Bayesian Analysis (2017) **12**, Number 2, pp. 533–555

Data-Dependent Posterior Propriety of a Bayesian Beta-Binomial-Logit Model

Hyungsuk Tak* and Carl N. Morris†

Abstract. A Beta-Binomial-Logit model is a Beta-Binomial model with covariate information incorporated via a logistic regression. Posterior propriety of a Bayesian Beta-Binomial-Logit model can be data-dependent for improper hyper-prior distributions. Various researchers in the literature have unknowingly used improper posterior distributions or have given incorrect statements about posterior propriety because checking posterior propriety can be challenging due to the complicated functional form of a Beta-Binomial-Logit model. We derive data-dependent necessary and sufficient conditions for posterior propriety within a class of hyper-prior distributions that encompass those used in previous studies. When a posterior is improper due to improper hyper-prior distributions, we suggest using proper hyper-prior distributions that can mimic the behaviors of improper choices.

Keywords: objective Bayes, hierarchical models, random effects, overdispersion, logistic regression, beta-binomial, uniform shrinkage prior.

1 Introduction

Binomial data from several independent groups sometimes have more variability than the assumed Binomial distribution for each group's count data. To account for this extra-Binomial variability, called overdispersion, a Beta-Binomial (BB) model (Skellam, 1948) puts a conjugate Beta prior distribution on unknown success probabilities by treating them as random effects. A Beta-Binomial-Logit (BBL) model (Williams, 1982; Kahn and Raftery, 1996) is one way to incorporate covariate information into the BB model. The BBL model has a two-level structure as follows: For each of k independent groups $(j = 1, 2, \ldots, k)$,

$$y_j \mid p_j \overset{\text{indep.}}{\sim} \text{Bin}(n_j,\ p_j), \tag{1}$$

$$p_j \mid r, \boldsymbol{\beta} \overset{\text{indep.}}{\sim} \text{Beta}(rp_j^E,\ rq_j^E), \tag{2}$$

$$p_j^E = 1 - q_j^E \equiv\ E(p_j \mid r, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})} \tag{3}$$

where y_j is the number of successful outcomes out of n_j trials, a sufficient statistic for the random effect p_j, $p_j^E = 1 - q_j^E$ denotes the expected random effect, $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jm})^\top$ is a covariate vector of length m for group j, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^\top$ is a vector of m logistic regression coefficients, and r represents the amount of prior

*Department of Statistics, Harvard University, hyungsuk.tak@gmail.com
†Department of Statistics, Harvard University, morris@stat.harvard.edu

 DOI: 10.1214/16-BA1012

information on p_j^E, considering that the Beta prior distribution in (2) concentrates on p_j^E as r increases (Albert, 1988). We focus only on a logit link function in (3) because it is canonical and is well defined for both binary ($n_j = 1$) and aggregate ($n_j \geq 2$) data. When there is no covariate with an intercept term, i.e., $\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$, the conjugate Beta distribution in (2) is exchangeable, and the BBL model reduces to the BB model.

A Bayesian approach to the BBL model needs a joint hyper-prior distribution of r and $\boldsymbol{\beta}$ that affects posterior propriety. Though a proper joint hyper-prior distribution guarantees posterior propriety, various researchers have used improper hyper-prior distributions hoping for minimal impact on the posterior inference. The articles of Albert (1988) and Daniels (1999) use $dr/(t+r)^2$ with a positive constant t as a hyper-prior probability density function (PDF) for r, and independently an improper flat hyper-prior PDF for $\boldsymbol{\beta}$, $d\boldsymbol{\beta}$. Chapter 5 of Gelman et al. (2013) suggests putting an improper hyper-prior PDF on r, $dr/r^{1.5}$, and independently a proper standard Logistic distribution on β_1 when $\boldsymbol{x}^\top \boldsymbol{\beta} = \beta_1$. (They use a different parameterization: $p_j \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ and $d\alpha d\beta/(\alpha+\beta)^{2.5}$. Transforming $r = \alpha + \beta$ and $p^E = \alpha/(\alpha+\beta)$, we obtain $dp^E dr/r^{1.5}$.) However, the paper of Albert (1988) does not address posterior propriety, the proposition in Daniels (1999) incorrectly concludes that posterior propriety holds regardless of the data, and Chapter 5 of Gelman et al. (2013) specifies an incorrect condition for posterior propriety.

To illustrate with an overly simple example for data-dependent conditions for posterior propriety, we toss two biased coins twice each ($n_j = 2$ for $j = 1, 2$). Let y_j indicate the number of Heads for coin j, and assume a BB model with $\boldsymbol{x}^\top \boldsymbol{\beta} = \beta_1$. If we use any proper hyper-prior PDF for r together with an improper flat density on an intercept term β_1 independently, posterior propriety holds except when both coins land either all Heads ($y_1 = y_2 = 2$) or all Tails ($y_1 = y_2 = 0$) as shown by an X in the diagram. Here the notation O means that the resulting posterior is proper. See Section 4.1 for details.

$y_1 \backslash y_2$	0	1	2
0	X	O	O
1	O	O	O
2	O	O	X

Also, there is a hyper-prior PDF for r that always leads to an improper posterior distribution regardless of the data. The article of Kass and Steffey (1989) adopts an improper joint hyper-prior PDF, $d\boldsymbol{\beta} dr/r$, without addressing posterior propriety. The paper of Kahn and Raftery (1996) uses the same improper hyper-prior PDF for r, dr/r, which they show is a Jeffreys' prior, and independently a proper multivariate Gaussian hyper-prior PDF for $\boldsymbol{\beta}$, declaring posterior propriety without a proof. However, the hyper-prior PDF dr/r used in both articles always leads to an improper posterior regardless of the data.

Making an inference unknowingly based on an improper posterior distribution is dangerous because the improper posterior distribution is not a probability distribution, and thus Markov chain Monte Carlo methods may draw samples from a nonexistent probability distribution (Hobert and Casella, 1996). We derive data-dependent necessary

and sufficient conditions for posterior propriety of a Bayesian BBL model equipped with various joint hyper-prior distributions, and summarize these conditions in Figure 1, the centerpiece of this article. We mainly work on a class of hyper-prior PDFs for r, $dr/(t+r)^{u+1}$, where t is non-negative and u is positive. It includes a proper $dr/(1+r)^2$ (Albert, 1988; Daniels, 1999) and an improper $dr/r^{1.5}$ (Gelman et al., 2013) as special cases. Independently the hyper-prior PDF for $\boldsymbol{\beta}$ that we consider is improper flat (Lebesgue measure) for its intended minimal impact on posterior inference or any proper one. When a posterior distribution is improper due to improper hyper-prior distributions, one possible alternative is to use proper hyper-prior distributions that can mimic the behavior of improper choices, e.g., $dr/(t+r)^{u+1}$ with a small constant t to mimic dr/r^{u+1} and a diffuse Gaussian distribution for $\boldsymbol{\beta}$ to mimic its improper flat choice.

The article is organized as follows. We derive an equivalent inferential model of the Bayesian BBL model in Section 2. We derive necessary and sufficient conditions for posterior propriety, address posterior propriety in past studies, and discuss possible alternatives when posterior distributions are improper in Section 3. In Section 4, we check posterior propriety and investigate posterior properties using two examples.

2 Inferential model

One advantage of the BBL model is that it allows the shrinkage interpretation in inference (Kahn and Raftery, 1996). For $j = 1, 2, \ldots, k$, the conditional posterior distribution of a random effect p_j given hyper-parameters and data is

$$p_j \mid r, \boldsymbol{\beta}, \boldsymbol{y} \overset{\text{indep.}}{\sim} \text{Beta}(rp_j^E + y_j,\ rq_j^E + (n_j - y_j)) \tag{4}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_k)^\top$. The posterior mean of the conditional posterior distribution in (4) is $\hat{p}_j \equiv (1 - B_j)\bar{y}_j + B_j p_j^E$. This mean is a convex combination of the observed proportion $\bar{y}_j = y_j/n_j$ and the expected random effect p_j^E weighted by the relative amount of information in the prior compared to the data, called a shrinkage factor $B_j = r/(r+n_j)$; r determines the precision of p_j^E and n_j determines the precision of $\bar{y}_j$. If the conjugate prior distribution contains more information than the observed data, i.e., ensemble sample size r exceeds individual sample size n_j, then the posterior mean shrinks more towards p_j^E than towards $\bar{y}_j$. The posterior variance of this conditional posterior distribution in (4) is a quadratic function of $\hat{p}_j$, i.e., $\hat{p}_j(1 - \hat{p}_j)/(r + n_j + 1)$.

The conjugate Beta prior distribution of random effects in (2) has unknown hyper-parameters, r and $\boldsymbol{\beta}$. Assuming r and $\boldsymbol{\beta}$ are independent a priori, we introduce their joint hyper-prior PDF as follows:

$$\pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) = f(r)g(\boldsymbol{\beta}) \propto \frac{g(\boldsymbol{\beta})}{(t+r)^{u+1}},\ \text{for } t \geq 0 \text{ and } u > 0. \tag{5}$$

This class of hyper-prior PDFs for r, i.e., $dr/(t+r)^{u+1}$, is proper if $t > 0$ and improper if $t = 0$. A hyper-prior PDF for a uniform shrinkage prior on r, transformed from a uniform prior on a shrinkage factor $dB = d\{r/(t+r)\}$, is $dr/(t+r)^2$ with $u = 1$ for

any positive constant t (Christiansen and Morris, 1997; Daniels, 1999). This uniform shrinkage prior is known to have good frequentist properties for Bayesian estimates (Strawderman, 1971; Christiansen and Morris, 1997; Daniels, 1999). A special case of the uniform shrinkage prior density function is $dr/(1+r)^2$ corresponding to $t = 1$ used by Albert (1988). As t goes to zero, a proper uniform shrinkage prior density, $dr/(t+r)^2$, becomes close to an improper hyper-prior PDF dr/r^2. This improper choice, dr/r^2, is free of an arbitrary constant t and is the most conservative choice that leads to the widest posterior intervals for random effects compared to those obtained by any uniform shrinkage prior (Christiansen and Morris, 1997). Chapter 5 of Gelman et al. (2013) suggests using $dr/r^{1.5}$ as a diffuse hyper-prior PDF, which corresponds to $u = 0.5$ and $t = 0$, together with a standard Logistic distribution on $\boldsymbol{\beta}$. Jeffreys' prior dr/r (Kahn and Raftery, 1996) is not included in the class because it always leads to an improper posterior distribution regardless of the data;[1] see Section 3.2. The hyper-prior PDF for $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, can be any proper PDF or an improper flat density.

The marginal distribution of the data follows independent Beta-Binomial distributions (Skellam, 1948) with random effects integrated out. The probability mass function for the Beta-Binomial distribution is, for $j = 1, 2, \ldots, k$,

$$\pi_{\text{obs}}(y_j \mid r, \boldsymbol{\beta}) = \binom{n_j}{y_j} \frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(rp_j^E,\ rq_j^E)} \tag{6}$$

where the notation $B(a, b)$ indicates a beta function defined as $\int_0^1 v^{a-1}(1-v)^{b-1}dv$ for positive constants a and b. The probability mass function in (6) depends on $\boldsymbol{\beta}$ because the expected random effects, $\{p_1^E, p_2^E, \ldots, p_k^E\}$, are a function of $\boldsymbol{\beta}$ as shown in (3). The likelihood function of r and $\boldsymbol{\beta}$ is the product of these Beta-Binomial probability mass functions being treated as expressions in r and $\boldsymbol{\beta}$, i.e.,

$$L(r, \boldsymbol{\beta}) = \prod_{j=1}^{k} \pi_{\text{obs}}(y_j \mid r, \boldsymbol{\beta}) = \prod_{j=1}^{k} \binom{n_j}{y_j} \frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(rp_j^E,\ rq_j^E)}. \tag{7}$$

When $n_j = 1$, this likelihood function reduces to the one of a logistic regression model:

$$L(r, \boldsymbol{\beta}) = \prod_{j=1}^{k} (p_j^E)^{y_j}(1-p_j^E)^{1-y_j} = \prod_{j=1}^{k} \left(\frac{\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}\right)^{y_j} \left(\frac{1}{1+\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}\right)^{1-y_j}, \tag{8}$$

which is free of r. Since the data tell nothing about r when $n_j = 1$ for all j, it is better not to make any inference on the random effects, $p_1, p_2, \ldots, p_k$, via a Bayesian BBL model unless we have prior information on r.

The joint posterior density function of hyper-parameters, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$, is proportional to their likelihood function in (7) multiplied by the joint hyper-prior PDF

[1] If the symbol A represents a second-level variance component in a two-level Gaussian multilevel model, e.g., $y_j \mid \mu_j \sim \text{Normal}(\mu_j, 1)$ and $\mu_j \mid A \sim \text{Normal}(0, A)$, then A is proportional to $1/r$. The improper hyper-prior PDF $dr/r^2 = -d(1/r)$ corresponds to dA leading to Stein's harmonic prior (Morris and Tang, 2011), $dr/r^{1.5}$ corresponds to $dA/\sqrt{A}$ (Gelman et al., 2013), and dr/r is equivalent to an inappropriate choice dA/A (Morris and Lysy, 2012).

in (5):

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y}) \propto \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \times L(r, \boldsymbol{\beta}). \tag{9}$$

Finally, the full posterior density function of $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)^\top$, r, and $\boldsymbol{\beta}$ is

$$\begin{aligned}\pi_{\text{full.post}}(\boldsymbol{p}, r, \boldsymbol{\beta} \mid \boldsymbol{y}) &\propto \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \times \prod_{j=1}^{k} \pi_{\text{obs}}(y_j \mid p_j) \times \pi_{\text{prior}}(p_j \mid r, \boldsymbol{\beta}) \\ &\propto \pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y}) \times \prod_{j=1}^{k} \pi_{\text{cond.post}}(p_j \mid r, \boldsymbol{\beta}, \boldsymbol{y}) \end{aligned} \tag{10}$$

where the distribution for the prior density function of random effect j, $\pi_{\text{prior}}(p_j \mid r, \boldsymbol{\beta})$, is specified in (2), and the distribution of the conditional posterior density of random effect j, $\pi_{\text{cond.post}}(p_j \mid r, \boldsymbol{\beta}, \boldsymbol{y})$, is specified in (4).

3 Posterior propriety

The full posterior density function in (10) is proper if and only if $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is proper because $\prod_{j=1}^{k} \pi_{\text{cond.post}}(p_j \mid r, \boldsymbol{\beta}, \boldsymbol{y})$ is a product of independent and proper Beta density functions. We therefore focus on posterior propriety of $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$.

Definition 1. *Group j whose observed number of successes is neither 0 nor n_j, i.e., $1 \le y_j \le n_j - 1$, is called an interior group. Similarly, group j is extreme if its observed number of successes is either 0 or n_j. The symbol W_y denotes the set of indices corresponding to interior groups, i.e., $W_y \subseteq \{1, 2, \ldots, k\}$, and k_y is the number of interior groups, i.e., the number of indices in W_y. We use W_y^c to represent the set of $k - k_y$ indices for extreme groups. The notation $X \equiv (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)^\top$ refers to the $k \times m$ covariate matrix of all groups ($k \ge m$) and X_y is the $k_y \times m$ covariate matrix of the interior groups.*

The subscript y emphasizes the data-dependence of k_y, W_y, and X_y. The rank of X_y can be smaller than m when X is of full rank m because we obtain X_y by removing rows of extreme groups from X. If all groups are interior, then $k_y = k$ and $X_y = X$. If all groups are extreme, then $k_y = 0$ and X_y is not defined.

3.1 Conditions for posterior propriety

In Figure 1, we summarize the necessary and sufficient conditions for posterior propriety according to different hyper-prior PDFs, $f(r)$ and $g(\boldsymbol{\beta})$, under two settings: The data contain at least one interior group ($1 \le k_y \le k$) and the data contain only extreme groups ($k_y = 0$).

To prove these conditions, we divide the first setting ($1 \le k_y \le k$) into two: A setting where at least one interior group and at least one extreme group exist ($1 \le k_y \le k - 1$) and a setting where all groups are interior ($k_y = k$). The key to proving conditions for posterior propriety is to derive certain lower and upper bounds for $L(r, \boldsymbol{\beta})$ that factor

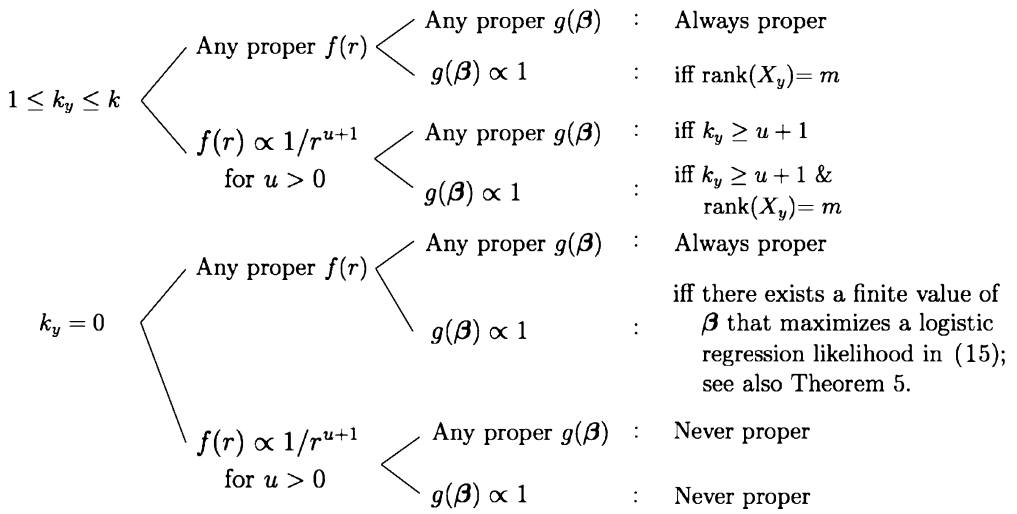

Figure 1: Necessary and sufficient conditions for posterior propriety of $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ according to $\pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) = f(r)g(\boldsymbol{\beta})$ under two settings: The data contain at least one interior group ($1 \le k_y \le k$) and the data contain only extreme groups ($k_y = 0$). The condition, rank$(X_y) = m$, implicitly requires $k_y \ge m$ because X_y is a $k_y \times m$ matrix.

into a function of r and a function of $\boldsymbol{\beta}$. We first derive lower and upper bounds for the Beta-Binomial probability mass function of group j with respect to r and $\boldsymbol{\beta}$ because $L(r, \boldsymbol{\beta})$ is just the product of these probability mass functions of all groups.

Lemma 1. *Lower and upper bounds for the Beta-Binomial probability mass function for interior group j with respect to r and $\boldsymbol{\beta}$ are $rp_j^E q_j^E/(1+r)^{n_j-1}$ and $rp_j^E q_j^E/(1+r)$, respectively, up to a constant multiple. Those for extreme group j with $y_j = n_j$ are $(p_j^E)^{n_j}$ and p_j^E, each, and those for extreme group j with $y_j = 0$ are $(q_j^E)^{n_j}$ and q_j^E, respectively, up to a constant multiple.*

Proof. See Section 6.1. □

Lemma 1 shows that our bounds for the Beta-Binomial probability mass function for either interior or extreme group j with respect to r and $\boldsymbol{\beta}$ factor into a function of r and a function of $\boldsymbol{\beta}$. Because $L(r, \boldsymbol{\beta})$ is a product of these Beta-Binomial probability mass functions of all groups, bounds for $L(r, \boldsymbol{\beta})$ also factor into a function of r and a function of $\boldsymbol{\beta}$. Next we derive certain lower and upper bounds for $L(r, \boldsymbol{\beta})$ with respect to r and $\boldsymbol{\beta}$ under the first setting where all groups are interior.

Lemma 2. *When all groups are interior ($k_y = k$), $L(r, \boldsymbol{\beta})$ can be bounded by*

$$c_1 \frac{r^k \prod_{j=1}^k p_j^E q_j^E}{(1+r)^{\sum_{j=1}^k (n_j - 1)}} \le L(r, \boldsymbol{\beta}) \le c_2 \frac{r^k \prod_{j=1}^k p_j^E q_j^E}{(1+r)^k} \tag{11}$$

where c_1 and c_2 are constants that do not depend on r and $\boldsymbol{\beta}$.

Proof. See Section 6.2. □

When all groups are interior, the joint posterior density function $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ equipped with any joint hyper-prior PDF $\pi_{\text{hyp.prior}}(r, \boldsymbol{\beta})$ is proper if

$$\int_{\mathbf{R}^m} \int_0^\infty \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \times \frac{r^k \prod_{j=1}^k p_j^E q_j^E}{(1+r)^k} dr d\boldsymbol{\beta} < \infty \tag{12}$$

because $r^k \prod_{j=1}^k p_j^E q_j^E / (1+r)^k$ is the upper bound for $L(r, \boldsymbol{\beta})$ specified in (11). Also, the joint posterior density function $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is improper if

$$\int_{\mathbf{R}^m} \int_0^\infty \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \times \frac{r^k \prod_{j=1}^k p_j^E q_j^E}{(1+r)^{\sum_{j=1}^k (n_j - 1)}} dr d\boldsymbol{\beta} = \infty \tag{13}$$

because $r^k \prod_{j=1}^k p_j^E q_j^E / (1+r)^{\sum_{j=1}^k (n_j - 1)}$ is the lower bound for $L(r, \boldsymbol{\beta})$ in (11).

Theorem 1. *When all groups are interior ($k_y = k$), the joint posterior density function of hyper-parameters, $\pi_{hyp.post}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$, equipped with a proper hyper-prior density function on r, $f(r)$, and independently an improper flat hyper-prior density function on $\boldsymbol{\beta}$, $g(\boldsymbol{\beta}) \propto 1$, is proper if and only if $rank(X) = m$.*

Proof. See Section 6.3. □

The condition for posterior propriety with a proper hyper-prior PDF for r is the same as the condition for posterior propriety when r is a completely known constant due to the factorization of the bounds for $L(r, \boldsymbol{\beta})$ in (11). Thus, the condition for posterior propriety in Theorem 1 arises only from the improper hyper-prior PDF for $\boldsymbol{\beta}$.

Theorem 2. *When all groups are interior ($k_y = k$), the joint posterior density function of hyper-parameters, $\pi_{hyp.post}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$, equipped with $f(r) \propto 1/r^{u+1}$ for positive u and independently a proper hyper-prior density function on $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, is proper if and only if $k \geq u + 1$.*

Proof. See Section 6.4. □

The condition for posterior propriety when $\boldsymbol{\beta}$ has a proper hyper-prior distribution is the same as the condition for posterior propriety when $\boldsymbol{\beta}$ is not a parameter to be estimated ($m = 0$) due to the factorization of bounds for $L(r, \boldsymbol{\beta})$ in (11). Thus, the condition for posterior propriety arises solely from the improper hyper-prior PDF for r.

Theorem 3. *When all groups are interior ($k_y = k$), the joint posterior density function of hyper-parameters, $\pi_{hyp.post}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$, equipped with the joint hyper-prior density function $\pi_{hyp.prior}(r, \boldsymbol{\beta}) \propto 1/r^{u+1}$ for positive u is proper if and only if* (i) *$k \geq u + 1$ and* (ii) *$\text{rank}(X) = m$.*

Proof. See Section 6.5. □

The conditions for posterior propriety in Theorem 3 are the combination of the condition in Theorem 1 and that in Theorem 2 because of the factorization of bounds for $L(r, \boldsymbol{\beta})$.

We begin discussing the conditions for posterior propriety under the second setting with at least one interior group and at least one extreme group in the data ($1 \leq k_y \leq k-1$).

Corollary 1. *With at least one interior group and at least one extreme group in the data* ($1 \leq k_y \leq k-1$)*, posterior propriety is determined solely by interior groups, not by extreme groups.*

Proof. See Section 6.6. □

Corollary 1 means that we can remove all the extreme groups from the data to determine posterior propriety, treating the remaining interior groups as a new data set ($k_y = k$). Then we can apply Theorem 1, 2, or 3 to the new data set. If posterior propriety holds with only the interior groups, then posterior propriety with the original data with the combined interior and extreme groups ($1 \leq k_y \leq k-1$) also holds. Corollary 1 justifies combining the first and second settings as shown in Figure 1.

We start specifying the conditions for posterior propriety under the third setting where there are no interior groups in the data ($k_y = 0$).

Lemma 3. *When all groups are extreme* ($k_y = 0$)*, $L(r, \boldsymbol{\beta})$ can be bounded by*

$$c_3 \prod_{j=1}^{k} (p_j^E)^{n_j \times I_{\{y_j = n_j\}}} (q_j^E)^{n_j \times I_{\{y_j = 0\}}} \;\leq\; L(r, \boldsymbol{\beta}) \;\leq\; c_4 \prod_{j=1}^{k} (p_j^E)^{I_{\{y_j = n_j\}}} (q_j^E)^{I_{\{y_j = 0\}}} \quad (14)$$

where c_3 and c_4 are constants that do not depend on r and $\boldsymbol{\beta}$, and $I_{\{D\}}$ is the indicator function of D.

Proof. See Section 6.7. □

The upper and lower bounds in (14) are free of r, indicating that the hyper-prior distribution of r must be proper for posterior propriety in this case ($k_y = 0$). If the hyper-prior distribution of $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, is also proper, the resulting posterior is automatically proper and we do not need to check posterior propriety. However, the posterior can be improper when $g(\boldsymbol{\beta})$ is improper. The next theorem deals with a case where $g(\boldsymbol{\beta}) \propto 1$.

Theorem 4. *When all groups are extreme* ($k_y = 0$)*, the posterior density function of hyper-parameters, $\pi_{hyp.post}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$, equipped with a proper hyper-prior density function for r, $f(r)$, and independently $g(\beta_1) \propto 1$, is proper if and only if there exists a finite value of $\boldsymbol{\beta}$ that maximizes the upper bound in (14) up to a constant, i.e.,*

$$\prod_{j=1}^{k} \left(\frac{\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})} \right)^{I_{\{y_j = n_j\}}} \left(\frac{1}{1 + \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})} \right)^{I_{\{y_j = 0\}}}. \quad (15)$$

Proof. See Section 6.8. □

The function in (15) is essentially the same as the likelihood function of a logistic regression in (8) because the powers in (15) are either one or zero with $I_{\{y_j=0\}} = 1 - I_{\{y_j=n_j\}}$. Thus, the value of $\boldsymbol{\beta}$ that maximizes (15) is the same as the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ in (8) for which we set $y_j = 1$ if $y_j \geq 1$. A quick way to check whether there exists a finite value of $\boldsymbol{\beta}$ that maximizes (15) is to fit a logistic regression model after setting $y_j = 1$ if $y_j \geq 1$, using any statistical software, e.g., `glm` in R (R Development Core Team, 2016). If no errors emerge, then the finite MLE of $\boldsymbol{\beta}$ exists; its uniqueness is guaranteed if the MLE exists in a logistic regression (Jacobsen, 1989). However, Theorem 4 is inconvenient in that we need to fit a logistic regression model to check posterior propriety. The next theorem specifies more descriptive sufficient conditions for posterior propriety that do not require fitting a logistic regression, which are also necessary conditions when there is only an intercept term, $\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$ for all j.

Theorem 5. *When all groups are extreme ($k_y = 0$), the posterior density function of hyper-parameters, $\pi_{hyp.post}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$, equipped with a proper hyper-prior density function for r, $f(r)$, and independently $g(\beta_1) \propto 1$, is proper if* (i) *there are at least m clusters of groups whose covariate values are the same within each cluster and different between clusters, and* (ii) *in each cluster there are at least one group of all successes and at least one group of all failures. The $k \times m$ covariate matrix X is assumed to be of full rank m. These two conditions are also necessary conditions when $\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$.*

Proof. See Section 6.9. □

When $\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$, the necessary and sufficient conditions in Theorem 5 simply reduce to having at least one group with all successes and at least one group with all failures in the data. Theorem 4 of Natarajan and Kass (2000) shows that this reduced condition is the same as the condition in Theorem 4, i.e., there exists a finite value of $\boldsymbol{\beta}$ that maximizes (15).

The two conditions in Theorem 5 are only sufficient conditions when there are covariates. For necessary conditions in this case, we need to show that integration of the lower bound in (14) with respect to $\boldsymbol{\beta}$ is not finite when either conditions in Theorem 5 are not met. However, the integration itself seems mathematically intractable. If either conditions in Theorem 5 are not met, we need to go back to Theorem 4, checking the existence of a finite value of $\boldsymbol{\beta}$ that maximizes (15) by fitting a logistic regression.

Theorem 6. *When all groups are extreme ($k_y = 0$), the posterior density function of hyper-parameters $\pi_{hyp.post}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$, equipped with any improper hyper-prior density function $f(r)$ and independently any hyper-prior density $g(\boldsymbol{\beta})$, is always improper.*

Proof. Because the lower bound for $L(r, \boldsymbol{\beta})$ in Lemma 3 is free of r, $L(r, \boldsymbol{\beta})$ cannot make the integration of $f(r)$ finite when $f(r)$ is improper. Thus, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ should always be improper under this setting. □

3.2 Posterior propriety in previous studies

The article of Albert (1988) does not address posterior propriety for $d\boldsymbol{\beta}dr/(1+r)^2$. Our work in Figure 1 shows that the condition for posterior propriety is that $\text{rank}(X_y) = m$ when $1 \leq k_y \leq k$, i.e., the covariate matrix of interior groups is of full rank m, and that there exists a finite value of $\boldsymbol{\beta}$ that maximizes (15) when $k_y = 0$.

The proposition (1c to be specific) in Daniels (1999) for posterior propriety of the Bayesian BBL model with hyper-prior distributions $d\boldsymbol{\beta}dr/(t+r)^2$ argues that the posterior distribution is always proper. However, its proof is based on a limited case with only an intercept term, $\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$. Under this simplified setting, if there is only one extreme group with two trials ($y_1 = 2$, $n_1 = 2$), the resulting joint posterior density function of r and β_1 is

$$\pi_{\text{hyp.post}}(r, \beta_1 \mid \boldsymbol{y}) \propto \frac{(1 + rp^E)p^E}{(1+r)(t+r)^2}. \tag{16}$$

The integration of (16) with respect to β_1 is not finite because $p^E = \exp(\beta_1)/(1 + \exp(\beta_1))$ converges to one as β_1 approaches infinity. Figure 1 shows that at least one interior group is required in the data for posterior propriety of the Bayesian BBL model under the simplified setting ($\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$) of Daniels (1999). Moreover, if all groups are extreme in the data under the simplified setting with an intercept term, the posterior is proper if and only if there exist at least one extreme group with all successes and at least one extreme group with all failures. In our counter-example, there is only one extreme group with all successes, and thus the resulting posterior in (16) is improper.

With only an intercept term ($\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$), Chapter 5 of Gelman et al. (2013) specifies that the joint posterior density function $\pi_{\text{hyp.post}}(r, \beta_1 \mid \boldsymbol{y})$ with $dr/r^{1.5}$ and independently with a proper standard Logistic distribution on β_1 is proper if there is at least one interior group. However, the resulting posterior can be improper with this condition. For example, when there is only one interior group with two trials ($y_1 = 1$, $n_1 = 2$), the joint posterior density function of r and β_1 is

$$\pi_{\text{hyp.post}}(r, \beta_1 \mid \boldsymbol{y}) \propto \pi_{\text{hyp.prior}}(r, \beta_1) \times L(r, \beta_1) \propto \frac{p^E q^E}{r^{1.5}} \times \frac{rp^E q^E}{(1+r)}, \tag{17}$$

where $p^E = 1 - q^E = \exp(\beta_1)/(1 + \exp(\beta_1))$. The integration of this joint posterior density function with respect to r is not finite because the density function goes to infinity as r approaches zero. (The integral of $dr/r^{0.5}$ over the range $[0, 0 + \epsilon]$ for a positive constant ϵ is not finite.) To achieve posterior propriety in this setting, we need at least two interior groups in the data as shown in Figure 1.

The posterior distributions of Kass and Steffey (1989) and Kahn and Raftery (1996) are always improper regardless of the data due to their hyper-prior PDF dr/r. This is because the likelihood function in (7) approaches $c(\boldsymbol{\beta})$, a positive constant with respect to r, as r increases to infinity. Then the hyper-prior PDF dr/r, whose integration becomes infinite over the range $[\epsilon, \infty)$ for a positive constant ϵ, governs the right tail behavior of the conditional posterior density function of r, $\pi_{\text{hyp.cond.post}}(r \mid \boldsymbol{\beta}, \boldsymbol{y})$. It indicates that $\pi_{\text{hyp.cond.post}}(r \mid \boldsymbol{\beta}, \boldsymbol{y})$ is improper, and thus the joint posterior density $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is improper.

3.3 Inference when a posterior distribution is improper

Making an inference based on an improper posterior distribution is dangerous because most statistical inferential tools assume that the target distribution is a probability distribution but the improper posterior distribution is not a probability distribution. For example, Hobert and Casella (1996) call attention to running a Gibbs sampler on an improper posterior distribution because the Gibbs sampler may seem to work well even when the posterior distribution is improper. They emphasize checking posterior propriety in advance to prevent a (non-recurrent) Gibbs chain from converging to some nonexistent probability distribution. Athreya and Roy (2014) also show that Markov chain Monte Carlo methods can be misleading when the posterior is improper because a standard average estimator based on Markov chains converges to zero with probability one. They introduce regenerative sequence Monte Carlo methods that enable a valid inference even when a posterior distribution is improper.

When it comes to a BBL model, the conditions for posterior propriety in Figure 1 can be met in most cases because in practice the data are composed of a suitably large number of groups, k. However, improper hyper-prior PDFs may result in posterior impropriety when the data are composed of a small number of groups. In this case, we recommend using proper hyper-prior PDFs for r and $\boldsymbol{\beta}$, e.g., a uniform shrinkage prior on r, $dr/(t+r)^2$, which is known to produce good frequentist properties (Strawderman, 1971; Christiansen and Morris, 1997), and a diffuse Gaussian prior on $\boldsymbol{\beta}$ with relatively large standard deviations (Kahn and Raftery, 1996). Setting a small constant t in a uniform shrinkage prior is considered as a conservative choice that allows the data to speak more with smaller shrinkage factors (Christiansen and Morris, 1997). Another possibility (except when $n_j = 1$ for all j) is to estimate MLEs of r and $\boldsymbol{\beta}$ via (7) and plug these estimates into the conditional Beta distributions of random effects in (4). This approach can be considered as an empirical Bayes (EB) approach (Efron and Morris, 1975) with $\pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \propto 1$. However, this EB approach tends to be over-confident in estimating random effects when k is small because the EB approach does not account for the uncertainties of unknown r and $\boldsymbol{\beta}$ though these uncertainties are large when k is small.

4 Numerical illustration

4.1 Data of two bent coins

We have two biased coins; a bent penny and a possibly differently bent nickel ($k = 2$). We flip these coins twice for each ($n_1 = n_2 = 2$) and record the number of Heads for the penny (y_1) and also for the nickel (y_2). We model this experiment as $y_j \mid p_j \sim \text{Bin}(2, p_j)$ independently, where p_j is the unknown probability of observing Heads for coin j. We assume an i.i.d. prior distribution for random effects, $p_j \mid r, \beta_1 \sim \text{Beta}(rp^E, rq^E)$, where $p^E = 1 - q^E = \exp(\beta_1)/[1 + \exp(\beta_1)]$, i.e., a BB model.

We look into posterior propriety under four different settings depending on whether the hyper-prior distribution for β_1 (or equivalently p^E) is proper or improper flat $d\boldsymbol{\beta}$, and on whether the hyper-prior distribution of r is proper or dr/r^2.

Table 1 shows when the posterior distribution is proper (denoted by O) and when it is not (denoted by X). The posterior distribution in case (a) is always proper because both hyper-prior distributions for r and β_1 are proper. In case (b) where β_1 has the Lebesgue measure and r has a proper hyper-prior PDF, the posterior is proper unless both coins land either all Heads ($y_1 = y_2 = 2$) or all Tails ($y_1 = y_2 = 0$). This is because the condition for posterior propriety is that the covariate matrix of interior coins is of full rank and this condition without any covariates is met if at least one coin is interior; see Figure 1. In cases (c) and (d), where r has the improper hyper-prior PDF, dr/r^2, posterior propriety holds only when each coin shows one Head and one Tail, i.e., both coins are interior ($y_1 = y_2 = 1$); see Figure 1. Cases (c) and (d) have the same condition for posterior propriety because the condition that arises from the improper flat hyper-prior PDF for β_1 in case (d) is automatically met if the condition arising from the improper hyper-prior PDF for r, i.e., $k_y \geq 2$, is met.

Next, we check the effect of different joint hyper-prior PDFs used in cases (a)–(d) on the random effect estimation, e.g., p_1. For this purpose, we set $g(\beta_1) = \mathrm{N}(\beta_1 \mid 0, 10^{10})$, a diffuse Gaussian distribution with mean zero and variance 10^{10} for a proper hyper-prior PDF of β_1, and set $f(r) \propto 1/(10^{-5} + r)^2$ for a proper hyper-prior PDF of r. We draw 55,000 posterior samples of r and β_1 from their joint posterior distribution, $\pi_{\text{hyp.post}}(r, \beta_1 \mid \boldsymbol{y})$, using a Metropolis within Gibbs sampler (Tierney, 1994), discarding the first 5,000 samples as burn-in. We adjust proposal scales of independent Gaussian proposal distributions to obtain a reasonable acceptance probability around 0.35 for each parameter. Using the posterior samples of r and β_1, we draw the posterior sample of p_1 from its marginal posterior distribution $\pi_{\text{marg.post}}(p_1 \mid \boldsymbol{y})$ via a Monte Carlo integration:

$$\pi_{\text{marg.post}}(p_1 \mid \boldsymbol{y}) = \int_{\mathbf{R}} \int_0^{\infty} \pi_{\text{cond.post}}(p_1 \mid r, \beta_1, \boldsymbol{y}) \times \pi_{\text{hyp.post}}(r, \beta_1 \mid \boldsymbol{y}) dr d\beta_1, \quad (18)$$

i.e., sampling p_1 from $\pi_{\text{cond.post}}(p_1 \mid r, \beta_1, \boldsymbol{y})$ given already sampled r and β_1. In addition, we estimate p_1 via an EB approach for a comparison; estimating MLEs of r and β_1,

(a) Any proper $f(r)$ and any proper $g(\beta_1)$

$y_1 \backslash y_2$	0	1	2
0	O	O	O
1	O	O	O
2	O	O	O

(b) Any proper $f(r)$ and $g(\beta_1) \propto 1$

$y_1 \backslash y_2$	0	1	2
0	X	O	O
1	O	O	O
2	O	O	X

(c) $f(r) \propto 1/r^2$ and any proper $g(\beta_1)$

$y_1 \backslash y_2$	0	1	2
0	X	X	X
1	X	O	X
2	X	X	X

(d) $f(r) \propto 1/r^2$ and $g(\beta_1) \propto 1$

$y_1 \backslash y_2$	0	1	2
0	X	X	X
1	X	O	X
2	X	X	X

Table 1: The symbol O indicates that the posterior distribution is proper on corresponding data, and the symbol X indicates that the posterior distribution is not proper on corresponding data.

Data \ Model	Case (a)	Case (b)	Case (c)	Case (d)	EB
$y_1 = y_2 = 1$	(0.048, 0.950)	(0.048, 0.951)	(0.049, 0.951)	(0.049, 0.950)	(0.490, 0.510)
$y_1 = 0, y_2 = 1$	(0.000, 0.247)	(0.000, 0.242)	Improper	Improper	(0.218, 0.284)

Table 2: The 95% posterior intervals of p_1 obtained by Bayesian BBL models equipped with joint hyper-prior PDFs in cases (a)–(d), and those obtained by an empirical Bayes (EB) approach. We set $g(\beta_1) = \mathrm{N}(\beta_1 \mid 0, 10^{10})$ and $f(r) \propto 1/(10^{-5} + r)^2$ for proper hyper-prior PDFs of β_1 and r, respectively.

inserting these into $\pi_{\text{cond.post}}(p_1 \mid r, \beta_1, \boldsymbol{y})$, and calculating (0.025, 0.975) quantiles of this conditional Beta posterior distribution $\pi_{\text{cond.post}}(p_1 \mid r, \beta_1, \boldsymbol{y})$.

We fit these models on the data $\{y_1 = y_2 = 1\}$ for which posterior distributions in cases (a)–(d) are all proper. The resulting 95% posterior intervals for p_1 are summarized in the first row of Table 2. All these intervals are similar because the proper hyper-prior PDF of r, $dr/(10^{-5} + r)^2$, used in cases (a) and (b) mimics well its improper limit, dr/r^2, used in cases (c) and (d), and because the diffuse Gaussian hyper-prior PDF of β_1 behaves similarly to an improper flat density function. These intervals are wide, reflecting on the lack of information about r and β_1 in two observations. However, the EB interval centered at 0.5 is much too narrow because it does not account for the uncertainties of unknown r and β_1.

The hyper-prior PDFs in cases (c) and (d) result in an improper posterior for the data $\{y_1 = 0, y_2 = 1\}$. Thus, we fit models equipped with hyper-prior PDFs in cases (a) and (b) and an EB model on these data. The posterior intervals for p_1 are summarized in the second row of Table 2. The intervals in cases (a) and (b) are similar because the diffuse Gaussian prior for β_1 is close to an improper flat prior. The EB interval centered at around 0.25 is again much narrower than the full Bayesian intervals in (a) and (b).

4.2 Data of five hospitals

New York State Cardiac Advisory Committee (2014) has reported the outcomes for the Valve Only and Valve/CABG surgeries. The data are based on the patients discharged between December 1, 2008, and November 30, 2011 in 40 non-federal hospitals in New York State. We select the smallest five hospitals with respect to the number of patients for simplicity. Table 3 shows the data including the number of cases (n_j), the number of deaths (y_j), and expected mortality rate (EMR_j). The EMR_j is a hospital-wise average over the predicted probabilities of death for each patient; the larger the EMR_j is, the more difficult cases hospital j handles. We use the EMR_j as a continuous covariate. We assume $y_j \mid p_j \sim \mathrm{Bin}(n_j, p_j)$ independently. We also assume that the unknown true mortality rates p_j come from independent conjugate Beta prior distributions in (2) with $x_j^T \boldsymbol{\beta} = \beta_1 x_{1j} + \beta_2 x_{2j}$, where $x_{1j} = 1$ and $x_{2j} = \mathrm{EMR}_j$.

This time we consider four joint hyper-prior densities: $d\boldsymbol{\beta} dr/r^2$, $d\boldsymbol{\beta} dr/(10^{-5} + r)^2$, $d\boldsymbol{\beta} dr/r^{1.5}$ and $d\boldsymbol{\beta} dr/(10^{-5} + r)^{1.5}$. The condition for posterior propriety is the same as $\mathrm{rank}(X_y) = 2$ for all four joint hyper-prior PDFs because this condition automatically meets $k_y \geq 2$. The data in Table 3 satisfy the condition for posterior propriety because

j	1	2	3	4	5
n_j	54	75	93	104	105
y_j	3	4	1	1	1
EMR_j (x_{2j})	4.30	2.21	2.59	4.73	3.28

Table 3: Data of five hospitals. The number of patients in hospital j is denoted by n_j, the number of deaths in hospital j is denoted by y_j, and the expected mortality rate (%) for hospital j is denoted by EMR_j, which is a continuous covariate.

all the hospitals are interior ($1 \le y_j \le n_j - 1$ for all j and thus $k = k_y = 5$) and their covariate matrix $X = X_y$ is of full rank.

Based on the data in Table 3, we make two hypothetical data sets in Table 4. In the first hypothetical data set, only one hospital is interior. The resulting posterior distribution is improper for the four joint hyper-prior PDFs because the rank of the covariate matrix of this interior hospital is not two ($\text{rank}(X_y) = 1$). In the second hypothetical data set, two hospitals are interior but their EMRs are the same, meaning that the rank of the covariate matrix of these two interior hospitals is one. Thus, the resulting posterior is improper for the four joint hyper-prior PDFs.

Next we compare several models using these data sets in Table 3 and Table 4 to see the effect of different constants, t and u, in $dr/(t+r)^{u+1}$; we consider using either $u = 1$ or $u = 0.5$ and either $t = 0$ or $t = 10^{-5}$. The sampling configurations are the same as those in the previous section except that we set $g(\boldsymbol{\beta}) = \text{N}(\boldsymbol{\beta} \mid 0 \times \mathbf{1}_2, 10^{10} \times I_2)$ for all models, where $\mathbf{1}_2$ is a vector of ones and I_2 is a 2×2 identity matrix. Table 5 summarizes the 95% posterior intervals for p_1.

When models are all proper based on the data in Table 3, the interval estimates are similar between $t = 10^{-5}$ and $t = 0$, but quite different depending on whether $u = 1$ or $u = 0.5$. Clearly, intervals based on $u = 1$ are wider (more conservative) than those based on $u = 0.5$. This is because dr/r^2 puts more weight at zero than $dr/r^{1.5}$ a priori, and thus dr/r^2 produces smaller posterior samples of r that leads to wider interval estimates in turn; the variance of a conditional Beta posterior distribution for p_j in (4), $\hat{p}_j(1-\hat{p}_j)/(r+n_j+1)$, increases as r decreases, where $\hat{p}_j$ is its posterior mean. The improper hyper-prior PDFs, dr/r^2 and $dr/r^{1.5}$, lead to posterior impropriety for the data in Table 4 due to the reasons specified above. The EB approach leads to much narrower intervals for all three data sets.

j	1	2	3	4	5
n_j	54	75	93	104	105
y_j	1	0	0	0	0
EMR_j	4.30	2.21	2.59	4.73	3.28

j	1	2	3	4	5
n_j	54	75	93	104	105
y_j	1	2	0	0	0
EMR_j	4.30	4.30	2.59	4.73	3.28

Table 4: Two hypothetical data sets of five hospitals. The number of patients in hospital j is denoted by n_j, the number of deaths in hospital j is denoted by y_j, and the expected mortality rate (%) for hospital j is denoted by EMR_j, which is a continuous covariate. In the first data set, only the first hospital is interior. In the second data set, the first two hospitals are interior but their EMRs are the same, i.e., their covariate matrix, X_y, is not of full rank.

Data\Model	$1/r^2$	$1/(10^{-5}+r)^2$	$1/r^{1.5}$	$1/(10^{-5}+r)^{1.5}$	EB
Table 3	(0.011, 0.116)	(0.011, 0.115)	(0.008, 0.099)	(0.008, 0.100)	(0.012, 0.046)
Table 4 (L)	Improper	(0.000, 0.067)	Improper	(0.000, 0.066)	(0.003, 0.005)
Table 4 (R)	Improper	(0.000, 0.068)	Improper	(0.001, 0.062)	(0.002, 0.030)

Table 5: The 95% posterior intervals of p_1 obtained by Bayesian BBL models equipped with hyper-prior PDFs, $g(\boldsymbol{\beta}) = \mathrm{N}(\boldsymbol{\beta} \mid 0 \times \mathbf{1}_2, 10^{10} \times I_2)$, which is the same for all models, and $dr/(t+r)^{u+1}$ with $u = 1$ or $u = 0.5$ and with $t = 0$ or $t = 10^{-5}$. The 95% intervals obtained by an empirical Bayes (EB) approach appear in the last column. The left hypothetical data in Table 4 are denoted by Table 4 (L) and the right one by Table 4 (R).

5 Concluding remarks

The Beta-Binomial-Logit (BBL) model accounts for the overdispersion in the Binomial data obtained from several independent groups with their covariate information considered. From a Bayesian perspective, we derive data-dependent necessary and sufficient conditions for posterior propriety of the Bayesian BBL model equipped with a joint hyper-prior density, $g(\boldsymbol{\beta})d\boldsymbol{\beta}dr/(t+r)^{u+1}$, where $t \geq 0$, $u > 0$, and $g(\boldsymbol{\beta})$ can be any proper density or an improper flat density. This joint hyper-prior density encompasses those used in the literature. Using two numerical illustrations, we look into posterior propriety and posterior properties, suggesting conservative and diffuse choices of proper hyper-prior densities be used when the posterior is improper due to improper hyper-prior probability density functions.

There are several opportunities to build upon our work. First of all, it is not clear whether the necessary and sufficient conditions specified in Figure 1 hold for other link functions, e.g., a complementary log-log link function; a probit link function is not appropriate for a BBL model because it is defined on binary data ($n_j = 1$) not on aggregate data ($n_j \geq 2$). As for frequency coverage properties, the data-dependent conditions for posterior propriety make it hard to evaluate these properties because some models with improper hyper-prior distributions do not define a frequency procedure for all possible data sets; the resulting posterior can be improper for some data sets. Thus, in a repeated sampling simulation, we may evaluate frequency properties given only the simulated data sets that achieve posterior propriety. If the probability of generating the data sets that lead to an improper posterior is negligible, this frequency evaluation procedure will be justified. We leave these for our future research.

6 Proofs

6.1 Proof of Lemma 1

If group j is interior ($1 \leq y_j \leq n_j - 1$, $n_j \geq 2$), we can derive an upper bound for the Beta-Binomial probability mass function of interior group j with respect to r and $\boldsymbol{\beta}$ as follows. All bounds in this proof are up to a constant multiple. With notation $q_j^E = 1 - p_j^E$,

$$\pi_{\text{obs}}(y_j \mid r, \boldsymbol{\beta}) \propto \frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(rp_j^E,\ rq_j^E)} \tag{19}$$

$$= \frac{B(1 + rp_j^E,\ 1 + rq_j^E)}{B(rp_j^E,\ rq_j^E)} \frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(1 + rp_j^E,\ 1 + rq_j^E)} \tag{20}$$

$$= \frac{rp_j^E q_j^E}{1+r} \frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(1 + rp_j^E,\ 1 + rq_j^E)} \tag{21}$$

$$= \frac{rp_j^E q_j^E}{1+r} \frac{\int_0^1 v^{y_j - 1 + rp_j^E} (1-v)^{n_j - y_j - 1 + rq_j^E} dv}{\int_0^1 v^{rp_j^E} (1-v)^{rq_j^E} dv} \ \leq\ \frac{rp_j^E q_j^E}{1+r}. \tag{22}$$

The ratio of the two beta functions in (22) is less than or equal to one because the integrand of the beta function in the numerator is less than or equal to the integrand of the beta function in the denominator, considering that $0 \leq y_j - 1 \leq n_j - 2$ and $0 \leq n_j - y_j - 1 \leq n_j - 2$.

A lower bound for the ratio of the two beta functions in (19) is

$$\frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(rp_j^E,\ rq_j^E)} \tag{23}$$

$$= \frac{(y_j - 1 + rp_j^E) \cdots (1 + rp_j^E) rp_j^E (n_j - y_j - 1 + rq_j^E) \cdots (1 + rq_j^E) rq_j^E}{(n_j - 1 + r)(n_j - 2 + r) \cdots (1 + r) r} \tag{24}$$

$$\geq \frac{r^2 p_j^E q_j^E}{(n_j - 1 + r)(n_j - 2 + r) \cdots (1 + r) r} \geq \frac{rp_j^E q_j^E}{(n_{\max} + r)^{n_j - 1}} \geq \frac{rp_j^E q_j^E}{(1 + r)^{n_j - 1}} \tag{25}$$

where $n_{\max} \equiv \max\{n_1, \ldots, n_k\}$. The first inequality in (25) holds because each factor (except rp_j^E and rq_j^E) in the numerator of (24) is greater than or equal to one. The third inequality holds up to a constant multiple, $1/n_j^{n_j - 1}$, because $(n_{\max} + r)/(1 + r) \leq n_j$.

If group j is extreme with all successes ($y_j = n_j \geq 1$), the upper bound for the Beta-Binomial probability mass function of group j with respect to r and $\boldsymbol{\beta}$ is

$$\pi_{obs}(y_j = n_j \mid r, \boldsymbol{\beta}) \propto \frac{B(n_j + rp_j^E,\ rq_j^E)}{B(rp_j^E,\ rq_j^E)} \ \leq\ \frac{B(1 + rp_j^E,\ rq_j^E)}{B(rp_j^E,\ rq_j^E)} = p_j^E. \tag{26}$$

The inequality holds because the integrand of the beta function in the numerator becomes the largest when $n_j = 1$. The lower bound for the Beta-Binomial probability mass function of this extreme group with respect to r and $\boldsymbol{\beta}$ is

$$\frac{B(n_j + rp_j^E,\ rq_j^E)}{B(rp_j^E,\ rq_j^E)} = \frac{(n_j - 1 + rp_j^E)(n_j - 2 + rp_j^E) \cdots (1 + rp_j^E) p_j^E}{(n_j - 1 + r)(n_j - 2 + r) \cdots (1 + r)} \geq (p_j^E)^{n_j}. \tag{27}$$

The inequality holds because the ratio of the two beta functions in (27) is a decreasing function of r, and thus the lower bound is achieved as r goes to infinity.

Similarly, when group j is extreme with all failures ($y_j = 0$, $n_j \geq 1$), we can bound the ratio of the two beta functions of this extreme group by

$$(q_j^E)^{n_j} \leq \frac{B(rp_j^E,\ n_j + rq_j^E)}{B(rp_j^E,\ rq_j^E)} \leq q_j^E. \tag{28}$$

6.2 Proof of Lemma 2

Without any extreme groups in the data, an upper bound for $L(r, \boldsymbol{\beta})$ is the product of the k upper bounds for the Beta-Binomial probability mass function of each interior group in (22), i.e., $r^k(\prod_{j=1}^k p_j^E q_j^E)/(1+r)^k$. Similarly, a lower bound for $L(r, \boldsymbol{\beta})$ is the product of the k lower bounds for the Beta-Binomial probability mass function of each interior group in (25), i.e., $r^k(\prod_{j=1}^k p_j^E q_j^E)/(1+r)^{\sum_{j=1}^k (n_j-1)}$. It is clear that both bounds factor into a function of r and a function of $\boldsymbol{\beta}$.

6.3 Proof of Theorem 1

Because the r part of the upper bound for $L(r, \boldsymbol{\beta})$ in Lemma 2, i.e., $r^k/(1+r)^k$, is always less than one, an upper bound for $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$, up to a normalizing constant, factors into a function of r and a function of $\boldsymbol{\beta}$ as follows:

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y}) \propto f(r)g(\boldsymbol{\beta})L(r, \boldsymbol{\beta}) < f(r) \times \prod_{j=1}^k p_j^E q_j^E. \tag{29}$$

The integration of $f(r)$ with respect to r is finite because it is a proper hyper-prior PDF. The integration of $\prod_{j=1}^k p_j^E q_j^E$ with respect to $\boldsymbol{\beta}$ is finite if and only if the covariate matrix of all groups, X, is of full rank m. To show the sufficient condition, let us choose m sub-groups, whose index set is denoted by W_{sub}, such that the $m \times m$ covariate matrix of the sub-groups is still of full rank m. Then,

$$\prod_{j=1}^k p_j^E q_j^E \leq \prod_{j \in W_{\text{sub}}} p_j^E q_j^E = \prod_{j \in W_{\text{sub}}} \frac{\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}{[1 + \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})]^2}. \tag{30}$$

The integration of this upper bound in (30) with respect to $\boldsymbol{\beta}$ factors into m separate integrations after linear transformations, $h_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}$ for all $j \in W_{\text{sub}}$, whose Jacobian is a constant:

$$\int_{R^m} \prod_{j \in W_{\text{sub}}} \frac{\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}{[1 + \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})]^2} d\boldsymbol{\beta} \propto \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j \in W_{\text{sub}}} \frac{\exp(h_j)}{[1 + \exp(h_j)]^2} dh_j = 1. \tag{31}$$

Each integration on the right hand side leads to one because each integrand is a proper density function of the standard Logistic distribution with respect to h_j.

Next, we show that if the rank of X is not of full rank m, then the integration of the $\boldsymbol{\beta}$ part of the lower bound for $L(r, \boldsymbol{\beta})$ in Lemma 2, i.e., $\prod_{j=1}^k p_j^E q_j^E$, cannot be finite.

Without loss of generality, let us assume that the rank of X is $m-1$ and that the last column of X can be expressed as a linear function of the first $m-1$ columns. Due to the singularity of X, we can always find $m-1$ linear functions, $t_i(\beta_i, \beta_m)$, $i = 1, 2, \ldots, m-1$, such that $\boldsymbol{x}_j^\top \boldsymbol{\beta} = x_{j1}t_1(\beta_1, \beta_m) + x_{j2}t_2(\beta_2, \beta_m) + \cdots + x_{j,m-1}t_{m-1}(\beta_{m-1}, \beta_m)$. As a result, the integration of $\prod_{j=1}^k p_j^E q_j^E$ with respect to $\boldsymbol{\beta}$ is infinity after a linear transformation from $\boldsymbol{\beta}$ to $(\beta_1^* = t_1(\beta_1, \beta_m),\ \beta_2^* = t_2(\beta_2, \beta_m),\ \ldots,\ \beta_{m-1}^* = t_{m-1}(\beta_{m-1}, \beta_m),\ \beta_m)^\top$, whose Jacobian is one. For notational simplicity, we use two $(m-1) \times 1$ vectors, $\boldsymbol{x}_j^* \equiv (x_1, x_2, \ldots, x_{m-1})^\top$ and $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \ldots, \beta_{m-1}^*)^\top$:

$$\int_{R^m} \prod_{j=1}^k \frac{\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})}{[1 + \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})]^2} d\boldsymbol{\beta} = \int_{R^{m-1}} \prod_{j=1}^k \frac{\exp(\boldsymbol{x}_j^{*T} \boldsymbol{\beta}^*)}{[1 + \exp(\boldsymbol{x}_j^{*T} \boldsymbol{\beta}^*)]^2} d\boldsymbol{\beta}^* \times \int_R d\beta_m, \tag{32}$$

where $\int_R d\beta_m = \infty$.

6.4 Proof of Theorem 2

The $\boldsymbol{\beta}$ part of the upper bound for $L(r, \boldsymbol{\beta})$ in Lemma 2, i.e., $\prod_{j=1}^k p_j^E q_j^E$, is always less than one. Thus, the upper bound for $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ up to a normalizing constant factors into a function of r and a function of $\boldsymbol{\beta}$ as follows:

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y}) \propto f(r) g(\boldsymbol{\beta}) L(r, \boldsymbol{\beta}) < \frac{r^{k-(u+1)} g(\boldsymbol{\beta})}{(1+r)^k}. \tag{33}$$

The integration of this upper bound with respect to r is finite if $k \geq u+1$ because in this case we can bound the r part by $1/(1+r)^{u+1}$ whose integration with respect to r is always finite. The integration of $g(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is finite because $g(\boldsymbol{\beta})$ is a proper probability density function.

If $k < u+1$, then the integration of the lower bound for $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is not finite because there is r^k in the numerator of the lower bound for $L(r, \boldsymbol{\beta})$ in Lemma 2. Specifically, once multiplying $f(r)$ ($\propto dr/r^{u+1}$) by r^k, we know that $r^{k-(u+1)}$ goes to infinity as r approaches zero if $k < u+1$.

6.5 Proof of Theorem 3

Based on the upper bound for $L(r, \boldsymbol{\beta})$ in Lemma 2, the upper bound for $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ up to a normalizing constant factors into a function of r and a function of $\boldsymbol{\beta}$ as follows:

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y}) \propto \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) L(r, \boldsymbol{\beta}) \leq \frac{r^{k-(u+1)}}{(r+1)^k} \prod_{j=1}^k p_j^E q_j^E. \tag{34}$$

The double integration on the upper bound in (34) with respect to r and $\boldsymbol{\beta}$ is finite if and only if (i) $k \geq u+1$ for the r part as proved in Theorem 2 and (ii) the $k \times m$ covariate matrix of all groups X has a full rank m for the $\boldsymbol{\beta}$ part as proved in Theorem 1.

If at least one condition is not met, then $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ becomes improper as proved in Theorem 1 and 2.

6.6 Proof of Corollary 1

Regarding the sufficient conditions for posterior propriety, an upper bound for $L(r, \boldsymbol{\beta})$ up to a constant multiple is

$$L(r,\boldsymbol{\beta}) \propto \prod_{j=1}^{k} \frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(rp_j^E,\ rq_j^E)} < \prod_{j\in W_y} \frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(rp_j^E,\ rq_j^E)} \tag{35}$$

$$= \prod_{j\in W_y} \frac{rp_j^E q_j^E}{1+r} \frac{B(y_j + rp_j^E,\ n_j - y_j + rq_j^E)}{B(1+rp_j^E,\ 1+rq_j^E)} \leq \frac{r^{k_y} \prod_{j\in W_y} p_j^E q_j^E}{(1+r)^{k_y}}. \tag{36}$$

The inequality in (35) holds because the upper bound for the ratio of two beta functions for extreme group j is either p_j^E (< 1) in (26) or q_j^E (< 1) in (28). The inequality in (36) holds because the integrand of the beta function in the numerator is less than or equal to the integrand of the beta function in the denominator.

The upper bound for $L(r, \boldsymbol{\beta})$ in (36) would be the same as the upper bound for $L(r, \boldsymbol{\beta})$ in Lemma 2 if we removed all extreme groups from the data and treated the interior groups as a new data set ($k_y = k$). Thus, if the joint posterior density function $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is proper with the new data set of k_y interior groups based on Theorem 1, 2, or 3, then posterior propriety with the original data with all interior and all extreme groups combined ($1 \leq k_y \leq k-1$) also holds. In other words, the extreme groups do not affect the sufficient condition for posterior propriety no matter how many of them are in the data as long as there exists at least one interior group in the data.

For the necessary conditions for posterior propriety, we will show that if a new data set with all the extreme groups removed does not meet the conditions for posterior propriety based on Theorem 1, 2, or 3, then $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is still improper even after we add extreme groups into the new data.

Because a lower bound for the Beta-Binomial probability mass function for extreme group j is either $(p_j^E)^{n_j}$ in (27) or $(q_j^E)^{n_j}$ in (28), the extra product term for extreme groups to the lower bound for the likelihood function based only on interior groups is $\prod_{i\in W_y^c} (p_i^E)^{n_i I_{\{y_i = n_i\}}} (q_i^E)^{n_i I_{\{y_i = 0\}}}$.

Specifically, let us consider a proper hyper-prior PDF for r, $f(r)$, and an improper flat hyper-prior PDF for $\boldsymbol{\beta}$, $g(\boldsymbol{\beta}) \propto d\boldsymbol{\beta}$ as in Theorem 1. Suppose we removed all the extreme groups in the data. If the rank of X_y is not of full rank, e.g., $\text{rank}(X_y) = m-1$, then we see the term $\int_R d\beta_m$ in (32). This term does not disappear even after we add all the extreme groups to the data because multiplying $\prod_{i\in W_y^c} (p_i^E)^{n_i I_{\{y_i = n_i\}}} (q_i^E)^{n_i I_{\{y_i = 0\}}}$ by the first integrand in (32) cannot make the term, $\int_R d\beta_m$, disappear. It means that $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is still improper.

Next, we consider $f(r) \propto dr/r^{u+1}$ for positive u and a proper hyper-prior PDF on $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, as in Theorem 2. Because contribution of extreme groups to the lower bound for the likelihood function, i.e., $\prod_{i\in W_y^c} (p_i^E)^{n_i I_{\{y_i = n_i\}}} (q_i^E)^{n_i I_{\{y_i = 0\}}}$, is free of r, if k_y is smaller than $u+1$, then $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is still improper even after we add all the extreme groups into the data.

If the data of interior groups do not meet the condition for posterior propriety specified in Theorem 3, then adding the extreme groups cannot change the result of posterior propriety. This is because Theorem 3 is an improper mixture of Theorem 1 and 2 and we already showed that extreme groups can be ignored in determining posterior propriety in Theorem 1 and 2.

6.7 Proof of Lemma 3

A lower bound for the Beta-Binomial probability mass function of extreme group j is either $(p_j^E)^{n_j}$ in (27) or $(q_j^E)^{n_j}$ in (28) depending on whether $y_j = n_j$ or $y_j = 0$. Thus, the product of k lower bounds for the Beta-Binomial probability mass functions of extreme groups, i.e., $\prod_{j=1}^{k}(p_j^E)^{n_j \times I_{\{y_j=n_j\}}}(q_j^E)^{n_j \times I_{\{y_j=0\}}}$, bounds $L(r, \boldsymbol{\beta})$ from below.

The product of the k upper bounds for the Beta-Binomial probability mass functions of extreme groups in (26) or (28), i.e., $\prod_{j=1}^{k}(p_j^E)^{I_{\{y_j=n_j\}}}(q_j^E)^{I_{\{y_j=0\}}}$, bounds $L(r, \boldsymbol{\beta})$ from above.

6.8 Proof of Theorem 4

Considering the upper bound of the likelihood function in (14) when all groups are extreme ($k_y = 0$), the upper bound of $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ up to a constant multiple is

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y}) \propto f(r)g(\boldsymbol{\beta})L(r, \boldsymbol{\beta}) \leq f(r)\prod_{j=1}^{k}(p_j^E)^{I_{\{y_j=n_j\}}}(q_j^E)^{I_{\{y_j=0\}}}. \tag{37}$$

The integration of $f(r)$ with respect to r is finite because $f(r)$ is proper. The integration of the $\boldsymbol{\beta}$ part in (37), i.e.,

$$\prod_{j=1}^{k}(p_j^E)^{I_{\{y_j=n_j\}}}(q_j^E)^{I_{\{y_j=0\}}}, \tag{38}$$

with respect to $\boldsymbol{\beta}$ is finite if there exists a finite value of $\boldsymbol{\beta}$ that maximizes (38). This is because (38) is essentially a likelihood function of a logistic regression in (8) in that the powers in (38) are either one or zero with $I_{\{y_j=0\}} = 1 - I_{\{y_j=n_j\}}$. Thus, we can use the fact that the posterior distribution of $\boldsymbol{\beta}$ with its constant prior (Lebesgue measure) in a logistic regression is proper if there exists a finite MLE of $\boldsymbol{\beta}$ (Albert and Anderson, 1984; Speckman et al., 2009). (Jacobsen (1989) shows that the MLE of a logistic regression is unique if it exists.) Consequently, the integration of (38) with respect to $\boldsymbol{\beta}$ is finite if there exists a finite value of $\boldsymbol{\beta}$ that maximizes (38).

The lower bound of $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ up to a constant multiple can be derived from the lower bound of the likelihood function in (14), i.e.,

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y}) \propto f(r)g(\boldsymbol{\beta})L(r, \boldsymbol{\beta}) \geq f(r)\prod_{j=1}^{k}\left[(p_j^E)^{I_{\{y_j=n_j\}}}(q_j^E)^{I_{\{y_j=0\}}}\right]^{n_j}. \tag{39}$$

The integration of the $\boldsymbol{\beta}$ part in (39) with respect to $\boldsymbol{\beta}$ can be bounded from below by

$$\int_{\mathbf{R}^m} \prod_{j=1}^{k} \left[(p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \right]^{n_j} d\boldsymbol{\beta} \geq \int_{\mathbf{R}^m} \left[\prod_{j=1}^{k} (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \right]^{n_{\max}} d\boldsymbol{\beta}$$
$$\geq \left[\int_{\mathbf{R}^m} \prod_{j=1}^{k} (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} d\boldsymbol{\beta} \right]^{n_{\max}}, \tag{40}$$

where the first inequality holds because of the largest power $n_{\max} \equiv \max(n_1, n_2, \ldots, n_k)$ and the second inequality holds via Jensen's inequality because the power function is convex. The integrand in (40) is the same as (38). This indicates that the integration in (40) is not finite (and thus $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \boldsymbol{y})$ is improper) if a finite value of $\boldsymbol{\beta}$ that maximizes (38) does not exist (Albert and Anderson, 1984; Speckman et al., 2009).

6.9 Proof of Theorem 5

First, we show that the integration of (38) with respect to $\boldsymbol{\beta}$ is finite if (i) there are at least m clusters of groups whose covariate values are the same within each cluster and different between clusters, and (ii) in each cluster there are at least one group of all successes and at least one group of all failures. We define c_i as the index set of cluster i, e.g., $c_i = \{2, 5\}$ means that groups 2 and 5 are in cluster i. Then we can bound (38) with groups only in the m clusters as follows.

$$\prod_{j=1}^{k} (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \leq \prod_{j \in \{c_i, i=1,2,\ldots,m\}} (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \leq \prod_{i=1}^{m} p_{c_i}^E q_{c_i}^E, \tag{41}$$

where $p_{c_i}^E = 1 - q_{c_i}^E = \exp(\boldsymbol{x}_{c_i}^\top \boldsymbol{\beta}) / \{1 + \exp(\boldsymbol{x}_{c_i}^\top \boldsymbol{\beta})\}$ is the same expected random effect for all groups in cluster i and $\boldsymbol{x}_{c_i}$ is the same covariate vector for all groups in cluster i. The first equality holds because some groups may not be included in one of m clusters. The second inequality holds for two reasons. First, groups in the same cluster share the same covariate values, meaning that every group in cluster i has the same expected random effect, $p_{c_i}^E = 1 - q_{c_i}^E$. Second, in each cluster there are at least one group with all successes and at least one group with all failures, indicating that in cluster i, $p_{c_i}^E q_{c_i}^E$ is the largest value of $\prod_{j \in c_i} (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}}$. The integration of the upper bound in (41) is finite with a linear transformation, $h_i = \boldsymbol{x}_{c_i}^\top \boldsymbol{\beta}$, as follows:

$$\int_{R^m} \prod_{i=1}^{m} p_{c_i}^E q_{c_i}^E d\boldsymbol{\beta} \propto \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{m} \frac{\exp(h_i)}{[1 + \exp(h_i)]^2} dh_i = 1. \tag{42}$$

The last equality holds because $\exp(h_i)/[1 + \exp(h_i)]^2$ is a PDF of a standard Logistic distribution with respect to h_i.

These conditions also become necessary conditions when $\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$ for all j. In this case, the conditions simply reduce to having at least one group with all successes and at least one group with all failures. Let us use notation $p_j^E = p^E = 1 - q^E =$

$\exp(\beta_1)/(1+\exp(\beta_1))$. If all the extreme groups have only successes ($y_j = n_j$ for all j), then we can bound $\pi_{\text{hyp.post}}(r, \beta_1 \mid \boldsymbol{y})$ from below using the lower bound in (14) up to a normalizing constant as follows:

$$\pi_{\text{hyp.post}}(r, \beta_1 \mid \boldsymbol{y}) \propto f(r)g(\beta_1)L(r, \beta_1) \geq f(r)(p^E)^{\sum_{j=1}^{k} n_j}. \tag{43}$$

The integration of this lower bound in (43) with respect to β_1 is not finite because p^E converges to one as β_1 approaches infinity. Similarly, $\pi_{\text{hyp.post}}(r, \beta_1 \mid \boldsymbol{y})$ is improper if all the extreme groups have only failures ($y_j = 0$ for all j).

References

Albert, A. and Anderson, J. A. (1984). "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika*, 71(1): 1–10. MR0738319. doi: http://dx.doi.org/10.1093/biomet/71.1.1. 552, 553

Albert, J. H. (1988). "Computational Methods Using a Bayesian Hierarchical Generalized Linear Model." *Journal of the American Statistical Association*, 83(404): 1037–1044. MR0997579. 534, 535, 536, 542

Athreya, K. B. and Roy, V. (2014). "Monte Carlo Methods for Improper Target Distributions." *Electronic Journal of Statistics*, 8(2): 2664–2692. MR3292953. doi: http://dx.doi.org/10.1214/14-EJS969. 543

Christiansen, C. L. and Morris, C. N. (1997). "Hierarchical Poisson Regression Modeling." *Journal of the American Statistical Association*, 92(438): 618–632. MR1467853. doi: http://dx.doi.org/10.2307/2965709. 536, 543

Daniels, M. J. (1999). "A Prior for the Variance in Hierarchical Models." *The Canadian Journal of Statistics*, 27(3): 567–578. MR1745822. doi: http://dx.doi.org/10.2307/3316112. 534, 535, 536, 542

Efron, B. and Morris, C. N. (1975). "Data Analysis Using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association*, 70(350): 311–319. MR0391403. 543

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, 3rd edition. MR3235677. 534, 535, 536, 542

Hobert, J. P. and Casella, G. (1996). "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models." *Journal of the American Statistical Association*, 91(436): 1461–1473. MR1439086. doi: http://dx.doi.org/10.2307/2291572. 534, 543

Jacobsen, M. (1989). "Existence and Unicity of MLEs in Discrete Exponential Family Distributions." *Scandinavian Journal of Statistics*, 335–349. MR1039287. 541, 552

Kahn, M. J. and Raftery, A. E. (1996). "Discharge Rates of Medicare Stroke Patients to Skilled Nursing Facilities: Bayesian Logistic Regression With Unobserved Heterogeneity." *Journal of the American Statistical Association*, 91(433): 29–41. 533, 534, 535, 536, 542, 543

Kass, R. E. and Steffey, D. (1989). "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)." *Journal of the American Statistical Association*, 84(407): 717–726. MR1132587. 534, 542

Morris, C. N. and Lysy, M. (2012). "Shrinkage Estimation in Multilevel Normal Models." *Statistical Science*, 27(1): 115–134. MR2953499. doi: http://dx.doi.org/10.1214/11-STS363. 536

Morris, C. N. and Tang, R. (2011). "Estimating Random Effects via Adjustment for Density Maximization." *Statistical Science*, 26(2): 271–287. MR2858514. doi: http://dx.doi.org/10.1214/10-STS349. 536

Natarajan, R. and Kass, R. E. (2000). "Reference Bayesian Methods for Generalized Linear Mixed Models." *Journal of the American Statistical Association*, 95(449): 227–237. MR1803151. doi: http://dx.doi.org/10.2307/2669540. 541

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. 541

Skellam, J. G. (1948). "A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials." *Journal of the Royal Statistical Society – Series B*, 10: 257–261. MR0028539. 533, 536

Speckman, P. L., Lee, J., and Sun, D. (2009). "Existence of the MLE and Propriety of Posteriors for a General Multinomial Choice Model." *Statistica Sinica*, 731–748. MR2514185. 552, 553

Strawderman, W. E. (1971). "Proper Bayes Minimax Estimators of the Multivariate Normal Mean." *The Annals of Mathematical Statistics*, 42(1): 385–388. MR0397939. 536, 543

Tierney, L. (1994). "Markov Chains for Exploring Posterior Distributions." *The Annals of Statistics*, 22(4): 1701–1728. MR1329166. doi: http://dx.doi.org/10.1214/aos/1176325750. 544

Williams, D. A. (1982). "Extra-Binomial Variation in Logistic Linear Models." *Journal of the Royal Statistical Society – Series C*, 31(2): 144–148. MR0673714. doi: http://dx.doi.org/10.2307/2347977. 533

Acknowledgments

The authors thank Joseph Kelly for productive discussions, Steven Finch for his careful proofreading, and an editor, an associate editor, and a referee for their insightful comments that have significantly improved this paper.