requires \(\theta_1 < \theta_2\) the normalization renders the inequality constraint to unconstraint for the positiveness of the \(h_t(\theta_1, \theta_2, \varphi)\). The unconstraint is very useful both theoretically and computationally because otherwise the maximum likelihood estimation of the parameters is a cumbersome inequality-constrained maximization. Summing up our findings, we have the following theorem.

**Theorem 3** Suppose \(\{\bar{h}(t; \varphi) : t \geq 0\}\) satisfies (2.8). If (2.19) holds with \(h_t\) given in (2.18), then \(h(\theta) = \{h_j(\theta) : j = 0, 1, ..., m\}\) well defines a mixture model \(EB(\theta; h)\). Further, this mixture model recovers the Binomial model when \(\theta_1 = \theta_2\).

**CM Links from Random Variables.** Here we describe another method about obtaining CM links used in literature. If a random variable \(\xi\) has a range of \([0, 1]\), then its distribution is concentrated on \([0, 1]\) and can be taken as the mixing distribution to obtain a CM link. One convenient way to obtain such a random variable \(\xi\) is from the composite
\[ \xi = F(\eta) \] where \( F \) is a nonnegative measurable function with range \([0, 1]\) and \( \eta \) is a random variable taking values in the domain of \( F \). For example, equating \( F \) to the cdf \( \Phi \) and \( \eta \) to the standard normal random variable \( Z \) gives the \textit{probit-normal mixing} \( \xi = \Phi(\mu + \sigma Z) \), where \( \mu \) and \( \sigma \) are location and scale parameters. Setting \( \xi = 1/(1 + \exp(-\mu - \sigma Z)) \) yields the \textit{logit-normal mixing}. These two mixing distributions are widely used in practice, for example, in the risk analysis in Finance. It should point out that the resulted Bernoulli mixtures from this method usually do not include the Binomial as a special case, while the proposed incomplete mixtures in the next section do include. We shall not go into the details of this method.

**CM-Binomials as Markov Chains.** Let \( Y_m \sim \text{EB}(\theta, m; h) \) and consider the stochastic process \( \{Y_m : m = 1, 2, \ldots\} \). We give a theorem below with the proof delayed to the Appendix. Denote by \( G_{\theta}^* \) the probability measure concentrated on \([0, 1]\) determined by the CM sequence \( \{h_j(\theta) : j = 0, 1, \ldots\} \) and write \( u \sim G_{\theta}^* \) the random variable distributed like \( G_{\theta}^* \).

**Theorem 4** Suppose that \( h(\theta) : \theta \in [0, 1]^2 \times \Theta \) is CM. Then for every \( \theta \) the following hold.

1. There exist random variables \( \tilde{B}_1, \ldots, \tilde{B}_m \) which, conditional on latent random variable \( u \sim G_{\theta}^* \), are i.i.d. with the Bernoulli with probability \( h_1(\theta) \) of success such that \( Y_m = \tilde{B}_1 + \ldots + \tilde{B}_m \) with \( \mathbb{P}(Y_m = y | u \sim G_{\theta}^*) = \binom{m}{y} h_1^y(\theta)(1 - h_1(\theta))^{m-y}, \ y = 0, 1, \ldots, m \) and \( \mathbb{P}(Y_1 = y) = h_1^y(\theta)(1 - h_1(\theta))^{1-y} \) for \( y = 0, 1 \).

2. \( \{Y_m\} \) forms a non-homogenous Markov chain with transition probabilities

\[
\mathbb{P}(Y_{m+1} = y + 1 | Y_m = y, Y_{m-1} = y_{m-1}, \ldots, Y_1 = y_1) = \mathbb{P}(Y_{m+1} = y + 1 | Y_m = y) = 1 - \mathbb{P}(Y_{m+1} = y | Y_m = y) = \sum_{k=0}^{m-y} (-1)^k \binom{m-y}{k} h_{y+1+k}(\theta) / \sum_{k=0}^{m-y} (-1)^k \binom{m-y}{k} h_{y+k}(\theta)
\]

3. \( \text{Cov}(Y_i, Y_m) = lh(\theta)(1 - h_1(\theta)) + l(m - 1)(h_1(\theta) - h_1^2(\theta)) \) for \( m \geq l \).

4. \( \{Y_m - mh(\theta) : m = 1, 2, \ldots\} \) is a martingale.

5. \( Y_m/m \xrightarrow{p} Q_{\theta} \) as \( m \to \infty \).

**The Random Generator.** Based on Theorem 4 we have a recipe for generating exchangeable random variables \( B_1, \ldots, B_m \) and hence \( Y_m \) for a parsimonious \( \text{EB}(h) \). First generate \( \lambda \) from \( G_{\theta}^* \) determined by \( h_1(\theta) \). Then generate \( B_1, \ldots, B_m \) i.i.d. from Bernoulli with probability \( \lambda \) of success and \( Y_m = B_1 + \ldots + B_m \). Another approach to generate \( Y_m \) is the usual method of simulating discrete random variables.

### 3 Various CM-Binomials

In this section, we give various CM links using (2.11)-(2.15) and (2.16), (2.17). The CM links resulted from (2.16), (2.17) usually contain incomplete special functions and are thus
referred to as *incomplete CM links*, while the CM links resulted from (2.11)-(2.15) are the special cases of the incomplete CM links and referred to as *(complete) CM links*.

**Complete CM Links.** By Theorem 2, every distribution concentrated on \([0, \infty)\) gives a CM link. An explicit formula for a CM link may be found from either the Laplace transform, the mgf, or the characteristic function if it exists. For example, the *Gamma Binomial (Gamma-Bin)* link is obtained from (2.11) or (2.12) with \(H\) being a gamma distribution. As a special case of Gamma Binomial link, taking \(H\) in the Laplace transform to be exponential distribution, we obtain the *Michaelis-Menten (MM-Bin)* link which is used frequently in biology. An explicit formula for a CM link also can be found from the moment formula (2.15). The independent binomial distribution *(Ind-Bin)* is the special case of exchangeable distribution with \(G\) being a mass point distribution. Letting \(G\) be a beta distribution yields the *Beta Binomial (Beta-Bin)* link in Table 1. Besides, even though the normal distribution with mean zero is not concentrated on \([0, \infty)\), it is symmetric about zero so that it can be folded to be concentrated on \([0, \infty)\). The folded distribution from normal with mean zero and variance \(\sigma^2\) yields the *normal Binomial (Normal-Bin)* link given in Table 1. In this way, any symmetric distribution yields a CM link. More generally, any distribution can be used to produce a CM link. We list some of CM links in Table 1 which are referred to as XX-Binomial link, where XX is the name of the resulting distribution such as Beta, Gamma, Poisson, etc. Additional CM links can be analogously derived, for example, the Inverse-Gamma-Bin, the Negative-Bin-Bin, the Noncentral-\(\chi^2\)-Bin, the Discrete-Unif-Bin, the Positive-stable-Bin, the Weibull-Bin, etc.

**Incomplete CM Links.** We now derive incomplete CM links from formulas (2.16) and (2.17). Our first incomplete CM link is from the independence link \(h_t(p) = p^t, 0 < p < 1\). Consider for given \(t\) the area \(A_t(\theta)\) under the polynomial \(y = p^{t-1}\) over \([\theta_1, \theta_2]\), so that \(A_t(\theta) = \int_{\theta_1}^{\theta_2} p^{t-1} dp = (\theta_2^t - \theta_1^t)/t\) for \(t > 0\) with \(\theta = (\theta_1, \theta_2)\). The \(k\)th derivative w.r.t. \(t\) is

\[
A_t^{(k)}(\theta) = \int_{\theta_1}^{\theta_2} (\ln p)^k p^{t-1} dp, \quad t > 0, k = 1, 2, \ldots, \theta \in [0, 1]^2, \tag{3.21}
\]

Table 1: **Complete Monotone Links, \(N = \{0, 1, \ldots\}\).**

<table>
<thead>
<tr>
<th>Name</th>
<th>Link</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind-Bin</td>
<td>(\theta^t)</td>
<td>(\theta \in (0, 1))</td>
</tr>
<tr>
<td>MM-Bin</td>
<td>(\theta/(\theta + t))</td>
<td>(\theta \in (0, \infty))</td>
</tr>
<tr>
<td>Beta-Bin</td>
<td>(B(\theta_1 + t, \theta_2)/B(\theta_1, \theta_2))</td>
<td>(\theta \in (0, \infty)^2)</td>
</tr>
<tr>
<td>Gamma-Bin</td>
<td>((1 + \theta_2 t)^{-\theta_1})</td>
<td>(\theta \in (0, \infty)^2)</td>
</tr>
<tr>
<td>Poisson-Bin</td>
<td>(\exp(\theta(e^{-t} - 1)))</td>
<td>(\theta \in (0, \infty))</td>
</tr>
<tr>
<td>Binomial-Bin</td>
<td>((p \exp(-t) + 1 - p)^N)</td>
<td>((p, N) \in [0, 1] \times \mathbb{N})</td>
</tr>
<tr>
<td>Normal-Bin</td>
<td>(2 \exp((\sigma t)^2/2)(1 - \Phi(\sigma t)))</td>
<td>(\sigma^2 \in (0, \infty))</td>
</tr>
</tbody>
</table>
The range of the second order correlation

This is a nice polynomial in

\( t \) with the exponential

F

Alternatively, we can also normalize at

so \((-1)^k A^{(k)}(\theta) \geq 0\). This shows that \( A_t(\theta) \) is CM. Since \( \lim_{t \to 0} A_t(\theta) = \ln(\theta_2/\theta_1) \), normalizing at \( t = 0 \) and by Remark 1, we immediately obtain a CM link of two parameters, referred to as the Inc-A binomial link in Table 2. From (3.21) it also follows that \( A_t(\theta) \) is absolutely monotone (AM) for \( 1 < \theta_1 < \theta_2 \), i.e., \( d^k \phi(t)/dt^k \geq 0, k = 0, 1, ..., \text{for } 1 < \theta_1 < \theta_2 \). AM functions can be used to construct new CM links, see Theorem 5 in Section 4. Interestingly the Inc-A binomial link is a special case of the incomplete Beta-Binomial link below. Alternatively, we can also normalize at \( t = 1 \), then we have the CM link,

\[
1 \frac{1}{t + 1} (\theta_2^k + \theta_2^{k-1} \theta_1 + ... + \theta_2 \theta_1^{k-1} + \theta_1^k), \quad t \geq 0, \theta \in [0, 1]^2.
\]

This is a nice polynomial in \( \theta_1, \theta_2 \) while the previous link contains a nonlinear log function. The range of the second order correlation \( \phi \), however, reduces from \([0, 1/2]\) to \([0, 1/3]\), see more discussions about the range of correlation in the Section 4. Hereafter we shall normalize at \( t = 0 \).

Mixing with the Beta \( dG_{\theta}(\lambda) = \lambda^{\theta_3-1}(1-\lambda)^{\theta_1-1}/B(\theta_3, \theta_4) d\lambda \), \( \lambda \in (0, 1) \) for \( \theta_3 > 0, \theta_4 > 0 \) in the (2.16) we obtain the CM link, referred to as the incomplete Beta (Inc-Beta) binomial link in Table 2. Here \( B(x; \theta_1, \theta_2) = \int_0^x u^{\theta_1-1}(1-u)^{\theta_2-1} du \) is the incomplete Beta function and \( F_{\text{beta}} \) is the distribution function of the Beta distribution. The case \( (\theta_1, \theta_2) = (0, 1) \) recovers the (complete) Beta-Bin link, which results in the Beta-Binomial distribution. Mixing with the exponential \( dH(x) = \theta_3 \exp(-\theta_3 x) dx \) for \( \theta_3 > 0 \) in (2.17), we obtain the Inc-MM binomial link. The case \( (\theta_1, \theta_2) = (0, \infty) \) corresponds to the Michaelis-Menten equation used frequently in biology and recovers the MM-Bin link. Mixing with the Gamma \( dH(x) = (x/\theta_4)^{\theta_3-1} \exp(-(x/\theta_4))/\theta_4 \Gamma(\theta_3) dx \), \( x \geq 0, \theta_3 \in (0, \infty) \), we obtain the incomplete Gamma (Inc-Gamma) binomial link. Here \( \Gamma(x, \theta) = \int_x^\infty z^{\theta-1} \exp(-z) dz \) is the incomplete Gamma function. The case \( (\theta_1, \theta_2) = (0, \infty) \) recovers the Gamma binomial link. Analogously, we can obtain other incomplete CM links.

<table>
<thead>
<tr>
<th>Name</th>
<th>Link</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc-A-Bin</td>
<td>( \frac{\theta_3 - \theta_1}{t \log \theta_2 + \log \theta_1} )</td>
<td>( \theta \in [0, 1]^2 )</td>
</tr>
<tr>
<td>Inc-MM-Bin</td>
<td>( \theta_1 \frac{\theta_2^{\theta_3-1} - \theta_1^{\theta_3-1}}{\theta_2^{\theta_3} - \theta_1^{\theta_3}} )</td>
<td>( \theta \in (0, \infty)^3 )</td>
</tr>
<tr>
<td>Inc-Beta-Bin</td>
<td>( \frac{\beta_2 + \theta_1}{\beta_2 + \theta_1 - \beta_1 + \theta_1} )</td>
<td>( \theta \in [0, 1]^2 \times (0, \infty)^2 )</td>
</tr>
<tr>
<td>Inc-Gamma-Bin</td>
<td>( \frac{1}{(1+\theta_4)^3} \left( \frac{1}{(\theta_2 \theta_3 - 1)(\theta_1 \theta_3)} \right) )</td>
<td>( \theta \in (0, \infty)^4 )</td>
</tr>
<tr>
<td>Inc-Poisson-Bin</td>
<td>( \exp\left{ (1 - 1) \theta \right} \frac{F_{\text{pois}}(N_2, e^{-1}) - F_{\text{pois}}(N_1, e^{-1})}{F_{\text{pois}}(N_2, \theta) - F_{\text{pois}}(N_1, \theta)} )</td>
<td>( \theta \in \mathbb{N}^2 \times (0, \infty) )</td>
</tr>
<tr>
<td>Inc-Normal-Bin</td>
<td>( \exp\left( t^2 \sigma^2 / 2 \right) \frac{\Phi(\theta_2 / \sigma + \sigma) - \Phi(\theta_1 / \sigma)}{\Phi(\theta_2 / \sigma) - \Phi(\theta_1 / \sigma)} )</td>
<td>( \theta \in (0, \infty)^3 )</td>
</tr>
</tbody>
</table>
4 Linear combination, Product and Composites

In this section, we demonstrate that the class of parsimonious mixtures is closed under convex linear combination, product and two types of composites; as an application we show the complete monotonicity of the links in Tables 1 and 2. We now give the useful theorem, of which the first two results are the quick consequences of the Criteria 1, 2 of Feller(1971, pp. 441), and the third is from Widder(1941).

Theorem 5 (LP) If \( \varphi, \psi \) are CM, then the convex linear combination \( \alpha \varphi + (1 - \alpha) \psi \) for \( \alpha \in [0, 1] \) and the product \( \varphi \psi \) are also CM.

(C.I) If \( h \) is CM and \( \psi \) is a positive function with a CM derivative, then \( h(\psi) \) is CM.

(C.II) If \( h \) is CM and \( \varphi \) is absolutely monotone, then the composite \( \varphi(h) \) is CM.

The following remark gives several useful functions which satisfy Theorem 5 (C).

Remark 2 (1) The logarithm function \( \psi_1(t; \theta) = \theta \log(1 + t) \) with \( \theta > 0 \), the power function \( \psi_2(t; \nu) = t^\nu \) with \( 0 \leq \nu \leq 1 \), and \( \psi_3(t) = 1 - \exp(-t) \) are positive with CM derivatives.

(2) The positive polynomial \( \varphi_1(t; \theta) = \theta_1 + \theta_2 t + \ldots + \theta_k t^k \) for \( \theta = (\theta_1, \ldots, \theta_k)^\top \in [0, \infty)^k \) and the exponential function \( \varphi_2(t; \theta) = \theta^t \) with \( \theta > 1 \) are absolutely monotone.

Applying Theorem 5 (C.I) and (C.II) to the existing CM links, we can obtain numerous new CM links. For example, the composite of \( h_\nu(s) = \exp(-s) \) and \( \psi_2(t; \theta) = t^\theta \) yields a CM link, which is called the positive stable link, see Joe(1997). The composite of logarithm function \( \ln(1 + t) \) on Gamma-Bin link results in the Gamma-Log-Bin link. Moreover, additional parameters can be introduced via the two types of composites. The composite of the power function \( t^\nu \) and the independence binomial link \( h(t) = p^t \) gives the p-power link of Kuk (2004). For each link in Tables 1 and 2, we have the following remark.

Remark 3 If \( t \) in Tables 1 and 2 is replaced with logarithm \( \ln(1 + t) \) or power \( t^\nu \), then the resulting links are still CM and referred as XX-Log-Bin links and XX-power-Bin links respectively, where XX is the corresponding name in the tables such as MM, Gamma, Poisson, etc.

The aforementioned multivariate parametric parsimonious mixtures offer great flexibility in fitting data. Unlike the MM-Binomial, the Gamma-Binomial and the Beta-Binomial, the probability curves of the parsimonious distributions resulted from the convex linear combinations, products and composites allow very flexible shapes, see Fig. 1 and 2.

Moment and Correlation. There are convenient formulas for the means and variances of the parsimonious mixtures.

Theorem 6 Suppose that \( g, h \) are two CM links. Then we have the following.

(1) The mgf of \( \text{EB}(h) \) resulted from \( h \) is \( M_h(t) = \sum_{k=0}^n (-1)^k \binom{n}{k} h_k(\theta)(1 - e^t)^k \). Hence the mean is \( \mu_{1,h} = mh_1(\theta) \) and the variance is \( \sigma_h^2 = m(h_1(\theta) - h_2(\theta)) + m^2(h_2(\theta) - h_1^2(\theta)) \).

(2) The mgf of \( \text{EB}(\alpha_1 g + \alpha_2 h) \) is \( M(t) = \alpha_1 M_g(t) + \alpha_2 M_h(t) \) where \( \alpha_1, \alpha_2 \in [0, 1] \), \( \alpha_1 + \alpha_2 = 1 \). Hence the first and second moments are \( \mu_1 = \alpha_1 \mu_{1,g} + \alpha_2 \mu_{1,h} \), \( \mu_2 = \alpha_1 \mu_{2,g} + \alpha_2 \mu_{2,h} \).
The Range of Correlation. The kth order correlation $\phi_k$ of the exchangeable binary random variables $B_1, ..., B_m$ are an important quantity which was given in Theorem 2.2 of George and Bowman(1995). The second order correlation of the parsimonious model $EB(\theta; h)$ is

$$\phi(\theta) = \phi_2(\theta) = (h_2(\theta) - h_1^2(\theta))/(h_1(\theta) - h_1^2(\theta)), \quad \theta \in \Theta.$$  

(4.23)

This is useful, for example, it can be used to test exchangeability, i.e., $\phi(\theta) = 0$ if and only if the infinite sequence $B_1, B_2, ...$ is independent, see the simulation in section 6. It is easily calculated that the range of the correlation $\phi(\theta)$ for the Kuk’s power family, the MM-power-Bin, Gamma-power-Binomial, and A-power Binomial are $[0, 1]$, which is the full range of the correlation for an infinite exchangeable binary sequence. The range of correlation $\phi(\theta_1, \theta_2, 1)$ for the Gamma-power Binomial is only $[0, 1/2]$. This corresponds to the Gamma link normalized at $t_0 = 0$ in Remark 1. If the Gamma link normalized at $t_0 = 1$ in Remark 1, then the range of the correlation $\phi(\theta_1, \theta_2, 1)$ reduces to $[0, 1/3]$.

5 Regression and Parameter Estimation

Suppose now that given litter size $M$, we have observation $Y$ from parsimonious mixture $EB(\theta, M; h)$, associated with a covariate $X$. Typically $X$ is a vector of such as dose, weight, etc. Let $(Y_i, M_i, X_i), i = 1, ..., n$ be observations of $(Y, M, X)$. Write $\theta = (\theta, \vartheta)$, where $\theta \in \mathbb{R}$ is a parameter of interest, while $\vartheta$ is treated as a nuisance parameter. Allowing $\theta$ to depend on the linear systematic part $\eta = \beta^\top X$ gives $\theta = \eta$, where $\beta \in \mathbb{B}$ is a regression parameter for some nonempty open $\mathbb{B}$. A common but equivalent expression is

$$\lambda_j = h_j(\beta^\top X; \theta), \quad j = 0, 1, 2, ...$$

These equations suggests two extensions of the proposed mixture modeling from the generalized linear models, namely, the binomial mixtures generalize the Binomial distribution and the completely monotone links generalize the usual links.

Relation to the Logistic Regression. By (2.19), the Binomial model is a special case of the proposed incomplete mixture models, so that $\theta_1 = \theta_2$ yields the logistic regression. This property is very useful in modeling. We can always fit a dataset with the proposed parsimonious binomial mixture. If the data is indeed resulted from independent binary responses (i.e., binomial data), then the estimators of the parameters $\hat{\theta}_1, \hat{\theta}_2$ of $\theta_1, \theta_2$ should be very close $\hat{\theta}_1 \approx \hat{\theta}_2$. Our simulations and applications in section 6 validate these results.

The Exponential Technique. To ensure the complete monotonicity of the link $h(\theta)$, $\theta$ hence the $\beta$ in $\eta = \beta^\top X$ is inequality constrained. The constrained maximization can be circumvented to an unconstrained one by replacing $\eta$ with $\exp(\eta)$. The reason to do so and still keep the resulting link CM is application of Remark 2 in Theorem 5.

Model Selection. The Laplace transforms (including moment generating functions, characteristic functions) together with the convex linear combinations, the products and the composites provide a vast class of parsimonious parametric binomial mixtures, furnishing various choices and great flexibility for statistical inference such as model fitting and regression. From a practitioner’s point of view, it is always helpful to have some guidelines about
Table 3: Estimated Death Probabilities and Intra-litter Correlations. The upper
and middle table are from Brooks et al.(1997), Brooks (2001) and Kuk (2005). The Stan-
dard deviations are included when they are available. The lower table is for the proposed
procedures. The highlighted correspond to the optimal models.

<table>
<thead>
<tr>
<th>Models (number of parameters)</th>
<th>ˆp (s.d.)</th>
<th>ˆφ (s.d.)</th>
<th>−2logL</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial (1)</td>
<td>0.113 (.006)</td>
<td>0.000</td>
<td>765.6</td>
<td>767.6</td>
<td>767.9</td>
</tr>
<tr>
<td>Correlated Binomial (2)</td>
<td>0.131 (.010)</td>
<td>0.073 (.012)</td>
<td>720.5</td>
<td>724.5</td>
<td>731.2</td>
</tr>
<tr>
<td>Beta-Binomial (2)</td>
<td>0.112 (.009)</td>
<td>0.101 (.017)</td>
<td>689.8</td>
<td>693.8</td>
<td>700.5</td>
</tr>
<tr>
<td>Two Binomial (3)</td>
<td>0.111</td>
<td>0.114</td>
<td>682.4</td>
<td>688.4</td>
<td>698.5</td>
</tr>
<tr>
<td>Three Binomial (5)</td>
<td>0.121</td>
<td>0.101</td>
<td>679.7</td>
<td>689.7</td>
<td>706.5</td>
</tr>
<tr>
<td>Beta-Binomial with Binomial (4)</td>
<td>0.135</td>
<td>0.189</td>
<td>680.2</td>
<td>688.2</td>
<td>701.6</td>
</tr>
<tr>
<td>Kuk’s q-power (2)</td>
<td>0.119</td>
<td>0.209</td>
<td>687.1</td>
<td>691.1</td>
<td>697.8</td>
</tr>
<tr>
<td>Kuk’s p-power (2)</td>
<td>0.109</td>
<td>0.080</td>
<td>698.8</td>
<td>702.8</td>
<td>709.5</td>
</tr>
<tr>
<td>Gamma-Binomial with θ_2 = 1 (1)</td>
<td>0.118 (.0015)</td>
<td>0.191 (.0009)</td>
<td>697.8</td>
<td>699.8</td>
<td>700.1</td>
</tr>
<tr>
<td>Gamma-Binomial (2)</td>
<td>0.110 (.0087)</td>
<td>0.093 (.0181)</td>
<td>679.9</td>
<td>683.9</td>
<td><strong>684.5</strong></td>
</tr>
<tr>
<td>Incomplete Gamma-Binomial (4)</td>
<td>0.109 (.0118)</td>
<td>0.101 (.0301)</td>
<td><strong>675.2</strong></td>
<td><strong>683.2</strong></td>
<td>696.6</td>
</tr>
<tr>
<td>Piecewise-Flogit (1)</td>
<td>0.111 (.0154)</td>
<td>0.112 (.0362)</td>
<td>680.8</td>
<td><strong>682.8</strong></td>
<td><strong>683.1</strong></td>
</tr>
<tr>
<td>Piecewise-Flogit Power (2)</td>
<td>0.111 (.0175)</td>
<td>0.112 (.0411)</td>
<td>680.8</td>
<td>684.8</td>
<td>691.5</td>
</tr>
<tr>
<td>Incomplete Beta-Binomial (4)</td>
<td>0.110 (.0055)</td>
<td>0.102 (.0384)</td>
<td>676.4</td>
<td>684.4</td>
<td>697.8</td>
</tr>
<tr>
<td>Incomplete A-Binomial (2)</td>
<td>0.115 (.0085)</td>
<td>0.096 (.0172)</td>
<td>681.6</td>
<td>685.6</td>
<td>692.3</td>
</tr>
</tbody>
</table>

obtaining a possible best model from this vast class. Here we suggest a procedure, referred to
as **Forward Model Selection**. We borrow the idea from **forward variable selection** in multiple
linear regression.

We start from simple one-parameter links such as MM link, MM-log link, one-parameter
gamma links, one-parameter gamma-log-bin links or poisson-log link, etc. According to cer-
tain criteria such as AIC, BIC or Chi-square statistics, the best two or three models are
selected. Next, consider larger two-parameter models such that they include the selected
models as sub-models. This step considers models including the linear combination, power
composite, and nonnegative polynomial composite of previous ones. We can test the differ-
ence between a larger model and reduced model by the asymptotic likelihood ratio test. If
the hypothesis is rejected, add more parameters into models and test the significance of addi-
tional parameters until the hypothesis is accepted or the model reaches four parameters since
the model with more than four parameters is of no interest due to its complexity. Through
this procedure, choosing from a long list of model catalog, we may obtain the possible op-
timal models with considerably less computing time than required with the procedure that
tries all possible models. The whole procedure is easy to implement and R code is available
upon the request.
Table 4: The Estimates Under Various Models. Link1=Gamma-Bin with $\theta_2 = 1$, Link2= Gamma-Log with $\theta_2 = 1$, Link3=Gamma-Bin, Link4=Gamma-Log

<table>
<thead>
<tr>
<th>Model (Number of Pars)</th>
<th>$\alpha$(S.D.)</th>
<th>$\beta$(S.D.)</th>
<th>-Loglike</th>
<th>$\chi^2$</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial (2)</td>
<td>-3.235(.113)</td>
<td>5.430(.217)</td>
<td>1147.5</td>
<td>1514.2</td>
<td>2298.9</td>
<td>2306.8</td>
</tr>
<tr>
<td>Beta-Bin (4)</td>
<td>0.433(.081)</td>
<td>3.788(.194)</td>
<td>732.4</td>
<td>336.43</td>
<td>1472.8</td>
<td>1488.5</td>
</tr>
<tr>
<td>GEE(Logit, Ex) (3)</td>
<td>-3.323(.205)</td>
<td>5.580(.438)</td>
<td>411.4</td>
<td>1472.8</td>
<td>1488.5</td>
<td></td>
</tr>
<tr>
<td>Williams’ (3)</td>
<td>-3.237(.231)</td>
<td>5.587(.442)</td>
<td>370.87</td>
<td>1472.8</td>
<td>1488.5</td>
<td></td>
</tr>
<tr>
<td>Kuk’s q-power (3)</td>
<td>0.884(.071)</td>
<td>-0.525(.115)</td>
<td>729.8</td>
<td>298.95</td>
<td>1465.6</td>
<td>1471.3</td>
</tr>
<tr>
<td>Link1 (2)</td>
<td>3.838(.174)</td>
<td>-4.712(.265)</td>
<td>730.4</td>
<td>372.22</td>
<td>1464.8</td>
<td>1472.7</td>
</tr>
<tr>
<td>Link2 (2)</td>
<td>4.633(.259)</td>
<td>-5.447(.411)</td>
<td>736.1</td>
<td>247.64</td>
<td>1468.8</td>
<td>1484.1</td>
</tr>
<tr>
<td>Link3 (3)</td>
<td>3.340(.161)</td>
<td>-4.082(.243)</td>
<td>729.5</td>
<td>343.81</td>
<td>1464.9</td>
<td>1476.7</td>
</tr>
<tr>
<td>Link4 (3)</td>
<td>15.03(.759)</td>
<td>-18.16(.116)</td>
<td>725.0</td>
<td>312.63</td>
<td>1456.1</td>
<td>1467.9</td>
</tr>
<tr>
<td>aLink1+(1-a)Link2 (3)</td>
<td>4.116(.203)</td>
<td>-4.975(.315)</td>
<td>722.9</td>
<td>310.53</td>
<td>1451.8</td>
<td>1463.6</td>
</tr>
<tr>
<td>aLink3+(1-a)Link4 (4)</td>
<td>7.884(.640)</td>
<td>-9.571(.878)</td>
<td>719.3</td>
<td>338.84</td>
<td>1446.6</td>
<td>1462.3</td>
</tr>
<tr>
<td>Link1*Link2 (2)</td>
<td>2.119(.105)</td>
<td>-2.568(.161)</td>
<td>729.7</td>
<td>309.66</td>
<td>1463.4</td>
<td>1471.3</td>
</tr>
<tr>
<td>Link3*Link4 (3)</td>
<td>2.404(.062)</td>
<td>-2.927(.292)</td>
<td>729.2</td>
<td>326.75</td>
<td>1464.5</td>
<td>1470.2</td>
</tr>
<tr>
<td>Piecewise-Flogit (2)</td>
<td>6.792(.265)</td>
<td>-8.441(.393)</td>
<td>750.9</td>
<td>510.81</td>
<td>1505.8</td>
<td>1513.7</td>
</tr>
</tbody>
</table>

6 Simulations and Applications

In this section, we present a small simulation and a quite comprehensive study of the applications of the proposed Binomial mixture models to two real data sets. One data was fitted with various mixture models and comparisons were made with exiting results. On the other data set we conducted a regression analysis based upon a variety of the mixture models. Again comparisons were carried out with exiting models. Parameter estimation is based upon the criterion of the usual maximum likelihood. The MLE’s of the parameters are the solutions of the score equations and the numerical solutions are found via the Newton-Raphson iteration. Under the usual regularity assumptions, one obtains the asymptotic normality of the MLE’s, in particular, the asymptotic covariance matrix which is used to calculate the standard deviations of the MLE’s. These results can be found in an ordinary textbook and we shall omit the details.

A Small Simulation. We conducted a small simulation to validate that the incomplete Binomial mixtures include the Binomial as a special case, namely, if the observations are indeed resulted from the independent binary responses (i.e., the Binomial) then the incomplete Binomial mixture will boil down to the Binomial through the equality of two parameters in the mixture, $\theta_1 = \theta_2$. We generated a sample of size 400 from Binomial $\text{Bin}(N, 0.1)$, where $N$ is a random variable from discrete Uniform distribution $U[5, 15]$. Here we consider $p = 0.1$ close to zero, because estimation in this case may become difficult. We obtained the MLE’s of the parameters in the incomplete A-, Beta- and Gamma-Binomial and the estimates are $(\hat{\theta}_1, \hat{\theta}_2) = (0.0940, 0.0940), (0.0940, 0.0940), \text{ and } (0.0941, 0.0944)$, respectively, which exhibit $\hat{\theta}_1 \approx \hat{\theta}_2$. Moreover, they all have the same log likelihood $-488.69$, the same estimate 0.094 of the probability (i.e. 0.1) from $h_1(\hat{\theta}_n)$, and the same estimated second correlation $\hat{\phi} \approx 0$ which, again, indicates that the data are from the Binomial.
Table 5: **Expected Number of Malformations.** link1, link2, link3=Gamma-Bin, Gamma-Log, Gamma-Log link with θ₁ = 1, θ₂ = 1, θ₁ = 1 respectively. \( \hat{a}_{mle} \) is the MLE of \( a \) treated as parameter. \( d_i \) is the difference between the expected and observed number of malformations in the \( i^{th} \) dose group.

| Models                      | Dose level | \( \sum |d_i| \) | \( \sqrt{\sum d_i^2} \) |
|-----------------------------|------------|----------------|--------------------------|
| Observed                    | 59 124 338 390 372 | | |
| Binomial(Logit)             | 29 159 351 408 336 | 132 | 62.56 |
| Beta-Binomial               | 80 189 323 359 387 | 147 | 77.96 |
| GEE(Logit,ex)               | 27 154 346 408 339 | 121 | 58.32 |
| Williams's \(^\dagger\)     | 27 153 345 408 339 | 119 | 57.68 |
| Kuk's q-power               | 90 260 396 347 246 | 394 | 201.36 |
| Link1                       | 54 177 342 400 390 | 90  | 57.22 |
| Link2                       | 68 196 356 393 361 | 113 | 75.62 |
| Link3                       | 56 180 344 400 387 | 90  | 59.21 |
| Link4                       | 58 183 345 396 378 | 79  | 60.02 |
| \( a \)Link1+(1-\( a \))Link2 | 59 182 342 393 377 | 70  | 58.43 |
| \( a \)Link3+(1-\( a \))Link4 | 58 179 339 391 379 | 65 55.47 |
| Link1*Link2                 | 59 184 347 398 381 | 86  | 61.85 |
| Link3*Link4                 | 57 182 345 399 384 | 88  | 60.35 |
| Piecewise-Flogit            | 50 169 333 400 405 | 102 | 57.61 |

**Fitting E2 Data.** The E2 data (Brooks et al., 1997) records fetal control mortality in mouse litters. There are total 211 litters with litter sizes varying from small (as to 3) to large (as to 19) with mean litter size 12.9 and standard deviation 2.68. The proportion of dead fetuses is 0.110. We start from one-parameter models. Among them, one-parameter Gamma-Binomial link with \( \theta_2 = 1 \) and Piecewise-Flogit link are the best two links according to the highest maximum likelihood. Those two models fit the data well in terms of the response probability but they little bit overestimate intra-correlation. The reason, as pointed out by Kuk (2004), is that there is no additional parameters to model correction structure. Next step, consider larger models which include the previous selected two models as sub-models. Composition with power function on Piecewise Flogit provides piecewise power link, i.e. replace \( t \) in (2.10) by \( t^\nu \), \( 0 \leq \nu \leq 1 \). The additional parameter \( \nu \), however, turns out to be 1 with p-value close to 1. Hence, we clearly are in favor of the reduced model. Two-parameter Gamma-Binomial model is significantly different from one-parameter Gamma-binomial with p-value of log-likelihood ratio test (LRT) 0.00002. Not only obtaining the gain of log-likelihood, two-parameter gamma-binomial also improves the modeling on correlation. We continue to consider the models by adding more parameters in the previous selected models. Two additional parameters in Incomplete Gamma-Binomial increase the log-likelihood slightly to -337.6, so there is no strong evidence to differentiate between the Incomplete Gamma-Bin and Complete Gamma-Bin with the p-value of LRT
Forward selection procedure ends with two or three optimal models: Piecewise-Flogit, Gamma-Binomial or Incomplete Gamma-Binomial model.

Besides, in order to compare the Beta-Binomial model considered by Brooks et al. (1997) and Brooks (2001), Incomplete Beta-Binomial model is included even it is not one of models selected from our forward procedure. So is Incomplete A-Binomial for its performance comparing with other incomplete links. Additional two parameters in the Incomplete Beta-Binomial give more modeling flexibility than the Beta-Binomial. The maximum likelihood of the Incomplete Beta-Binomial is $-338.2$, significantly larger than the likelihood $-344.9$ of the Beta-Binomial. The p-value of the likelihood ratio test is $0.001$, indicating significant improvement from the additional two parameters. Among all incomplete links, Incomplete A-link is appealing due to its model simplicity with fewer number of parameters.

Reported in Table 3 are the MLE’s of the marginal response probability $p$, the intra-litter correlation $\phi$, -2*log-likelihood, the Akaike and Bayesian Information Criterion (AIC and BIC). Note that AIC=$-2 \log L + 2 \times npr$ and BIC=$-2 \log L + npr \times \log(nobs)$. Both criteria are used for model comparison and penalize the model complexity in favor of a simple model. Comparing with Brook’s models and Kuk’s models, all proposed models fit the data well in terms of the likelihood as well as the match of the estimated response probability to the observed ones. In fact, Brook’s finite mixture models and Kuk’s power models are falling into our proposed frame. By taking the convex linear combination (Theorem 6.(LP)) with Binomial and Binomial (Beta-Binomial) we obtain Brook’s models, while Kuk’s p-power is obtained by Remark 2 with $h(t; \theta) = \theta^t$ and $\psi_2(t) = t^\nu, \nu \in [0, 1]$. Very broad choices and techniques allow us to achieve desirable performance in fitting a data set.

**Regression Analysis On CD1 Data.** The proposed regression is applied to a real data from a developmental toxicology study conducted at the National Center for Toxicological Research. The study involves replicate experiments with 9 strains of female mice exposed to the herbicide 2,4,5-Trichlorophenoxyacetic acid. We use the data for the CD1 mice, which is extensively analyzed e.g. by George and Bowman (1995) and Kuk (2004). As Gaylor and Razzaghi (1992) pointed out that the classical approach to bioassay presents a particular problem. The investigator in an animal study is usually interested in the results for low doses, where typical human exposure occurs. In a bioassay, though, there will be few responses in the treatment groups given low doses of the toxin. Thus, the parameter estimation for the dose-response curve is greatly affected by what happens at the high doses. However, two different distributions that fit the observed data adequately at the high doses may give very different estimates at the lower end of the curve. Thus we shall focus on the dose levels 0, 30, 45, 60, and 75 excluding the dose level 90 at which more than 95% fetuses were malformed.

As application on E2 data, we use the forward selection procedure for possible optimal models. Starting from one parameter links, we let the parameter $\theta$ linearly depends on the dose level $D$, i.e. $\theta = \alpha + \beta D$. Among all one-parameter links, in terms of log-likelihood, superior performance exhibited are the Gamma-Bin with $\theta_2 = 1$ and Gamma-Log-Bin with $\theta_2 = 1$ which are denoted by Link1, Link2 respectively in the Tables 4 and 5. The Gamma-type model outperformed other models in fitting this data. Indeed, comparison of two models
can also be made through the linear combination of two links. As an illustration, consider
\[ h(t; \theta) = a\text{MM}(t; \theta) + (1 - a)\text{Gamma}(t; \theta) = a\theta/(\theta + t) + (1 - a)(1 + t)^{-\theta} \]
for \( a \in [0, 1] \). Treating the weight \( a \) as parameter, the MLE of \( a \) is approximately 0. This means that the Gamma link is more suitable than the MM-link in fitting this data. This can also be easily observed from the plot of the inverse link function of observed probability of malformation versus dose level in Fig. 3. There is a clearly stronger linear relationship in Gamma-Bin model than in MM-Bin model. The next step is to consider larger models which include the previously selected ones as sub-models. We add \( \theta^2 \) into Gamma-Bin and Gamma-Log-Bin links referred as Link3 and Link4. We treat \( \theta^2 \) as nuisance parameter, in other word, \( \theta^2 \) doesn’t depend on the dose level \( D \). We test the null hypothesis \( \theta^2 = 1 \) by the LRT and find that \( \theta^2 \) in Link3 is not significant but it does in Link4 with the significance level 2.56e -16. We also consider the linear combination of Link1 and Link2, i.e. \( a\text{Link1} + (1-a)\text{Link2} \), which reduces to Link1 or Link2 if \( a = 1 \) or \( a = 0 \). The maximum likelihood estimate of \( a \) is \( 0.678(0.043) \) far away from 0 or 1, which indicates the difference between the larger model and reduce models. Again the LRT confirms that conclusion. The larger model of the previous linear combination model has extra parameter \( \theta^2 \) in Link1 and Link2, i.e. it is the linear combination of Link3 and Link4. The MLE’s of nuisance parameters \( \theta_2 = 0.491(8.9 e - 5), \hat{a} = 0.443(0.325) \). The log-likelihood of the larger model increases to -719.3 and LRTs give \( \theta_2 \) the significance level 0.007, reject \( a = 0 \) and \( a = 1 \) with p-values 7.34e-4, 6.28e-6, respectively. The forward selection procedure concludes that the possible optimal link is the linear combination of the Gamma-Bin and Gamma-Log-Bin links.

Besides the linear combination of two links, we also suggest to consider their product, even it would not be a larger model of either model. The only thing effected is that we can’t directly use LRT for model comparison, but other criteria are possible, for example, sum of absolute differences between estimated and observed number of malformations or square root of sum of squared differences which are introduced later. The MLE of the product of Link1 and Link2 is reported at Table 4. The model of the product of Link1 and Link2 is not significantly different from the product of Link1 and Link2. The modified version of folded logistic link (Piecewise Flogit) is also used to fit the data for the propose of correction.

We compare our proposed procedure with some existing models, i.e., the Binomial, the Beta-Binomial, Generalized Estimation Equations, Williams’s procedure and Kuk’s power family. Reported on Table 4 are the MLE’s of the parameters under various models, the Pearson \( \chi^2 \), and the AIC and BIC for goodness of fit and model comparison. Recall that the Pearson’s \( \chi^2 \) statistic is defined as \( \chi^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 / \text{Var}(\hat{Y}_i) \). Under mild regularity assumptions, it is asymptotically distributed as a Chi-square \( \chi^2(n-p) \), where \( p \) is the number of fitted parameters.

The Binomial model fits poorly because it ignores the intra-litter correlation. The Pearson \( \chi^2 \) of Binomial model is 1514, which is much greater than its expectation 376. This indicates over-dispersion in this data. The Beta-Binomial model assumes, in each dose group \( D \), that the probability of malformation has a Beta Binomial distribution with parameter \( \theta_1, \theta_2 \), which are modeled as \( \theta_1 = \alpha_1 + \alpha_2 D, \quad \theta_2 = \beta_1 + \beta_2 D \). The MLE’s of \( \alpha_1, \alpha_2, \beta_1, \beta_2 \) constrained on \( \theta_1, \theta_2 > 0 \) are reported at Table 4. For the GEE model, we use the R package
geepack. We specify the logistic link on the mean response, and the correlation structure as “exchangeable”. The estimator of intra-litter correlation is 0.2647(0.1166). Williams(1982) presented an iterative algorithm to estimate extra-binomial variation $\phi$, which was incorporated into the reweighed least squares procedure. Here we use software Arc developed by Cook and Weisberg(1999) to get the results of Williams procedure. The over-dispersion parameter in the Williams procedure is 0.2997(0.1231). It was estimated by equating Pearson $\chi^2$ to its degree of freedom, so that the goodness of fit statistic in the Williams procedure is no longer informative. As anticipated, the GEE and Williams procedure produce very similar results since both only use up to the second moment. With Kuk’s q-power model, we directly model $q = \alpha + \beta D$, and treat $\nu$ as nuisance parameter just as we do for other models. The MLE of $\hat{\nu} = 0.608(0.022)$ and MLEs of $\hat{\alpha}$ and $\hat{\beta}$ of Kuk’s model are reported at Table 4. Although Kuk’s q-power can achieve relatively high likelihood, it fails to estimate number of malformations in each dose level, especially there are severe overestimation in 30 mg kg$^{-1}$ dose level and under-prediction at 75 mg kg$^{-1}$ dose group. From this example, AIC or BIC sometimes may not be good criterion for comparison of models which are not nested one in the other one.

Two other criteria are included to assess the models. Both criteria consider overall deviation between estimated and observed number of malformation. The first one is the sum of absolute differences between estimated and observed number of malformations across all dose groups. From Table 5, we can see that most results under the proposed mixture models have superior estimates than the Binomial model, the Beta-Binomial, the GEE and Williams procedure. The second criterion is the square root of sum of squared differences. The proposed models still exhibit competitive performance among previously existed models. The model of linear combination of Gamma-Bin and Gamma-Log links which is chosen by the forward selection procedure demonstrates the best performance in terms of those two criteria as well as AIC and BIC.

**Concluding Remarks.** In this article, we have proposed a unified approach to obtain complete monotonic binomial links, which in turn yield parsimonious binomial mixtures. We have given a rich class of parsimonious mixtures which may meet the needs of various applications in modeling correlated binary data. This class includes many existing distributions such as binomial, beta-binomial, Kuk’s power family, Brook’s correlated binomial, William’s model as special cases. New members of this class based on incomplete functions have also been introduced. We have derived rules (i.e. linear combination, product, and composite) to obtain new parsimonious mixtures from existing ones. A stepwise forward procedure is presented which can serve as a guideline to practitioners for building a possible best model. Our application to two data sets from developmental toxicology literature has exhibited great flexibility and superior performance of the proposed procedure.

**APPENDIX: TECHNICAL DETAILS**

**Proof of Theorem 1.** By (2.3) we have $f(y; \theta) = \binom{m}{y}(-1)^{m-y} \Delta^{m-y} h_y(\theta)$ so that
\(f(y; \theta) \geq 0\) by (2.7). By the inverse formula (Feller, 1970),

\[
\sum_{y=0}^{m} f(y; \theta) = \sum_{y=0}^{m} \left( \begin{array}{c} m \\ y \end{array} \right) (-1)^{m-y} \Delta^{m-y} h_y(\theta) = h_0(\theta) = 1
\]

These verify that \(f\) well defines a probability distribution. \(\Box\)

(2.9) is not a CM function for \(\theta > 1\). By the Taylor expansion, we have

\[
h_\theta(t) = t^{-\theta}(1 + t^{-\theta})^{-1} = \sum_{k=1}^{\infty} (-1)^{k-1} t^{-k\theta}, \quad t \geq 2, \theta > 0.
\]

Then \((-1)^n h_\theta^{(n)}(t) = \sum_{k=1}^{\infty} (-1)^{k-1} a_{n,k}(t, \theta)\) where \(a_{n,k} = (k\theta)(k\theta + 1)\cdots(k\theta + n - 1)t^{-k\theta-n}\).

Since \((k\theta + j)/(k\theta + j + 1)\) is decreasing in \(\theta\) it follows that for \(\theta \geq 1\) and arbitrarily fixed \(t \geq 0\) and every \(k \geq 1\),

\[
a_{n,k}(t, \theta) = \frac{k\theta}{(k + 1)\theta} \cdot \frac{k\theta + 1}{(k + 1)\theta + 1} \cdots \frac{k\theta + n - 1}{(k + 1)\theta + n - 1} \cdot t^\theta
\]

\[
\leq \frac{k}{k + 1} \cdot \frac{k + 1}{k + 2} \cdots \frac{k + n - 1}{k + n} \cdot \frac{k}{k + n} \cdot t^\theta = \frac{k}{k + n} \cdot t^\theta \to 0, \quad n \to \infty.
\]

Then \((-1)^n h_\theta^{(n)}(t) \leq A_n(t)/2 < 0\) for large \(n\) which invalidates the sufficient condition (2.8).

**Proof of Theorem 4.** Set \(dQ_\theta^*(p) = 1[\theta_1, \theta_2]/\hat{h}_\theta(t, \theta)\frac{dQ_\theta(p)}{d\theta}\). Then this determines a probability measure on \([0, 1]\) and rewrites (2.16) as \(\hat{h}(t; \theta) = \int_0^1 p^t dQ_\theta^*(p)\) for \(\theta \in [0, 1]^2 \times \Theta\). By Kendall (1967), there exists a sequence of exchangeable events \(A_1, A_2, \ldots\) such that \(\hat{h}(k; \theta) = \mathbb{P}_{\theta}(A_{r_1} \cap A_{r_2} \cdots \cap A_{r_k})\) and \(\mathbb{P}_{\theta}(A_{r_1} \cap A_{r_2} \cdots \cap A_{r_k} | p) = p^k\) a.s. for some \(p \in [0, 1]\), where \(r_1, \ldots, r_k\) are different. Let \(\hat{B}_1 = 1[A_1]\) the indicator function of set \(A_1\) and \(\hat{Y}_m = \hat{B}_1 + \cdots + \hat{B}_m\).

Then the distribution of \(\hat{Y}_m\) is given by (2.6) so that \(\hat{Y}_m\) and \(Y_m\) have the same distribution.

Abusing notation to write \(\hat{B}_i = B_i\), we obtain a stochastic representation \(Y_m = B_1 + \cdots + B_m\).

Further, conditional on given \(p, B_1, \ldots, B_m\) are independent and identically distribution with the Bernoulli with probability \(p\) of success, so that

\[
\mathbb{P}(Y_m = y | p) = \left( \begin{array}{c} m \\ y \end{array} \right) p^y (1 - p)^{m-y}, \quad y = 0, 1, \ldots, m,
\]

where \(p\) has distribution \(Q_\theta\) which is determined by the CM sequence \(h\). The rest of the proof are straightforward verification analogous to Srivastava and Wu (1993). \(\Box\)

**Asymptotic Distribution of the Mean** \(Y_m/m\). One can easily show that the moment generating function of \(Y_m/m\) is

\[
M_{Y_m/m}(t) = \int_0^1 (pe^{t/m} + 1 - p)^m dQ(p) \to \int_0^1 e^{-pt} dQ(p) = M_Q(t),
\]

where \(M_Q\) is the mgf of \(Q\). Thus it follows from the continuity theorem of the mgf that \(Y_m/m\) converges in distribution to the mixing distribution \(Q\). \(\Box\)
References


Figure 1: The probability curves of the Gamma-power Binomial (upper left) and the linear combination 0.5*GPB+0.5*Binomial (upper right), (1 − α)*GPB + α*Bin for different weights α(lower left) and product GPB*Bin (lower right) for different parameter values. The curves of the GPB are L- or J shaped and close to zero in the middle while its linear combinations and product have various shapes, thus they are more flexible in fitting data. Upper left: the Gamma-power Binomial with θ = (.95, .65, 1.0), (1.0, .85, .95), (2.0, 1.0, 3.0), (.95, .90, 1.4), (.20, .99, .60), (.30, .99, .60), (.90, 1.0, .20). Upper right: the mixture of the 0.5*Gamma-power Binomial + 0.5*Binomial with θ = (0.3, .95, .65, 1.0), (0.6, .95, .65, 1.0), (0.9, .95, .65, 1.0), (0.3, 1.0, .95, .95), (0.3, 3.0, .95, .95), (.20, 3.0, 1.0, 1.0); Lower left: the mixture of Gamma-power Binomial and Binomial with weights α = 0.3, 0.9, 0.6; Lower right: the product of GPB*Bin with θ = (1.0, 0.9, 3.0, 0.9), (3.0, 0.1, 0.2, 0.3), (0.8, 0.1, 0.2, 0.3), (0.5, 0.9, 0.5, 0.5), (0.3, 0.5, 0.5, 0.8), (.05, 0.2, .15, .95).
Figure 2: The probability curves of the A-Binomial (left panel) and the product GPB*Bin (right) for different parameter values. The curves of the GPB and the product have very flexible shapes, thus they are more flexible in modeling data.

Figure 3: On the left panel is the inverse link of the observed probability of malformation versus dose level. Stronger linearity of the Gamma link than the MM link indicates its better suitability. On the right panel is the probability curve of malformation versus dose level under different models. The proposed model closely matches the observed curve and outperforms the others except a slight overestimation at dose level 30.