Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. http://www.jstatsoft.org/

Rgbp: An R Package for Conjugate Gaussian, Poisson, and Binomial Hierarchical Modeling and Frequency Method Checking on Overdispersed Data

Hyungsuk Tak
Harvard University

Joseph Kelly
Google

Carl Morris
Harvard University

Abstract

Rgbp is an R package that utilizes approximate Bayesian machinery to fit two-level conjugate hierarchical models on overdispersed Gaussian, Poisson, and Binomial data. The data that **Rgbp** assumes comprise of observed sufficient statistics for each random effect, such as sample means, possibly together with covariates of each group. The approximate Bayesian tool equipped with the adjustment for density maximization estimates random effects and hyper-parameters in the conjugate prior distributions. As for the Binomial model, the package has an option to draw their posterior samples via the acceptance-rejection method. The main goal of **Rgb** is to produce Bayesian interval estimates for the random effects that meet their nominal confidence levels. For this purpose, we adopt unique improper hyper-prior distributions. **Rgbp** provides a tool to check quickly whether the resulting Bayesian interval estimates for the random effects achieve the nominal confidence levels via a repeated sampling coverage evaluation, which we call "frequency method checking."

Keywords: overdispersion, hierarchical model, adjustment for density maximization, acceptance-rejection method, repeated sampling coverage evaluation, frequency method checking, R.

1. Introduction (v4; 28 July 2015)

Gaussian, Poisson, or Binomial data from several independent groups sometimes have more variation than the assumed Gaussian, Poisson, or Binomial distributions of the first-level observed data. To account for this extra-variability, called overdispersion, a two-level conjugate hierarchical model regards first-level mean parameters as random effects that come from a population-level conjugate prior distribution. The main goal of our two-level conjugate modeling is to estimate these random effects for a comparison between groups.

With an assumption of homogeneity within each group, the observed data are group-level aggregate data from k independent groups, composed of sufficient statistics for their k random effects without the population-level data. Specifically, the data for the Gaussian model comprise of each group's sample mean and its standard error, those for the Poisson model use each group's outcome count and an exposure measure, those for the Binomial model use the number of each group's successful outcomes together with the total number of trials. The data analyzed by **Rgbp** may incorporate each group's covariate information. These types of data are common, e.g., as for a biological analysis on litter data (Tamura and Young 1987), a meta analysis on independent studies (Chapter 5 in Gelman, Carlin, Stern, and Rubin (2014)), or small area estimation problems (Ghosh and Rao 1994; Rao 2003).

For such data, assuming homogeneity within each group, **Rgbp**'s two-level model may be viewed as a conjugate hierarchical generalized linear model (HGLM) (Lee and Nelder 1996; Lee, Nelder, and Pawitan 2006) where each random effect has a conjugate prior distribution. However, the HGLM focuses on estimating regression coefficients to explore associations between covariates and observed data. While **Rgbp** does that too, its emphasis concerns making valid point and interval estimates for the k random effects.

Rgbp employs a Bayesian approach with our unique improper hyper-prior distributions placed on the second-level parameters of the conjugate prior distribution. The hyper-prior is chosen to produce Bayesian interval estimates for the random effects that achieve nominal confidence levels. The hyper-prior distribution used in **Rgbp** is related to Stein's harmonic prior for two-level Gaussian models where it has been seen to produce good repeated sampling coverage rates for the Bayesian interval estimates for the k random effects (Morris and Tang 2011; Morris and Lysy 2012; Kelly 2014). This prior, as adapted further here for **Rgbp**'s Poisson and Binomial hierarchical models, has been seen for every data set tested thus far to meet (or exceed) the pre-assigned confidence level.

For fitting the hierarchical model, **Rgbp** uses ADM, i.e. "adjustment for density maximization" (Morris 1988a; Christiansen and Morris 1997; Morris and Tang 2011). ADM approximates a posterior density or a likelihood function by fitting a pre-specified (one dimensional) Pearson family, using the first two derivatives of the given function. When the Normal distribution is the chosen Pearson family, ADM produces exactly the same approximation as maximum likelihood estimation (MLE) does for fitting a Normal curve. Because shrinkage factors lie in [0,1], **Rgbp** approximates shrinkage factor distributions with Beta distributions because they are constrained to [0,1], unlike the Normal.

Rgbp estimates the first two (actually, three in the Gaussian case) posterior moments of each random effect. Finally, **Rgbp** approximates the posterior distribution of each random effect by a skewed Normal distribution for Gaussian case, by a Gamma distribution for the Poisson case, and by a Beta distribution for the Binomial case.

For the Binomial hierarchical model, **Rgbp** provides an option to draw independent posterior samples of all the model parameters, including random effects, via an acceptance-rejection method (Everson and Morris 2000; Tang 2002).

In addition to fitting the hierarchical models, **Rgbp** evaluates the repeated sampling coverage rates of the resulting Bayesian interval estimates for random effects (Christiansen and Morris 1997; Daniels 1999; Tang 2002; Morris and Tang 2011; Morris and Lysy 2012). This is a unique procedure that distinguishes **Rgbp** from any other R packages for similar hierarchical modeling such as **hglm** (Rönnegård, Shen, and Alam 2010, 2011) for conjugate hierarchical

generalized models. The evaluation procedure which we use for this process of "frequency method checking" uses a parametric bootstrapping method that many times generates mock data sets given the fitted values of the estimated hyper-parameters and uses the mock data to estimate the coverage rates. Using this procedure, we show that the coverage rates of the random effects obtained by our model achieve (or exceed) the nominal confidence level.

The rest of this paper is organized as follows. We specify the Bayesian hierarchical models and discuss their posterior propriety in Section 2. In Section 3, we explain the inferential models used to estimate the model parameters. We describe the estimation procedures including ADM and the acceptance-rejection method in Section 4 and 5, respectively. We introduce frequency method checking techniques in Section 6. We explain the usages of **Rgbp**'s main functions in **Rgbp** in Section 7, and apply them to three data examples in Section 8.

2. Conjugate hierarchical modeling structure

Rgbp allows users to choose one of three hierarchical models according to the type of data, namely Normal-Normal, Poisson-Gamma, and Binomial-Beta models. Although there are more hierarchical models, we choose the three models because these are based on the most common types of data we may encounter in practice. Also, their conjugacy leads to linear posterior means simplifying computations.

Our parametrization of the three hierarchical models leads to an intuitive shrinkage interpretation in inference and facilitates the estimation procedure because the shrinkage factors under our parametrization are a function of only a second-level variance component.

2.1. Normal-Normal model for Gaussian data

The following Normal-Normal hierarchical model (or the Gaussian model) assumed by **Rgbp** is useful when the group-level aggregate data from k independent groups are continuous (or approximately continuous) variables with known standard errors. The subscript j below indicates the jth group among k groups in the dataset. For $j = 1, 2, \ldots, k$,

$$y_j|\mu_j \overset{\text{indep.}}{\sim} \text{Normal}(\mu_j, V_j), \tag{1}$$

$$\mu_j|\boldsymbol{\beta}, A \overset{\text{indep.}}{\sim} \text{Normal}(\mu_j^E, A), \tag{2}$$

where y_j is an observed unbiased estimate, e.g., sample mean, for random effect μ_j, V_j is a completely known standard error of y_j, μ_j^E is an expected random effect defined as $\mathsf{E}(\mu_j|\boldsymbol{\beta}, A) = \boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1 x_{j,1} + \beta_2 x_{j,2} + \cdots + \beta_m x_{j,m}$, and m is the number of regression coefficients to be estimated. **Rgbp** sets $x_{j,1}$ to 1 for an intercept term as a default. It also provides a usage without the intercept term. It is assumed that the second-level variance A is unknown and that the $m \times 1$ regression coefficient vector $\boldsymbol{\beta}$ is also unknown unless otherwise specified. If no covariates are available, but with an unknown intercept term, then $\boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1$ $(m = 1)$ and thus $\mu_j^E = \mu^E = \beta_1$ for all j, resulting in an exchangeable conjugate prior distribution for the random effects. Based on these conjugate prior distributions for random effects, it is easy to derive the conditional posterior distribution of each random effect. For $j = 1, 2, \ldots, k$,

$$\mu_j|\boldsymbol{\beta}, A, \boldsymbol{y} \overset{\text{indep.}}{\sim} \text{Normal}(\ (1 - B_j)y_j + B_j\mu_j^E,\ (1 - B_j)V_j\ ), \tag{3}$$

where $B_j \equiv V_j/(V_j + A)$ is a shrinkage factor of group j and $\boldsymbol{y} = (y_1, y_2, \ldots, y_k)^\top$. Note that the conditional posterior mean $\mathsf{E}(\mu_j|\boldsymbol{\beta}, A, \boldsymbol{y})$, denoted by μ_j^*, is a convex combination of the observed sample mean y_j and the expected random effect μ_j^E weighted by the shrinkage factor B_j. If the variance of the conjugate prior distribution, A, is smaller than the variance of the observed distribution, V_j, then we expect the posterior mean to borrow more information from the second-level conjugate prior distribution.

2.2. Poisson-Gamma model for Poisson data

Rgbp can estimate a conjugate Poisson-Gamma hierarchical model (or the Poisson model) when the group-level aggregate data from k independent groups are comprised of non-negative count data without upper limit. However, in this case its usage is limited to the case where the expected random effect is known (or equivalently all the regression coefficients in $\lambda_j^E = \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})$ is known ($m = 0$)); we may be able to obtain this information from the past studies or from experts. For $j = 1, 2, \ldots, k$,

$$y_j|\lambda_j \overset{\text{indep.}}{\sim} \text{Poisson}(n_j\lambda_j), \tag{4}$$

$$\lambda_j|r \overset{i.i.d.}{\sim} \text{Gamma}(r\lambda^E, r), \tag{5}$$

where y_j is the number of events happening, n_j is the exposure of group j, which is not necessarily an integer, $\lambda^E = \mathsf{E}(\lambda_j|r)$ is the known expected random effect ($m = 0$), and r is the unknown second-level variance component. The mean and variance of this conjugate Gamma prior distribution are λ^E and λ^E/r, respectively. Albert (1988) interprets r as the amount of prior information as n_j represents the amount of observed information because the uncertainty of the conjugate prior distribution increases as r decreases and vice versa. The conditional posterior distribution of the random effect λ_j for this Poisson model is

$$\lambda_j|r, \boldsymbol{y} \overset{\text{indep.}}{\sim} \text{Gamma}(r\lambda^E + n_j\bar{y}_j,\ r + n_j), \tag{6}$$

where $\bar{y}_j \equiv y_j/n_j$. The mean and variance of the conditional posterior distribution are

$$\lambda_j^* \equiv \mathsf{E}(\lambda_j|r, \boldsymbol{y}) = (1 - B_j)\bar{y}_j + B_j\lambda^E \quad \text{and} \quad \mathsf{VAR}(\lambda_j|r, \boldsymbol{y}) = \frac{\lambda_j^*}{r + n_j}. \tag{7}$$

where $B_j \equiv r/(r + n_j)$ is the shrinkage factor of group j, the relative amount of information in the prior compared to the data. The conditional posterior mean is a convex combination of $\bar{y}_j$ and λ^E weighted by B_j. If the conjugate prior distribution contains more information than the observed data have, *i.e.*, ensemble sample size r exceeds individual sample size n_j, then the posterior mean shrinks towards the prior mean by more than 50%.

Note that the conditional posterior variance in (7) is linear in the conditional posterior mean, whereas a slightly different Poisson-Gamma model specification has been used elsewhere (Christiansen and Morris 1997) that makes the variances quadratic functions of means.

2.3. Binomial-Beta model for Binomial data

Rgbp can fit a conjugate Binomial-Beta hierarchical model when the group-level aggregate data from k independent groups are composed of each group's number of successes out of

total number of trials. The expected random effect in the Binomial model is either known ($m = 0$) or unknown ($m \geq 1$). For $j = 1, 2, \ldots, k$,

$$y_j | p_j \overset{\text{indep.}}{\sim} \text{Binomial}(n_j,\ p_j), \tag{8}$$

$$p_j | \boldsymbol{\beta}, r \overset{\text{indep.}}{\sim} \text{Beta}(r p_j^E,\ r(1 - p_j^E)), \tag{9}$$

where y_j is the number of successes out of n_j trials, p_j^E is the expected random effect of group j defined as $p_j^E \equiv \mathsf{E}(p_j | \boldsymbol{\beta}, r) = \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}))$. The $m \times 1$ vector of the logistic regression coefficient $\boldsymbol{\beta}$ and the second-level variance component r are unknown. The mean and variance of the conjugate Beta prior distribution for group j are p_j^E and $p_j^E(1-p_j^E)/(r+1)$, respectively. The resultant conditional posterior distribution of random effect p_j is

$$p_j | \boldsymbol{\beta}, r, \boldsymbol{y} \overset{\text{indep.}}{\sim} \text{Beta}(n_j \bar{y}_j + r p_j^E,\ n_j(1 - \bar{y}_j) + r(1 - p_j^E)), \tag{10}$$

where $\bar{y}_j = y_j / n_j$ is the observed proportion of group j. The mean and variance of the conditional posterior distribution are

$$p_j^* \equiv \mathsf{E}(p_j | \boldsymbol{\beta}, r, \boldsymbol{y}) = (1 - B_j)\bar{y}_j + B_j p_j^E \quad \text{and} \quad \mathsf{VAR}(p_j | \boldsymbol{\beta}, r, \boldsymbol{y}) = \frac{p_j^*(1 - p_j^*)}{r + n_j + 1}. \tag{11}$$

The conditional posterior mean p_j^* is a convex combination of $\bar{y}_j$ and p_j^E weighted by $B_j \equiv r/(r+n_j)$ like the Poisson model. If the conjugate prior distribution contains more information than the observed distribution does ($r > n_j$), then the resulting conditional posterior mean borrows more information from the conjugate Beta prior distribution.

2.4. Hyper-prior distribution

Hyper-prior distributions are the distributions assigned to the second-level parameters called hyper-parameters. Our choices for the hyper-prior distributions are

$$\boldsymbol{\beta} \sim \text{Uniform on } \mathbf{R}^m \quad \text{and} \quad A \sim \text{Uniform}(0, \infty) \quad \left(\text{or } \frac{1}{r} \sim \text{Uniform}(0, \infty)\right). \tag{12}$$

The improper flat hyper-prior distribution on $\boldsymbol{\beta}$ is a common non-informative choice. In the Gaussian case, the flat hyper-prior distribution on the second-level variance A is known to produce good repeated sampling coverage properties of the Bayesian interval estimates for the random effects (Morris and Tang 2011; Morris and Lysy 2012; Kelly 2014). The resulting full posterior distribution of the random effects and hyper-parameters is proper if $k \geq m + 3$ (Morris and Tang 2011; Kelly 2014).

In the other two cases, Poisson and Binomial, the flat prior distribution on $1/r$ induces the same improper prior distribution on shrinkages ($\pi(B_j) \propto B_j^{-2} dB_j$) as does A with the Uniform$(0, \infty)$ for the Gaussian case. The Poisson model with this hyper-prior distribution on r, i.e., dr/r^2, provides posterior propriety if there are at least two groups whose observed values y_j are non-zero and the expected random effect, λ^E, is a completely known constant ($m = 0$); see Appendix A for its proof. If λ^E is unknown, we recommend using the Binomial model with the same hyper-prior distributions because the Poisson model is an approximation to the Binomial model when n_j is large and p_j is small.

As for posterior propriety of the Binomial model, let's define an "interior group" as the group whose number of successes y_j are neither 0 nor n_j, and k_y as the number of interior groups among k groups. The full posterior distribution of random effects and hyper-parameters is proper if and only if there are at least two interior groups in the data and the $k_y \times m$ covariate matrix of the interior groups is of full rank m ($k_y \geq m$) (Tak and Morris in preparation).

3. The inferential model

The likelihood function of hyper-parameters, A and $\boldsymbol{\beta}$, for the Gaussian model is derived from the independent Normal distributions of the observed data with random effects integrated out;

$$L(A, \boldsymbol{\beta}) = \prod_{j=1}^{k} f(y_j|A, \boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi(A + V_j)}} \exp\left(-\frac{(y_j - \mu_j^E)^2}{2(A + V_j)}\right), \tag{13}$$

where $\mu_j^E = \boldsymbol{x}^\top \boldsymbol{\beta}$. The joint posterior density function of hyper-parameters $f(A, \boldsymbol{\beta}|\boldsymbol{y})$ for the Gaussian model is proportional to their likelihood function in (13) because we use flat improper hyper-prior density functions for A and $\boldsymbol{\beta}$;

$$f(A, \boldsymbol{\beta}|\boldsymbol{y}) \propto L(A, \boldsymbol{\beta}) dA d\boldsymbol{\beta}. \tag{14}$$

The likelihood function of hyper-parameters r and $\boldsymbol{\beta}$ for the Binomial model is derived from the independent Beta-Binomial distributions of the observed data with random effects integrated out (Skellam 1948);

$$L(r, \boldsymbol{\beta}) = \prod_{j=1}^{k} f(y_j|r, \boldsymbol{\beta}) = \prod_{j=1}^{k} \binom{n_j}{y_j} \frac{B(y_j + rp_j^E,\ n_j - y_j + r(1 - p_j^E))}{B(rp_j^E,\ r(1 - p_j^E))}, \tag{15}$$

where the notation $B(a, b)$ ($\equiv \int_0^1 v^{a-1}(1 - v)^{b-1} dv$) indicates a beta function for positive constants a and b. The joint posterior density function of hyper-parameters $f(r, \boldsymbol{\beta}|\boldsymbol{y})$ for the Binomial model is proportional to their likelihood function in (15) multiplied by the hyper-prior density functions of r and $\boldsymbol{\beta}$ in (12);

$$f(r, \boldsymbol{\beta}|\boldsymbol{y}) \propto L(r, \boldsymbol{\beta}) d\boldsymbol{\beta} dr/r^2. \tag{16}$$

Similarly, the likelihood function of r for the Poisson model comes from the independent Negative-Binomial distributions of the observed data with the random effects integrated out;

$$L(r) = \prod_{j=1}^{k} f(y_j|r) = \prod_{j=1}^{k} \frac{\Gamma(r\lambda^E + y_j)}{\Gamma(r\lambda^E)(y_j!)} (1 - B_j)^{y_i} B_j^{r\lambda^E}, \tag{17}$$

where $\Gamma(a)$ is a gamma function defined as $\int_0^\infty x^{a-1} \exp(x) dx$ for a positive constant a. The posterior density function of r, $f(r|\boldsymbol{y})$, for the Poisson hierarchical model is the likelihood function in (17) times the hyper-prior density function of r, i.e., dr/r^2;

$$f(r|\boldsymbol{y}) \propto L_p(r) dr/r^2. \tag{18}$$

Our goal is to obtain the point and interval estimates of the random effects from their unconditional posterior distributions. With subscripts suppressed, the strategy is to compute by using these formulas: for the Gaussian model,

$$f(\boldsymbol{\mu}|\boldsymbol{y}) = \int f(\boldsymbol{\mu}|A, \boldsymbol{\beta}, \boldsymbol{y}) \cdot f(A, \boldsymbol{\beta}|\boldsymbol{y}) dA d\boldsymbol{\beta}, \tag{19}$$

for the Binomial model,

$$f(\boldsymbol{p}|\boldsymbol{y}) = \int f(\boldsymbol{p}|r, \boldsymbol{\beta}, \boldsymbol{y}) \cdot f(r, \boldsymbol{\beta}|\boldsymbol{y}) dr d\boldsymbol{\beta}, \tag{20}$$

and lastly for the Poisson model,

$$f(\boldsymbol{\lambda}|\boldsymbol{y}) = \int f(\boldsymbol{\lambda}|r, \boldsymbol{y}) \cdot f(r|\boldsymbol{y}) dr. \tag{21}$$

4. Estimation via the adjustment for density maximization

We illustrate our estimation procedure which utilizes adjustment for density maximization (ADM) (Morris 1988a; Christiansen and Morris 1997; Morris and Tang 2011). ADM is a method to approximate a distribution by a member of Pearson family of distributions and obtain moment estimates via maximization. The ADM procedure for the Gaussian model adopted in **Rgbp** is well documented in Kelly (2014) and so in this section we will describe the ADM procedure using the Poisson and Binomial model.

4.1. Estimation for shrinkage factors and expected random effects

Our goal here is to estimate the unconditional posterior moments of the shrinkage factors and the expected random effects because they are used to estimate the unconditional posterior moments of the random effects.

Unconditional posterior moments of shrinkage factors

It is noted that the shrinkage factors are a function of r, i.e., $B_j = B_j(r) = r/(r + n_j)$ (or a function of A for the Gaussian model). A common method of estimation of B_j is to approximate the likelihood of r with two derivatives and use a Delta method for an asymptotic Normal distribution of $\hat{B}_j(\hat{r}_{MLE})$. This Normal approximation, however, is defined on $(-\infty, \infty)$ whereas B_j lies on the unit interval between 0 and 1, and hence in small sample sizes the Delta method can result in point estimates lying on the boundary of the parameter space, from which the restricted MLE procedure sometimes suffers (Morris and Tang 2011; Kelly 2014).

To continue with a maximization-based estimation procedure but to steer clear of aforementioned boundary issues we make use of ADM. The ADM approximates the distribution of the function of the parameter of interest by one of the Pearson family distributions using the first two derivatives as the Delta method does; the Delta method is a special case of the ADM based on the Normal distribution.

The ADM procedure specified in Morris and Tang (2011) assumes that the unconditional posterior distribution of a shrinkage factor follows a Beta distribution; for $j = 1, 2, \ldots, k$,

$$B_j|\boldsymbol{y} \sim \text{Beta}(a_{1j},\ a_{0j}). \tag{22}$$

Note that the mean of Beta distribution $a_{1j}/(a_{1j} + a_{0j})$ is not the same as its mode $(a_{j1} - 1)/(a_{j1} + a_{j0} - 2)$. The ADM works on an adjusted posterior distribution $f^A(B_j|\boldsymbol{y})dB_j \propto B_j(1 - B_j)f(B_j|\boldsymbol{y})dB_j$ so that the mode of $f^A(B_j|\boldsymbol{y})$ is the same as the mean of the original Beta distribution. The assumed posterior mean and variance of the jth shrinkage factor are

$$\mathsf{E}(B_j|\boldsymbol{y}) = \frac{a_{1j}}{a_{1j} + a_{0j}} = \arg\max_{B_j}\ f^A(B_j|\boldsymbol{y}) \equiv B'_j, \tag{23}$$

$$\mathsf{VAR}(B_j|\boldsymbol{y}) = \frac{B'_j(1 - B'_j)}{a_{1j} + a_{0j} + 1} = \frac{B'_j(1 - B'_j)}{B'_j(1 - B'_j)[-\frac{d^2}{dB_j^2}\log(A(B_j|\boldsymbol{y}))|_{B_j=B'_j}] + 1}. \tag{24}$$

The ADM estimates these mean and variance using the marginal posterior distribution of r, $f(r|\boldsymbol{y}) \propto L(r)dr/r^2$, where the marginal likelihood, $L(r) = \int L(\boldsymbol{\beta}, r)d\boldsymbol{\beta}$, for the Binomial model is obtained via the Laplace approximation with a Lebesque measure on $\boldsymbol{\beta}$ and that for the Poisson model is specified in (17).

Considering that (23) and (24) involve the maximization and Hessian calculation, we work on a logarithmic scale of r, i.e., $\alpha = -\log(r)$ (or $\alpha = \log(A)$ for the Gaussian model), because the distribution of α is more symmetric than that of r, and α is defined on a real line without any boundary issues. Because $f^A(B_j|\boldsymbol{y})$ is proportional to the marginal posterior density $f(\alpha|\boldsymbol{y}) \propto \exp(\alpha)L(\alpha)d\alpha$ (Morris and Tang 2011), the posterior mean in (23) is estimated by

$$\hat{B}'_j = \frac{\exp(-\hat{\alpha})}{n_j + \exp(-\hat{\alpha})}, \tag{25}$$

where $\hat{\alpha}$ is the mode of $f(\alpha|\boldsymbol{y})$, *i.e.*, $\arg\max_\alpha\{\alpha + \log(L(\alpha))\}$.

We need invariance information introduced in Morris and Tang (2011) to estimate the variance in (24), which is defined as

$$\begin{aligned} I_{\text{inv}} &\equiv -\frac{d^2\log(f^A(B_j|\boldsymbol{y}))}{d[\text{logit}(B_j)]^2}\bigg|_{B_j=\hat{B}'_j} = -\frac{d^2\log(f^A(B_j(r)|\boldsymbol{y}))}{d[\log(r)]^2}\bigg|_{r=\hat{r}} \\ &= -\frac{d^2\log(f^A(B_j(r(\alpha))|\boldsymbol{y}))}{d\alpha^2}\bigg|_{\alpha=\hat{\alpha}} \end{aligned} \tag{26}$$

Note that this invariance information is the negative Hessian value of $\alpha + \log(L(\alpha))$ at the mode $\hat{\alpha}$. Using the invariance information, we estimate the unconditional posterior variance of shrinkage factor in (24) by

$$\widehat{\mathsf{VAR}}(B_j|\boldsymbol{y}) = \frac{(\hat{B}'_j)^2(1 - \hat{B}'_j)^2}{I_{\text{inv}} + \hat{B}'_j(1 - \hat{B}'_j)}. \tag{27}$$

We obtain the estimates of a_{1j} and a_{0j}, the two parameters of the Beta distribution in (22), by matching them to the estimated unconditional posterior mean and variance of B_j specified in (25) and (27);

$$\hat{a}_{1j} = \frac{I_{\text{inv}}}{1 - \hat{B}'_j} \quad \text{and} \quad \hat{a}_{0j} = \frac{I_{\text{inv}}}{\hat{B}'_j}. \tag{28}$$

The moments of the Beta distribution are well defined as a function of a_{1j} and a_{0j}, i.e., $\mathsf{E}(B_j^c|\boldsymbol{y}) = B(a_{1j}+c, a_{0j})/B(a_{1j}, a_{0j})$ for $c \geq 0$. Their estimates are

$$\widehat{\mathsf{E}}(B_j^c|\boldsymbol{y}) = \frac{B(\hat{a}_{1j}+c,\ \hat{a}_{0j})}{B(\hat{a}_{1j},\ \hat{a}_{0j})}. \tag{29}$$

The ADM approximation to the shrinkage factors via Beta distributions is empirically proven to be more accurate than a Laplace approximation (Morris 1988a; Christiansen and Morris 1997; Morris and Tang 2011; Morris and Lysy 2012).

Unconditional posterior moments of expected random effects

We estimate the unconditional posterior moments of expected random effects using their relationship to the conditional posterior moments. For a non-negative constant c, the unconditional posterior moments are

$$\mathsf{E}((p_j^E)^c|\boldsymbol{y}) = \mathsf{E}(\mathsf{E}((p_j^E)^c|r, \boldsymbol{y})|\boldsymbol{y}). \tag{30}$$

Instead of averaging over r, we approximate the unconditional posterior moments by the conditional posterior moments with $\hat{r} = \exp(-\hat{\alpha})$ plugged-in (Kass and Steffey 1989).

However, calculating conditional posterior moments of each expected random effect involves an intractable integration. For example, the first conditional posterior moment of p_j^E is

$$\mathsf{E}(p_j^E|\hat{\alpha}, \boldsymbol{y}) = E\left(\frac{\exp(x_j^\top\boldsymbol{\beta})}{1+\exp(x_j^\top\boldsymbol{\beta})}\middle|\hat{\alpha}, \boldsymbol{y}\right) = \int_{\mathbf{R}^m} \frac{\exp(x_j^\top\boldsymbol{\beta})}{1+\exp(x_j^\top\boldsymbol{\beta})} f(\beta|\hat{\alpha}, \boldsymbol{y}) d\boldsymbol{\beta}. \tag{31}$$

Thus, we use another ADM, assuming the conditional posterior distribution of each expected random effect is a Beta distribution as follows;

$$p_j^E|\hat{\alpha}, \boldsymbol{y} = \frac{\exp(x_j^\top\boldsymbol{\beta})}{1+\exp(x_j^\top\boldsymbol{\beta})}\middle|\hat{\alpha}, \boldsymbol{y} \sim \text{Beta}(b_{1j},\ b_{0j}) \sim \frac{G(b_{1j})}{G(b_{1j})+G(b_{0j})}, \tag{32}$$

where $G(b_{1j})$ is a random variable following a Gamma$(b_{1j}, 1)$ distribution and independently $G(b_{0j})$ has a Gamma$(b_{0j}, 1)$ distribution. Note that the representation in (32) is equivalent to saying $\exp(x_j^\top\boldsymbol{\beta})|\hat{\alpha}, \boldsymbol{y} \sim G(b_{1j})/G(b_{0j})$, a ratio of two independent Gamma random variables. Its mean and variance are

$$\mathsf{E}(\exp(x_j^\top\boldsymbol{\beta})|\hat{\alpha}, \boldsymbol{y}) = E\left(\frac{G(b_{1j})}{G(b_{0j})}\right) = \frac{b_{1j}}{b_{0j}-1} \equiv \eta_j, \tag{33}$$

$$\mathsf{VAR}(\exp(x_j^\top\boldsymbol{\beta})|\hat{\alpha}, \boldsymbol{y}) = \mathsf{VAR}\left(\frac{G(b_{1j})}{G(b_{0j})}\right) = \frac{\eta_j(1+\eta_j)}{b_{0j}-2}. \tag{34}$$

In order to estimate b_{1j} and b_{0j}, we assume that the conditional posterior distribution of β given $\hat{\alpha}$ and $\boldsymbol{y}$ follows a Normal distribution with mean $\hat{\boldsymbol{\beta}}$ and variance-covariance matrix $\hat{\Sigma}$, where $\hat{\boldsymbol{\beta}}$ is the mode of $f(\boldsymbol{\beta}|\hat{\alpha}, \boldsymbol{y})$ and $\hat{\Sigma}$ is an inverse of the negative Hessian matrix at the mode. Thus, the posterior distribution of $x_j^\top\boldsymbol{\beta}$ is also Normal with mean $x_j^\top\hat{\boldsymbol{\beta}}$ and variance $x_j^\top\hat{\Sigma}x_j$.

Using the property of the log-Normal distribution for $\exp(x_j^\top \boldsymbol{\beta})$, we estimate the posterior mean and variance in (33) and (34) as

$$\hat{\eta}_j = \exp(x_j^\top \hat{\boldsymbol{\beta}} + x_j^\top \hat{\Sigma} x_j/2), \tag{35}$$

$$\widehat{\mathsf{VAR}}(\exp(x_j^\top \boldsymbol{\beta})|\boldsymbol{y}) = \hat{\eta}_j^2\big(\exp(x_j^T \hat{\Sigma} x_j) - 1\big). \tag{36}$$

We estimate the values of b_{1j} and b_{0j} by matching them to the estimated unconditional posterior mean and variance of $\exp(x_j^\top \boldsymbol{\beta})$ in (35) and (36);

$$\hat{b}_{1j} = \hat{\eta}_j(\hat{b}_{0j} - 1) \quad \text{and} \quad \hat{b}_{0j} = \frac{1 + \hat{\eta}_j}{\hat{\eta}_j(\exp(x_j^T \hat{\Sigma} x_j) - 1)} + 2. \tag{37}$$

Finally, we estimate the unconditional posterior moments of the expected random effects, $\mathsf{E}((p_j^E)^c|\boldsymbol{y})$, by $\hat{\mathsf{E}}((p_j^E)^c|\hat{\alpha}, \boldsymbol{y}) = B(\hat{b}_{1j} + c, \hat{b}_{0j})/B(\hat{b}_{1j}, \hat{b}_{0j})$ for $c \geq 0$.

The ADM approximation to a log-Normal density via a F distribution (represented by a ratio of two independent Gamma random variables) is known to be more accurate than the Laplace approximation (Morris 1988a).

For the Gaussian model (Morris and Tang 2011), the conditional posterior distribution of $\boldsymbol{\beta}$ given $\hat{A}$ and $\boldsymbol{y}$ is Normal whose mean and variance-covariance matrix are

$$(X^\top D_{V+\hat{A}}^{-1} X)^{-1} X^\top D_{V+\hat{A}}^{-1} \boldsymbol{y} \quad \text{and} \quad (X^\top D_{V+\hat{A}}^{-1} X)^{-1}, \tag{38}$$

respectively, where $X \equiv (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)^\top$ is a $k \times m$ covariate matrix and $D_{V+\hat{A}}$ is a $k \times k$ diagonal matrix with the j-th diagonal element equal to $V_j + \hat{A}$. Because $\boldsymbol{x}^\top \boldsymbol{\beta}$ given $\hat{A}$ and $\boldsymbol{y}$ is also Normally distributed, we easily obtain the conditional posterior moments of $\mu_j^E = \boldsymbol{x}^\top \boldsymbol{\beta}$ and use them to estimate unconditional posterior moments of μ_j^E.

4.2. Estimation for random effects

It is intractable to derive analytically the unconditional posterior distribution of each random effect for the three models. This motivates us to use the ADM again by assuming that each random effect has a skewed Normal distribution (Azzalini 1985) for the Gaussian model, a Beta distribution for the Binomial model, or a Gamma distribution for the Poisson model; for $j = 1, 2, \ldots, k$,

$$\mu_j|\boldsymbol{y} \sim \text{skewed-Normal}(\phi,\ \omega,\ \delta), \tag{39}$$

$$p_j|\boldsymbol{y} \sim \text{Beta}(t_{1j},\ t_{0j}), \tag{40}$$

$$\lambda_j|\boldsymbol{y} \sim \text{Gamma}(s_{1j},\ s_{0j}), \tag{41}$$

where (ϕ, ω, δ) of the skewed-Normal distribution are location, scale, and skewness parameters, respectively.

Morris and Lysy (2012) first note that the unconditional posterior distribution of the random effect in a two-level conjugate Gaussian model may be skewed. Kelly (2014) shows that the skewed-Normal approximation to the unconditional posterior distribution of the random effect is better than a Normal approximation ($\mu_j|\boldsymbol{y} \sim$ Normal) in terms of the repeated sampling coverage properties of random effects. Kelly (2014) uses the results derived in Azzalini (1985), i.e., $\mathsf{E}(\mu_j|\boldsymbol{y}) = \phi + \omega\delta\sqrt{2/\pi}$, $\text{Var}(\mu_j|\boldsymbol{y}) = \omega^2(1 - 2\delta^2/\pi)$, and

Skewness$(\mu_j|\boldsymbol{y}) = (4-\pi)\delta^3/[2(\pi/2-\delta^2)^{3/2}]$. He estimates three parameters of the skewed-Normal distribution using three approximately estimated unconditional posterior moments of each random effect; see Kelly (2014) for details.

To estimate the unconditional posterior moments of random effects in the Binomial model, we assume that hyper-parameters r and $\boldsymbol{\beta}$ are independent *a posteriori*. With this assumption, the unconditional posterior mean and variance of random effect p_j are

$$\begin{aligned}
\mathsf{E}(p_j|\boldsymbol{y}) &= \mathsf{E}(\mathsf{E}(p_j|r,\boldsymbol{\beta},\boldsymbol{y})|\boldsymbol{y}) = (1-\mathsf{E}(B_j|\boldsymbol{y}))\bar{y}_j + \mathsf{E}(B_j|\boldsymbol{y})\mathsf{E}(p_j^E|\boldsymbol{y}), && (42)\\
\mathsf{VAR}(p_j|\boldsymbol{y}) &= \mathsf{E}(\mathsf{VAR}(p_j|r,\boldsymbol{\beta},\boldsymbol{y})|\boldsymbol{y}) + \mathsf{VAR}(\mathsf{E}(p_j|r,\boldsymbol{\beta},\boldsymbol{y})|\boldsymbol{y}) && (43)\\
&= \mathsf{E}\big(p_j^*(1-p_j^*)/(r+n_j+1)|\boldsymbol{y}\big) + \mathsf{VAR}\big(B_j(\bar{y}_j - p_j^E)|\boldsymbol{y}\big) && (44)\\
&\approx \mathsf{E}\big(p_j^*(1-p_j^*)(1-B_j)/n_i|\boldsymbol{y}\big) + \mathsf{VAR}\big(B_j(\bar{y}_j - p_j^E)|\boldsymbol{y}\big) && (45)\\
&= \big\{(1-\bar{y}_j)\bar{y}_j[1-\mathsf{E}(B_j|\boldsymbol{y})] + (2\bar{y}_j-1)\mathsf{E}(B_j(1-B_j)|\boldsymbol{y})(\bar{y}_j - \mathsf{E}(p_j^E|\boldsymbol{y}))\\
&\quad + \mathsf{E}(B_j^2(1-B_j)|\boldsymbol{y})\mathsf{E}((\bar{y}_j-p_j^E)^2|\boldsymbol{y})\big\}/n_j + \mathsf{VAR}\big(B_j(\bar{y}_j-p_j^E)|\boldsymbol{y}\big), && (46)
\end{aligned}$$

where the approximation in (45) is a first-order Taylor approximation. By plugging-in the estimated unconditional posterior moments of shrinkage factors and those of expected random effect into (42) and (46), we obtain the estimates of the unconditional posterior mean and variance of each random effect, denoted by $\hat{\mu}_{p_j}$ and $\hat{\sigma}^2_{p_j}$, respectively. Finally, we obtain the estimates of two parameters t_{1j} and t_{0j} in (40) as follows;

$$\hat{t}_{1j} = \left(\frac{\hat{\mu}_{p_j}(1-\hat{\mu}_{p_j})}{\hat{\sigma}^2_{p_j}} - 1\right)\hat{\mu}_{p_j}, \text{ and } \hat{t}_{0j} = \left(\frac{\hat{\mu}_{p_j}(1-\hat{\mu}_{p_j})}{\hat{\sigma}^2_{p_j}} - 1\right)(1-\hat{\mu}_{p_j}). \tag{47}$$

The unconditional posterior mean and variance of random effect λ_j in the Poisson model are

$$\begin{aligned}
\mathsf{E}(\lambda_j|\boldsymbol{y}) &= \mathsf{E}(\mathsf{E}(\lambda_j|r,\boldsymbol{y})|\boldsymbol{y}) = (1-\mathsf{E}(B_j|\boldsymbol{y}))\bar{y}_j + \mathsf{E}(B_j|\boldsymbol{y})\lambda^E, && (48)\\
\mathsf{VAR}(\lambda_j|\boldsymbol{y}) &= \mathsf{E}(\mathsf{VAR}(\lambda_j|r,\boldsymbol{y})|\boldsymbol{y}) + \mathsf{VAR}(\mathsf{E}(\lambda_j|r,\boldsymbol{y})|\boldsymbol{y}) && (49)\\
&= \mathsf{E}\big(\lambda_j^*/(r+n_j)|\boldsymbol{y}\big) + \mathsf{VAR}\big(B_j(\bar{y}_j-\lambda_j^E)|\boldsymbol{y}\big) && (50)\\
&= \frac{1}{n_j}\left[\bar{y}_j\mathsf{E}\left((1-B_j)^2|\boldsymbol{y}\right) + \lambda^E\mathsf{E}\left((1-B_j)B_j|\boldsymbol{y}\right)\right] + (\bar{y}_j-\lambda_j^E)^2\mathsf{VAR}\big(B_j|\boldsymbol{y}\big). && (51)
\end{aligned}$$

Let $\hat{\mu}_{\lambda_j}$ and $\hat{\sigma}^2_{\lambda_j}$ denote the estimated unconditional posterior mean and variance in (48) and (51), respectively, with the estimated unconditional posterior moments of shrinkage factors plugged-in. The estimates of the two parameters s_{1j} and s_{0j} in (41) are

$$\hat{s}_{1j} = \frac{\hat{\mu}^2_{\lambda_j}}{\hat{\sigma}^2_{\lambda_j}}, \text{ and } \hat{s}_{0j} = \frac{\hat{\mu}_{\lambda_j}}{\hat{\sigma}^2_{\lambda_j}}. \tag{52}$$

Finally, the assumed unconditional posterior distribution of random effect for the Gaussian model is

$$\mu_j|\boldsymbol{y} \sim \text{skewed-Normal}(\hat{\phi},\hat{\omega},\hat{\delta}), \tag{53}$$

that for the Binomial model is

$$p_j|\boldsymbol{y} \sim \text{Beta}(\hat{t}_{1j},\hat{t}_{0j}), \tag{54}$$

and that for the Poisson model is

$$\lambda_j|\boldsymbol{y} \sim \text{Gamma}(\hat{s}_{1j},\hat{s}_{0j}). \tag{55}$$

Our point and interval estimates of each random effect are the mean and (2.5%, 97.5%) quantiles (if we assign 95% confidence level) of the assumed unconditional posterior distribution in (53), (54), or (55) depending on the type of data.

5. The acceptance-rejection method for the Binomial model

As for the Binomial model, the package **Rgbp** also provides a way to draw posterior samples of random effects and hyper-parameters via the acceptance-rejection (A-R) method. Unlike the approximate Bayesian machinery specified in the previous section, this method does not assume that hyper-parameters are independent *a posteriori*. We continue working on a logarithmic scale of r, i.e., $\alpha = \log(1/r)$. The joint posterior density function of α and $\boldsymbol{\beta}$ based on their joint hyper-prior density function in (12) is

$$f(\alpha, \boldsymbol{\beta}|\boldsymbol{y}) \propto f(\alpha, \boldsymbol{\beta}) L(\alpha, \boldsymbol{\beta}) \propto \exp(\alpha) L_b(\alpha, \boldsymbol{\beta}) d\alpha d\boldsymbol{\beta}. \tag{56}$$

The A-R method (Everson and Morris 2000; Tang 2002; Robert and Casella 2013) is useful when it is difficult to sample a parameter of interest θ directly from its target probability density $f(\theta)$, which is known up to a normalizing constant, but an easy-to-sample envelope function $g(\theta)$ is available. The A-R method samples θ from the envelope $g(\theta)$ and accepts it with a probability $\frac{f(\theta)}{Mg(\theta)}$, where M is a constant making $f(\theta)/g(\theta) \leq M$ for all θ. The distribution of the accepted θ exactly follows $f(\theta)$. The A-R method is stable as long as the tails of the envelop function are thicker than those of the target density function.

The goal of the A-R method for the Binomial model is to draw posterior samples of hyper-parameters from $f(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$, using an easy-to-sample envelop function $g(\alpha, \boldsymbol{\beta})$ that has thicker tails than the target density function.

We factor the envelope function into two parts, $g(\alpha, \boldsymbol{\beta}) = g_1(\alpha) g_2(\boldsymbol{\beta})$ to model the tails of each function separately. We consider the tail behavior of the conditional posterior density function $f(\alpha|\boldsymbol{\beta}, \boldsymbol{y})$ to come up with $g_1(\alpha)$; $f(\alpha|\boldsymbol{\beta}, \boldsymbol{y})$ behaves as $\exp(-\alpha(k-1))$ when α goes to ∞ and as $\exp(\alpha)$ when α goes to $-\infty$. It indicates that $f(\alpha|\boldsymbol{\beta}, \boldsymbol{y})$ is skewed to the left because the right tail touches the x-axis faster than the left tail does as long as $k > 1$. A skewed t-distribution is a good candidate for $g_1(\alpha)$ because it behaves as a power law on both tails, leading to thicker tails than those of $f(\alpha|\boldsymbol{\beta}, \boldsymbol{y})$.

It is too complicated to figure out the tail behaviors of $f(\boldsymbol{\beta}|\alpha, \boldsymbol{y})$. However, because $f(\boldsymbol{\beta}|\alpha, \boldsymbol{y})$ of the Gaussian hierarchical model has a multivariate Gaussian density function (Morris and Tang 2011; Kelly 2014), we consider a multivariate t-distribution with 4 degrees of freedom as a good candidate for $g_2(\boldsymbol{\beta})$.

Specifically, we assume

$$g_1(\alpha) = g_1(\alpha; \mu, \sigma, a, b) \equiv \text{Skewed-}t(\alpha|\mu, \sigma, a, b), \tag{57}$$

$$g_2(\boldsymbol{\beta}) = g_2(\boldsymbol{\beta}; \boldsymbol{\xi}, S_{(m\times m)}) \equiv t_4(\boldsymbol{\beta}|\boldsymbol{\xi}, S), \tag{58}$$

where the notation Skewed-$t(\alpha|\mu, \sigma, a, b)$ represents a density function of a skewed t-distribution at α with location μ, scale σ, degree of freedom $a + b$, and skewness $a - b$ for any positive constants a and b (Jones and Faddy 2003). The article of Jones and Faddy (2003) derives the mode of $g_1(\alpha)$ as

$$\mu + \frac{(a-b)\sqrt{a+b}}{\sqrt{(2a+1)(2b+1)}}. \tag{59}$$

and shows that the tails of the skewed-t density function follow a power law with $\alpha^{-(2a+1)}$ on the left and $\alpha^{-(2b+1)}$ on the right when $b > a$. It also provides a representation to generate the random variable following their skewed-t distribution as

$$\alpha \sim \mu + \sigma \frac{\sqrt{a+b}(2T-1)}{2\sqrt{T(1-T)}}, \text{ where } T \sim \text{Beta}(a, b). \tag{60}$$

The notation $t_4(\boldsymbol{\beta}|\boldsymbol{\xi}, S)$ in (58) indicates a density function of a multivariate t-distribution at $\boldsymbol{\beta}$ with 4 degrees of freedom, a location vector $\boldsymbol{\xi}$, and a $m \times m$ scale matrix S that leads to the variance-covariance matrix $2S$.

We determine the parameters of $g_1(\alpha)$ and $g_2(\boldsymbol{\beta})$, i.e., μ, σ, a, b, $\boldsymbol{\xi}$, and S, to make the product of $g_1(\alpha)$ and $g_2(\boldsymbol{\beta})$ similar to the target joint posterior density $f(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$. First, we obtain the mode of $f(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ and the inverse of the negative Hessian matrix at the modes, $-H^{-1}$. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ denote the modes of $f(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$, $-H^{-1}_{\hat{\alpha}}$ indicate (1, 1) element of $-H^{-1}$, and $-H^{-1}_{\hat{\boldsymbol{\beta}}}$ represent $-H^{-1}$ without the first row and column.

Next, we set $(a,\ b)$ to $(k,\ 2k)$ if $k < 10$ (or otherwise $(\log(k),\ 2\log(k))$) to maintain a left-skewness of $g_1(\alpha)$ and to keep a and b small enough for thick tails. We match the mode of $g_1(\alpha)$ specified in (59) to $\hat{\alpha}$ by setting the location parameter μ to $\hat{\alpha} - (a - b)\sqrt{a+b}/\sqrt{(2a+1)(2b+1)}$. We set the scale parameter σ to $(-H^{-1}_{\hat{\alpha}})^{0.5}\psi$, where ψ is a tuning parameter with 1.3 as a default value[1].

As for $g_2(\boldsymbol{\beta})$, we matches the location vector $\boldsymbol{\xi}$ to the mode $\hat{\boldsymbol{\beta}}$ and the scale matrix S to $-H^{-1}_{\hat{\boldsymbol{\beta}}}/2$ so that the variance-covariance matrix becomes $-H^{-1}_{\hat{\boldsymbol{\beta}}}$;

$$g_2(\boldsymbol{\beta}) \equiv t_4(\boldsymbol{\beta}|\boldsymbol{\xi} = \hat{\boldsymbol{\beta}}, S = -H^{-1}_{\hat{\boldsymbol{\beta}}}/2). \tag{61}$$

For the implementation of the acceptance-rejection method, we obtain four times more trial samples than the desired number of samples, denoted by N, independently from $g_1(\alpha)$ and $g_2(\boldsymbol{\beta})$. We calculate $4N$ weights, each of which is defined as

$$w_i \equiv w(\alpha^{(i)}, \boldsymbol{\beta}^{(i)}) = \frac{f(\alpha^{(i)}, \boldsymbol{\beta}^{(i)}|\boldsymbol{y})}{g_1(\alpha^{(i)})g_2(\boldsymbol{\beta}^{(i)})}, \text{ for } i = 1, 2, \ldots, 4N. \tag{62}$$

We accept each pair of $(\alpha^{(i)}, \boldsymbol{\beta}^{(i)})$ with a probability w_i/M where M is set to the maximum of all the $4N$ weights. In a case where we accept more than the desired number of samples N, we discard the redundant. If the number of accepted samples is smaller than N, then we additionally sample 6 times more pairs than the shortage and calculate a new maximum M' from all the previous and new weights, accepting or rejecting the entire pairs again with new probabilities w_i/M'.

Once we have posterior samples of hyper-parameters, it is easy to obtain posterior samples of random effects via a Monte Carlo integration below.

$$f(\boldsymbol{p}|\boldsymbol{y}) = \int f(\boldsymbol{p}|\alpha, \boldsymbol{\beta}, \boldsymbol{y}) \cdot f(\alpha, \boldsymbol{\beta}|\boldsymbol{y}) d\alpha d\boldsymbol{\beta}. \tag{63}$$

The integration can be done by sampling $(p_1, p_2, \ldots, p_k)$ from the independent Beta conditional posterior distributions, i.e., $f(p_j|\boldsymbol{\beta}, r, \boldsymbol{y})$ specified in (40), given r $(= \exp(-\alpha))$ and $\boldsymbol{\beta}$ that are already sampled from $f(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ via the A-R method.

[1]When the A-R method produces extreme weights defined in (62), we increase the value of ψ.

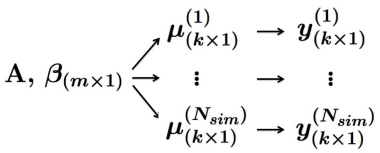

Figure 1: Pseudo-data generating process.

6. Frequency method checking

The question as to whether the interval estimates of random effects for given confidence level obtained by a specific model achieve the nominal coverage rate for any true parameter values is one of the key model evaluation criteria. Unlike standard model checking methods that test whether a two-level model is appropriate for data (Dean 1992; Christiansen and Morris 1996), *frequency method checking* is a procedure to evaluate the coverage properties of the model. Conditioning that the two-level model is appropriate, the frequency method checking generates pseudo-data sets given specific values of hyper-parameters and estimates unknown coverage probabilities based on these mock data sets (a parametric bootstrapping). We describe the frequency method checking based on the Gaussian model because the idea can be easily applied to the other two models.

6.1. Pseudo-data generation

Figure 1 displays the process of generating pseudo-data sets. It is noted that the conjugate prior distribution of each random effect in (2) is completely determined by two hyper-parameters, A and $\boldsymbol{\beta}$. Fixing these hyper-parameters at specific values, we generate N_{sim} sets of random effects from the conjugate prior distribution, i.e., $\{\boldsymbol{\mu}^{(i)},\ i = 1, \ldots, N_{\text{sim}}\}$, where the superscript (i) indicates the i-th simulation. Next, using the distribution of observed data in (1), we generate N_{sim} sets of observed data sets $\{\boldsymbol{y}^{(i)},\ i = 1, \ldots, N_{\text{sim}}\}$ given each $\boldsymbol{\mu}^{(i)}$. Note that we generate one observed data set per one set of random effects.

6.2. Coverage probability estimation

After fitting the Gaussian model on each simulated data set, we obtain interval estimates of random effects $\boldsymbol{\mu}$. Let $(\hat{\mu}^{(i)}_{j,\ \text{low}},\ \hat{\mu}^{(i)}_{j,\ \text{upp}})$ represent the lower and upper bounds of the interval estimate of random effect j based on the i-th simulation given a specific confidence level. Let's define a coverage indicator of random effect j on the i-th mock data set as

$$I_{A,\boldsymbol{\beta}}(\mu^{(i)}_j) = \begin{cases} 1, & \text{if } \mu^{(i)}_j \in (\hat{\mu}^{(i)}_{j,\ \text{low}},\ \hat{\mu}^{(i)}_{j,\ \text{upp}}) \\ 0, & \text{otherwise} \end{cases} \tag{64}$$

Note that an outcome of the coverage indicator depends on the simulated random effects and mock data generated by A and $\boldsymbol{\beta}$.

Simple unbiased coverage estimator.

When the confidence level is 95%, the proportion of 95% interval estimates that contain random effect j is an intuitive choice for the coverage rate estimator for random effect j. This

estimator implicitly assumes that there exist k unknown coverage probabilities of random effects, denoted by $C_{A,\boldsymbol{\beta}}(\mu_j)$ for $j = 1, 2, \ldots, k$, depending on the values of the hyper-parameters that generate random effects and mock data sets. The coverage indicators for random effect j in (64) is assumed to follow an independent and identically distributed Bernoulli distribution given the unknown coverage rate $C_{A,\boldsymbol{\beta}}(\mu_j)$. The sample mean of these coverage indicators is a simple unbiased coverage estimator for $C_{A,\boldsymbol{\beta}}(\mu_j)$;

$$\bar{I}_{A,\boldsymbol{\beta}}(\mu_j) = \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} I_{A,\boldsymbol{\beta}}(\mu_j^{(i)}), \ j = 1, 2, \ldots, k. \tag{65}$$

Note that $\bar{I}_{A,\boldsymbol{\beta}}(\mu_j)$ averages over possible values of μ_j and y_j generated by specific values of A and $\boldsymbol{\beta}$. For $j = 1, 2, \ldots, k$, the unbiased variance estimator of $\mathsf{VAR}(\bar{I}_{A,\boldsymbol{\beta}}(\mu_j))$ is

$$\widehat{\mathsf{VAR}}(\bar{I}_{A,\boldsymbol{\beta}}(\mu_j)) = \frac{1}{N_{\text{sim}}(N_{\text{sim}} - 1)} \sum_{i=1}^{N_{\text{sim}}} (I_{A,\boldsymbol{\beta}}(\mu_j^{(i)}) - \bar{I}_{A,\boldsymbol{\beta}}(\mu_j))^2. \tag{66}$$

Rao-Blackwellized unbiased coverage estimator.

The frequency method checking is computationally expensive in nature because it fits a model on every mock data set. The situation deteriorates if the number of simulations or the size of data is large, or the estimation method is computationally demanding. Christiansen and Morris (1997) and Tang (2002) used a Rao-Blackwellized (RB) unbiased coverage estimator for the unknown coverage rate of each random effect, which is more efficient than the simple unbiased coverage estimator. For $j = 1, 2, \ldots, k$,

$$C_{A,\boldsymbol{\beta}}(\mu_j) = \mathsf{E}(\bar{I}_{A,\boldsymbol{\beta}}(\mu_j)|A, \boldsymbol{\beta}) = E\left[\frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \mathsf{E}(I_{A,\boldsymbol{\beta}}(\mu_j^{(i)})|A, \boldsymbol{\beta}, \boldsymbol{y}^{(i)}) \middle| A, \boldsymbol{\beta}\right], \tag{67}$$

where the sample mean of the conditional expectations inside the outer expectation is the RB unbiased coverage estimator. To be specific,

$$\bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j) = \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \mathsf{E}(I_{A,\boldsymbol{\beta}}(\mu_j^{(i)})|A, \boldsymbol{\beta}, \boldsymbol{y}^{(i)}) \tag{68}$$

$$= \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \mathsf{P}(\mu_j^{(i)} \in (\hat{\mu}^{(i)}_{j,\,\text{low}},\ \hat{\mu}^{(i)}_{j,\,upp})|A, \boldsymbol{\beta}, \boldsymbol{y}^{(i)}). \tag{69}$$

We can easily compute the above conditional posterior probabilities using the cumulative density function of the Normal conditional posterior distribution of each random effect in (3). The variance of $\bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j)$ does not exceed the variance of a simple unbiased coverage estimator $\bar{I}_{A,\boldsymbol{\beta}}(\mu_j)$ (Rao 1945; Blackwell 1947).

If one dataset $\boldsymbol{y}^{(i)}$ is simulated per one set of random effects $\boldsymbol{\mu}^{(i)}$, the variance estimator below is an unbiased estimator of $\mathsf{VAR}(\bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j))$. For $j = 1, 2, \ldots, k$,

$$\widehat{\mathsf{VAR}}(\bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j)) \equiv \frac{1}{N_{\text{sim}}(N_{\text{sim}} - 1)} \sum_{i=1}^{N_{\text{sim}}} \left(\mathsf{E}(I_{A,\boldsymbol{\beta}}(\mu_j^{(i)})|A, \boldsymbol{\beta}, \boldsymbol{y}^{(i)}) - \bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j)\right)^2. \tag{70}$$

Rao-Blackwellized overall unbiased coverage estimator

Assuming that the unknown coverage probabilities are the same for all random effects, we use the Rao-Blackwellized overall unbiased coverage estimator and its variance estimator;

$$\bar{\bar{I}}^{RB}_{r,\boldsymbol{\beta}} = \frac{1}{k}\sum_{j=1}^{k} \bar{I}^{RB}_{r,\boldsymbol{\beta}}(p_j) \quad \text{and} \quad \widehat{\mathsf{VAR}}(\bar{\bar{I}}_{RB}) = \frac{1}{k^2}\sum_{j=1}^{k} \widehat{\mathsf{VAR}}(\bar{I}^{RB}_{r,\boldsymbol{\beta}}(p_j)). \tag{71}$$

7. Usage of functions in Rgbp

In this section, we describe the usage of the two main functions of **Rgbp**, i.e., `gbp` for model fitting and `coverage` for frequency method checking.

7.1. Model fitting

The function `gbp` creates an S3 object "gbp" on which three generic functions `plot`, `print`, and `summary` are defined.

There are two cases according to whether covariates are available or not. When no covariates are available, the function `gbp` requires fitting an intercept term or knowing the value of the the expected random effect, meaning that the intercept term must be either estimated or known. The default of `gbp` in this case is to fit an intercept term. We can assign the value of the known expected random effect through an optional argument `mean.PriorDist`. Note that `gbp` can fit the Poisson model only when the value of expected random effect, λ^E, is known. The usage of `gbp` to fit each model without any covariates is

```
R> g.output <- gbp(y, se.or.n, model = "gaussian")
R> b.output <- gbp(y, se.or.n, model = "binomial")
R> p.output <- gbp(y, se.or.n, mean.PriorDist, model = "poisson")
```

The argument `y` is a vector of k observed sample means for the Gaussian model, k observed numbers of successful outcomes for the Binomial model, and k observed numbers of events happening for the Poisson model. The argument `se.or.n` is a vector of k standard errors of each sample mean for the Gaussian model, k numbers of trials for the Binomial model, and k exposures for the Poisson model. If we want to designate a known value of the expected random effect $\mu^E = \beta_1$ for the Gaussian model or $p^E = \exp(\beta_1)/(1+\exp(\beta_1))$ for the Binomial model, then we put that value into `gbp` using an argument `mean.PriorDist`. For example, if μ^E is known as 10, then we use the following code.

```
R> g.output <- gbp(y, se.or.n, mean.PriorDist = 10, model = "gaussian")
```

If covariate information for each group is available, we can fit the Gaussian and Binomial models, using the following codes.

```
R> g.output <- gbp(y, se.or.n, X, model = "gaussian")
R> b.output <- gbp(y, se.or.n, X, model = "binomial")
```

The argument `X` is a matrix of covariate(s) each column of which corresponds to one covariate for k groups. For example, if we have two covariates for each group, the argument `X` must be $k \times 2$ matrix to estimate three regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ including an intercept term, β_1, as a default. If we do not want to include an intercept term ($\beta_1 = 0$), estimating two regression coefficients for the two covariates, we add an optional argument `intercept` as follows.

```
R> g.output <- gbp(y, se.or.n, X, model = "gaussian", intercept = FALSE)
```

The function `gbp` contains several optional arguments. The argument `Alpha`, whose default value is 0.95, sets the confidence level, producing $100\times$`Alpha`% interval estimates for the random effects. For the Gaussian model, setting the argument `normal.CI` to `TRUE` lets `gbp` use a Normal approximation to the unconditional posterior distribution of the random effect (Morris and Tang 2011). The default value of `normal.CI` is `FALSE` for the skewed-Normal approximation (Kelly 2014).

The function `gbp` uses the A-R method to fit the Binomial model if we assign the desired number of posterior samples (N in (62)) through the argument `n.AR`; its default value is 0. There are several arguments related to the A-R method. The argument `n.AR.factor` determines how many trial samples the method draws; its default value is 4, meaning that the method draws `n.AR` $\times 4$ trial samples and accepts or rejects them. The argument `trial.scale` is ψ determining the scale parameter of the skewed-t distribution (the envelope function); its default value is 1.3. The argument `save.result` indicates whether we save the whole posterior samples of the random effects and hyper-parameters; its default value is `TRUE`. The two arguments `t` and `u`, taking on positive values, allow users to choose the joint hyper-prior distribution $f(r, \boldsymbol{\beta}) \propto d\boldsymbol{\beta}dr/(t+r)^{u+1}$; the default values for `t` and `u` are 0 and 1, respectively, for the joint hyper-prior distribution specified in (12).

For example, when there are two covariates, the following code produces 2,000 posterior samples of random effects and hyper-parameters, r and $\boldsymbol{\beta}_{(3\times 1)}$ including an intercept term, via the A-R method with 8,000 trial samples.

```
R> b.output <- gbp(y, se.or.n, X, model = "binomial", n.AR = 2000)
```

The object `b.output` above contains 8,000 weights (`b.output$weight`), 2,000 posterior samples of α (`b.output$alpha`), a $2,000 \times 3$ matrix of β (`b.output$beta`) each column of which corresponds to 2,000 posterior samples of each regression coefficient, and a $k \times 2,000$ matrix of random effects (`b.output$p`) each row of which has posterior samples of each random effect.

The S3 object "gbp" benefits from three generic functions, `print`, `summary`, and `plot`. The estimation result for all the random effects appears if we type the "gbp" object at the R console, which plays the same role of the function `print` with its default argument "`sort = TRUE`". When the argument `sort` is set to `TRUE`, the function `print` prints out the estimation result for all the groups in the ascending order of n for the Binomial and Poisson model and the descending order of standard errors for the Gaussian model. When the argument `sort` is `FALSE`, the estimation result comes out in the order of data input.

```
R> b.output
R> print(b.output, sort = FALSE)
```

The function `summary` prints a detailed estimation result, including the estimation result for the hyper-parameters, A (or r) and $\boldsymbol{\beta}$.

```
R> summary(b.output)
```

The function `plot` draws a shrinkage plot and 100×`Alpha`% interval plot for random effects with its default argument "`sort = TRUE`" that draws the 100×`Alpha`% interval plot in the ascending order of n for the Binomial and Poisson model and the descending order of standard errors for the Gaussian model. When the argument `sort` is set to `FALSE` the 100×`Alpha`% interval plot is displayed in the order of data input.

```
R> plot(b.output)
R> plot(b.output, sort = FALSE)
```

7.2. Frequency method checking

The function `coverage` conducts a frequency method checking that estimates the coverage probabilities of random effects, conditioning on the values of the hyper-parameters that generate random effects and mock data sets. The basic usage of `coverage` needs a "gbp" object, such as `b.output` above, as the first argument;

```
R> cov <- coverage(b.output, nsim = 1000)
```

The argument `nsim` sets the number of simulations, N_{sim}, defined in Section 6.1. If we do not assign values of the hyper-parameters through the arguments `A.or.r` and `reg.coef`, then the function `coverage` automatically sets the estimated posterior modes of hyper-parameters saved in the "gbp" object (or their posterior medians if the acceptance-rejection method for the Binomial model is used) as the generative values of hyper-parameters. If we want to conduct the frequency method checking with different generated values of hyper-parameters, for example, $r = 100$ and $\boldsymbol{\beta} = (2, 5)^\top$ when one covariate was used with an intercept term, then we specify them via the arguments `A.or.r` and `reg.coef`;

```
R> cov <- coverage(b.output, A.or.r = 100, reg.coef = c(2, 5), nsim = 1000)
```

When we fit a model with a known value of the expected random effect, `coverage` conducts the frequency method checking based on the known value. However, we may sometimes want to conduct the frequency method checking with a different value for the expected random effect. For example, if we want to try a different value of the expected random effect, e.g., 30, we add the argument `mean.PriorDist` as follows.

```
R> cov <- coverage(p.output, mean.PriorDist = 30, nsim = 1000)
```

The resulting frequency method checking is based on the estimated posterior mode of r and the newly specified value of the expected random effect, 30.

Although the function `coverage` does not produce an S3 object, a detailed summary appears in a plot that automatically pops up. In addition, the variable `cov` where we save the result of `coverage` as above contains numerical details; k RB coverage estimates (`cov$coverageRB`)

and their standard errors (`cov$se.coverageRB`), k simple coverage estimates (`cov$coverageS`) and their standard errors (`cov$se.coverageS`), etc.

If we save the result into an object such as `cov` above, then we can always recall the coverage plot, using the function `coverage.plot`.

```
R> coverage.plot(cov)
```

8. Examples

8.1. Data of 31 hospitals with a known expected random effect

We analyze a data set of 31 hospitals in New York state comprising of the outcomes of the coronary artery bypass graft (CABG) surgery (Morris and Lysy 2012). The data set contains the number of deaths, $\boldsymbol{y}$, for a specified period after CABG surgeries out of the total number of patients, $\boldsymbol{n}$, receiving CABG surgeries in each hospital. Health care providers may use these data to improve their care qualities, and patients may refer to the data to select a better hospital. Our goal is to obtain the point and interval estimates for the unknown true mortality rates (random effects) of 31 hospitals to evaluate each hospital's reliability on the CABG surgery (Morris and Christiansen (1995) use a similar Poisson model to handle these hospital profile data). We interpret the caseloads, $\boldsymbol{n}$, as exposures and assume that the state-level death rate per exposure of this surgery (λ^E) is known as 0.03 to fit the Poisson model for an illustrative purpose. If covariate information is available or the information about the expected random effect is unknown, we recommend using the Binomial model.

These data can be loaded into R using the following code.

```
R> library("Rgbp")
R> data("hospital")
R> y <- hospital$d
R> n <- hospital$n
```

The function `gbp` fits the Poisson model with a Gamma conjugate prior distribution as the population distribution of the death rates in New York states whose mean (λ^E) is 0.030. The number of regression coefficients (m) is 0 because we do not estimate them via a log-linear regression.

```
R> p.output <- gbp(z, n, mean.PriorDist = 0.03, model = "poisson")
R> p.output

Summary for each unit (sorted by n):

      obs.mean    n prior.mean shrinkage low.intv post.mean upp.intv post.sd
1       0.0448   67       0.03     0.911   0.0199    0.0313   0.0454 0.00653
2       0.0294   68       0.03     0.910   0.0189    0.0299   0.0435 0.00631
3       0.0238  210       0.03     0.765   0.0185    0.0285   0.0407 0.00566
4       0.0430  256       0.03     0.728   0.0225    0.0335   0.0467 0.00619
```

```
5       0.0335  269       0.03     0.718   0.0208    0.0310   0.0432 0.00573
6       0.0438  274       0.03     0.714   0.0229    0.0339   0.0472 0.00621
7       0.0432  278       0.03     0.711   0.0228    0.0338   0.0469 0.00617
8       0.0136  295       0.03     0.699   0.0157    0.0250   0.0366 0.00534
9       0.0288  347       0.03     0.663   0.0200    0.0296   0.0410 0.00536
10      0.0372  349       0.03     0.662   0.0222    0.0325   0.0446 0.00571
11      0.0391  358       0.03     0.656   0.0228    0.0331   0.0454 0.00579
12      0.0177  396       0.03     0.633   0.0165    0.0255   0.0363 0.00506
13      0.0278  431       0.03     0.613   0.0200    0.0292   0.0400 0.00511
14      0.0249  441       0.03     0.608   0.0191    0.0280   0.0387 0.00502
15      0.0273  477       0.03     0.589   0.0199    0.0289   0.0394 0.00499
16      0.0455  484       0.03     0.585   0.0256    0.0364   0.0491 0.00601
17      0.0304  494       0.03     0.580   0.0211    0.0302   0.0409 0.00506
18      0.0220  501       0.03     0.577   0.0180    0.0266   0.0369 0.00483
19      0.0277  505       0.03     0.575   0.0202    0.0290   0.0395 0.00494
20      0.0204  540       0.03     0.559   0.0173    0.0258   0.0358 0.00474
21      0.0284  563       0.03     0.548   0.0206    0.0293   0.0395 0.00485
22      0.0236  593       0.03     0.535   0.0187    0.0270   0.0369 0.00466
23      0.0150  602       0.03     0.532   0.0147    0.0230   0.0329 0.00466
24      0.0238  629       0.03     0.521   0.0188    0.0271   0.0368 0.00460
25      0.0204  636       0.03     0.518   0.0173    0.0254   0.0351 0.00455
26      0.0480  729       0.03     0.484   0.0286    0.0393   0.0516 0.00587
27      0.0306  849       0.03     0.446   0.0223    0.0303   0.0397 0.00445
28      0.0274  914       0.03     0.428   0.0208    0.0285   0.0374 0.00423
29      0.0213  940       0.03     0.421   0.0176    0.0249   0.0335 0.00407
30      0.0293 1193       0.03     0.364   0.0223    0.0296   0.0379 0.00397
31      0.0201 1340       0.03     0.338   0.0170    0.0235   0.0310 0.00360
Mean             517      0.03     0.600   0.0201    0.0293   0.0403 0.00517
```

The output contains information about the observed death rates $\bar{y}_j$, caseloads n_j, known prior mean λ^E, shrinkage estimates $\hat{B}_j$, lower bounds of interval estimates $\hat{\lambda}_{j,\text{low}}$, posterior means $\hat{\lambda}_j$, upper bounds of interval estimates $\hat{\lambda}_{j,\text{upp}}$, and posterior standard deviations for each random effect based on the assumed unconditional Gamma posterior distributions.

A function `summary` shows selective information about hospitals with minimum, median, and maximum exposures and the estimation result of the hyper-parameter $\alpha = -\log(r)$.

```
R> summary(p.output)

Main summary:

                    obs.mean    n prior.mean shrinkage low.intv post.mean
Unit with min(n)      0.0448   67       0.03     0.911   0.0199    0.0313
Unit with median(n)   0.0455  484       0.03     0.585   0.0256    0.0364
Unit with max(n)      0.0201 1340       0.03     0.338   0.0170    0.0235
Overall Mean                  517       0.03     0.600   0.0201    0.0293

                    upp.intv  post.sd
```

```
0.0454  0.00653
0.0491  0.00601
0.0310  0.00360
0.0403  0.00517

Second-level Variance Component Estimation Summary:
alpha=log(A) for Gaussian or alpha=log(1/r) for Binomial and Poisson data:

post.mode.alpha post.sd.alpha post.mode.r
          -6.53         0.576         684
```

The output of `summary` shows that $\hat{r} = \exp(6.53) = 684$, which is an indicator of how valuable and informative the second-level hierarchy is. It means that observed sample means of hospitals whose caseloads are less than 684 shrink toward the prior mean (0.030) more than 50%. For example, the shrinkage estimate of the first hospital ($\hat{B}_1 = 0.911$) was calculated by 684 / (684 + 67), where 67 is its caseload (n_1), and its posterior mean is $(1 - 0.911) * 0.0448 + 0.911 * 0.030 = 0.0313$. As for this hospital, using more information from the conjugate prior distribution is an appropriate choice because the amount of observed information (67) is much less than the amount of state-level information (684).

To obtain a graphical summary, we use the function `plot`.

```
R> plot(p.output)
```

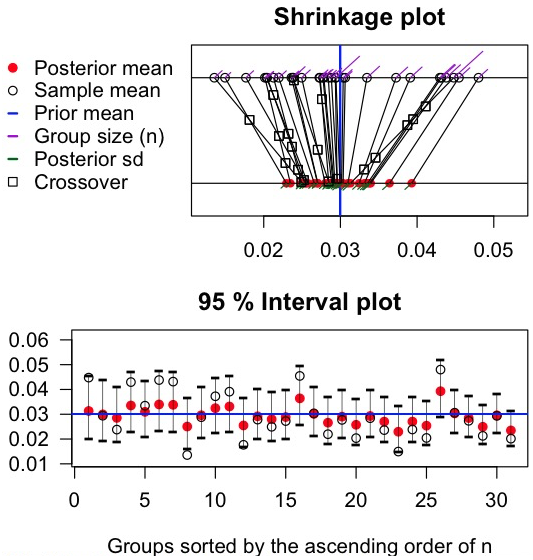

Figure 2: Shrinkage plot and 95% interval plot for 31 hospitals.

In Figure 2 the regression towards the mean (RTTM) is obvious in the first plot; the observed death rates, denoted by empty dots on the upper horizontal line, are shrinking towards the known expected random effect, denoted by a blue vertical line at 0.03, to the different extents. Note that some hospitals' ranks have changed by shrinking much harder towards 0.03 than

the others. For example, an empty square at the crossing point of the two left-most lines (8th and 23rd hospitals on the list above) indicates that the seemingly safest hospital in terms of the observed death rate is probably not the safest in terms of the estimated posterior mean accounting for the different caseloads of these two hospitals.

To be specific, their observed death rates (y_j, $j = 8, 23$) are 0.0136 and 0.0150 and caseloads (n_j, $j = 8, 23$) are 295 and 602, respectively. Considering solely the observed death rates may lead to an unfair comparison because the latter hospital handled twice the caseload. **Rgbp** accounts for this caseload difference, making the posterior mean for the random effect of the former hospital shrink toward the state-level mean (λ^E=0.03) much harder than that for the latter hospital.

Note that the point estimates are not enough to evaluate hospital reliability because one hospital may have a lower point estimate but larger uncertainty (variance) than the other. The second plot of Figure 2 displays the 95% interval estimates. Note that each posterior mean (red dot) is between the sample mean (empty dot) and the known expected random effect (a blue horizontal line).

This 95% interval plot reveals that the 31st hospital has the lowest upper bound even though its point estimate ($\hat{\lambda}_{31} = 0.0235$) is slightly larger than that of the 23rd hospital ($\hat{\lambda}_{23} = 0.0230$). The observed death rates for these two hospitals ($y_j, j = 23, 31$) are 0.0150 and 0.0201 and the caseloads ($n_j, j = 23, 31$) are 602 and 1340 each. The 31st hospital has twice the caseload, which leads to borrowing less information from the New York state-level hierarchy (or shrinking less toward the state-level mean, 0.03) with smaller variance. Based on the point and interval estimates, the 31st hospital seems the most reliable one.

We conduct the frequency method checking to question how reliable the estimation procedure is. The function `coverage` generates pseudo-datasets given the estimated value of r, 683.53, as a generative value. For reference, we can designate any generative values of r and λ^E by adding two arguments, `A.or.r` and `mean.PriorDist`, into the code below.

```
R> p.coverage <- coverage(p.output, nsim = 1000)
```

In Figure 3, the black horizontal line at 0.95 represents the nominal confidence level and the red circles indicate RB unbiased coverage estimates, $\bar{I}^{RB}_{r,\lambda^E}(\lambda_j)$ for $j = 1, 2, \ldots, 31$. The overall RB unbiased coverage estimate across all the hospitals ($\bar{\bar{I}}^{RB}_{r,\lambda^E}$) is 0.955. None of RB unbiased coverage estimates for 31 hospitals are less than 0.95 regardless of their caseloads. This result shows that the interval estimates for this particular dataset accurately achieves 95% confidence level.

The following code provides specific values of the 31 RB unbiased coverage estimates and their standard errors for each hospital (the output is omitted for a space concern).

```
R> p.coverage$coverageRB
R> p.coverage$se.coverageRB
```

And the code below shows 31 simple unbiased coverage estimates and their standard errors for each hospital.

```
R> p.coverage$coverageS
R> p.coverage$se.coverageS
```

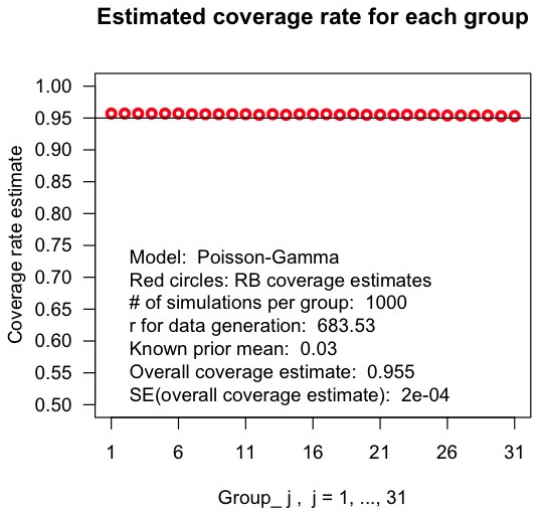

Figure 3: Coverage plot via frequency method checking for 31 hospitals.

It turns out that the variance estimate of the RB unbiased coverage estimate for the first hospital (0.0016^2) is about 19 times smaller than that of the simple one (0.0070^2). It means that the RB unbiased coverage estimates based on 1,000 simulations (N_{sim}) are as precise as the simple unbiased coverage estimates based on 19,000 simulations in terms of estimating the coverage probability for the first hospital, $C_{r,\lambda^{\mathsf{E}}}(\lambda_1)$.

8.2. Data of 8 schools with unknown expected random effect and no covariates

The Education Testing Service (ETS) conducted randomized experiments in eight separate schools (groups) to test whether students (units) SAT scores are effected by coaching. The dataset contains the estimated coaching effects on SAT scores ($y_j, j = 1, \ldots, 8$) and standard errors ($se_j, j = 1, \ldots, 8$) of the eight schools (Rubin 1981). We can load this data set into R by the following codes.

```
R> library("Rgbp")
R> data("schools")
R> y  <- schools$y
R> se <- schools$se
```

Due to the nature of the test each school's coaching effect has an approximately Normal sampling distribution with known sampling variance, i.e., standard error of each school is completely known. At the second hierarchy, the mean for each school is assumed to be drawn from a common Normal distribution and hence, we can use the Gaussian component of gbp to fit this Gaussian hierarchical model.

```
R> g.output <- gbp(y, se, model = "gaussian")
R> g.output
```

```
Summary for each group (sorted by the descending order of se):

     obs.mean   se prior.mean shrinkage low.intv post.mean upp.intv post.sd
8       12.00 18.0       8.17     0.734   -10.21      9.19     29.9   10.23
3       -3.00 16.0       8.17     0.685   -17.13      4.65     22.5   10.10
1       28.00 15.0       8.17     0.657    -2.32     14.98     38.8   10.56
4        7.00 11.0       8.17     0.507    -8.78      7.59     23.6    8.26
6        1.00 11.0       8.17     0.507   -13.03      4.63     20.1    8.44
2        8.00 10.0       8.17     0.459    -7.25      8.08     23.4    7.81
7       18.00 10.0       8.17     0.459    -1.29     13.48     30.8    8.18
5       -1.00  9.0       8.17     0.408   -13.30      2.74     16.7    7.63
Mean          12.5       8.17     0.552    -9.16      8.17     25.7    8.90
```

This output from `gbp` summarizes the results. In this Gaussian hierarchical model the amount of shrinkage for each unit is governed by the shrinkage factor, $B_j = V_j/(V_j + A)$. As such, schools whose variation within the school (V_j) is less than the between school variation (A) will shrink greater than 50%. The results provided by `gpb` suggests that there is little evidence that the training provided much added benefit due to the fact that every school's 95% posterior interval contains 0. In the case where the number of groups is large **Rgbp** provides a summary feature:

```
R> summary(g.output)

Main summary:

                     obs.mean   se prior.mean shrinkage low.intv post.mean
Unit with min(se)       -1.00  9.0       8.17     0.408   -13.30      2.74
Unit with median(se)1    1.00 11.0       8.17     0.507   -13.03      4.63
Unit with median(se)2    7.00 11.0       8.17     0.507    -8.78      7.59
Unit with max(se)       12.00 18.0       8.17     0.734   -10.21      9.19
Overall Mean                  12.5       8.17     0.552    -9.16      8.17

                     upp.intv post.sd
                         16.7    7.63
                         20.1    8.44
                         23.6    8.26
                         29.9   10.23
                         25.7    8.90

Second-level Variance Component Estimation Summary:
alpha=log(A) for Gaussian or alpha=log(1/r) for Binomial and Poisson data:

post.mode.alpha post.sd.alpha post.mode.A
           4.77          1.14         118


Regression Summary:
```

```
      estimate   se z.val p.val
beta1    8.168 5.73 1.425 0.154
```

The summary provides results regarding the second level hierarchy parameters. It can be seen that the estimate of the expected random effect, $\mu^E = \beta_1$ (`beta1`), is not significantly different from 0 suggesting that there was no effect of the coaching program on SAT math scores.

Rgbp also provides functionality to plot the results of the analysis as seen in Figure 4. Plotting the results provides a visual aid to understanding but is only largely beneficial when the number of groups (k) is small.

```
R> plot(g.output)
```

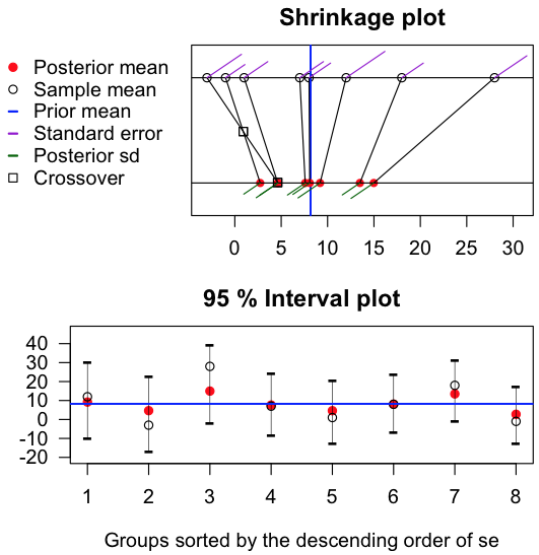

Figure 4: Shrinkage plot and 95% interval plot for 8 schools.

The frequency method checking generates new pseudo-data from our assumed model. Unless otherwise specified, the procedure fixes the hyper-parameter values at their estimates ($\hat{A}$ and $\hat{\boldsymbol{\beta}}_1$ in this example) and then simulates random effects θ_j for each group j. The model is then estimated and this is repeated an N_{sim} (`nsim`) number of times to estimate the coverage probabilities of the procedure.

```
R> g.coverage <- coverage(g.output, nsim = 1000)
```

As seen in Figure 5 the desired 95% confidence level, denoted by a black horizontal line at 0.95, is achieved for each school in this example. Note that all the coverage estimates depend on the chosen generative values of A and β_1, and the assumption that the model is valid.

In addition, Rao-Blackwellized (RB) unbiased coverage estimate and its standard error for each school can be gotten with the command below.

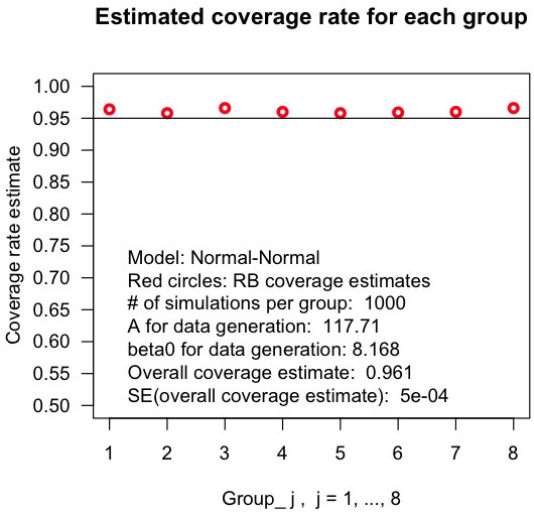

Figure 5: Coverage plot via frequency method checking for 8 schools.

```
R> g.coverage$coverageRB

 [1] 0.966 0.959 0.967 0.960 0.959 0.962 0.960 0.966

R> g.coverage$se.coverageRB

 [1] 0.0013 0.0012 0.0013 0.0013 0.0011 0.0011 0.0010 0.0017
```

All the individual RB coverage estimates are saved in the $8 \times 1,000$ matrix, each row of which is about each school. We can load this matrix by the following code.

```
R> g.coverage$raw.resultRB
```

8.3. Data of 18 baseball players with unknown expected random effect and one covariate

The data of 18 major league baseball players contain the batting averages through their first 45 official at-bats of the 1970 season (Efron and Morris 1975). We add a binary covariate that takes on one if a player is an outfielder and zero otherwise. We can load the data into R with the following codes.

```
R> library("Rgbp")
R> data("baseball")
R> y <- baseball$Hits
R> n <- baseball$At.Bats
R> x <- ifelse(baseball$Position == "fielder", 1, 0)
```

Conditioning on the unknown true batting average (random effect) of each player we assume that the at-bats are independent and therefore, $y_j|p_j \overset{\text{indep.}}{\sim} \text{Binomial}(45,\ p_j),\ j = 1, \ldots, 18$. Our goal is to obtain point and interval estimates of each random effect whilst considering the additional information on whether the player is an outfielder or not. The function `gbp` provides a way to incorporate such covariate information seamlessly into the model so that the regression towards the mean (RTTM) occurs within outfielders and non-outfielders separately.

```
R> b.output <- gbp(z, n, x, model = "binomial")
R> b.output
```

```
Summary for each unit (sorted by n):
```

	obs.mean	n	X1	prior.mean	shrinkage	low.intv	post.mean	upp.intv	post.sd
1	0.400	45	1.0	0.310	0.715	0.248	0.335	0.429	0.0462
2	0.378	45	1.0	0.310	0.715	0.244	0.329	0.420	0.0448
3	0.356	45	1.0	0.310	0.715	0.240	0.323	0.411	0.0437
4	0.333	45	1.0	0.310	0.715	0.236	0.316	0.403	0.0429
5	0.311	45	1.0	0.310	0.715	0.230	0.310	0.396	0.0424
6	0.311	45	0.0	0.233	0.715	0.179	0.256	0.341	0.0415
7	0.289	45	0.0	0.233	0.715	0.175	0.249	0.331	0.0400
8	0.267	45	0.0	0.233	0.715	0.171	0.243	0.323	0.0388
9	0.244	45	0.0	0.233	0.715	0.166	0.237	0.315	0.0380
10	0.244	45	1.0	0.310	0.715	0.210	0.291	0.379	0.0432
11	0.222	45	0.0	0.233	0.715	0.161	0.230	0.308	0.0377
12	0.222	45	0.0	0.233	0.715	0.161	0.230	0.308	0.0377
13	0.222	45	0.0	0.233	0.715	0.161	0.230	0.308	0.0377
14	0.222	45	1.0	0.310	0.715	0.202	0.285	0.375	0.0441
15	0.222	45	1.0	0.310	0.715	0.202	0.285	0.375	0.0441
16	0.200	45	0.0	0.233	0.715	0.155	0.224	0.302	0.0377
17	0.178	45	0.0	0.233	0.715	0.148	0.218	0.297	0.0381
18	0.156	45	0.0	0.233	0.715	0.140	0.211	0.292	0.0389
Mean		45	0.4	0.267	0.715	0.191	0.267	0.351	0.0410

Note that the shrinkage estimates are the same for all players because all players have the same 45 at-bats.

```
R> summary(b.output)
```

```
Main summary:
```

	obs.mean	n	X1	prior.mean	shrinkage	low.intv
Unit with min(obs.mean)	0.156	45	0.000	0.233	0.715	0.140
Unit with median(obs.mean)1	0.244	45	0.000	0.233	0.715	0.166
Unit with median(obs.mean)2	0.244	45	1.000	0.310	0.715	0.210
Unit with max(obs.mean)	0.400	45	1.000	0.310	0.715	0.248
Overall Mean		45	0.444	0.267	0.715	0.191

```
post.mean upp.intv post.sd
    0.211    0.292  0.0389
    0.237    0.315  0.0380
    0.291    0.379  0.0432
    0.335    0.429  0.0462
    0.267    0.351  0.0410

Second-level Variance Component Estimation Summary:
alpha=log(A) for Gaussian or alpha=log(1/r) for Binomial and Poisson data:

post.mode.alpha post.sd.alpha post.mode.r
          -4.73         0.957         113


Regression Summary:

      estimate    se  z.val p.val
beta1   -1.194 0.131 -9.129 0.000
beta2    0.389 0.187  2.074 0.038
```

The regression coefficient for the outfielder indicator is significant, considering that p value for $\hat{\beta}_2$ is 0.038. It means that the two estimates for the expected random effects for the outfielders and infielders are significantly different. Also, the positive sign of $\hat{\beta}_2$ indicates that the population batting average for all outfielders tends to be higher than that for infielders. The estimated odds ratio is $\exp(0.389) = 1.48$.

```
R> plot(b.output)
```

The shrinkage plot in Figure 6 shows that the observed batting averages (empty dots) on the upper horizontal line shrink towards the two expected random effects, 0.233 and 0.310. The short red line symbols near some empty dots are for when two or more points have the same mean and are plotted over each other. For example, five players (from the 11th player to the 15th) have the same batting average, 0.222, and at this point on the upper horizontal line, there are short red lines toward five directions.

The 95% interval plot in Figure 6 shows the range of true batting average for each player, which clarifies the regression toward the mean (RTTM) within two groups. The 10th, 14th, and 15th players, for example, are outfielders but their observed batting averages are far lower than the first five outfielders. This can be attributed to their bad luck because their observed batting averages are close to the lower bounds of their interval estimates. The RTTM indicates that their batting averages shrink towards the expected random effect of outfielders (0.310) in the long run.

To check the level of trust in these interval estimates, we proceed to frequency method checking by assuming the estimates, 112.95 for $\hat{r}$ and (-1.194, 0.389) for $\hat{\boldsymbol{\beta}}$, are the generative values.

```
R> b.coverage <- coverage(b.output, nsim = 1000)
```

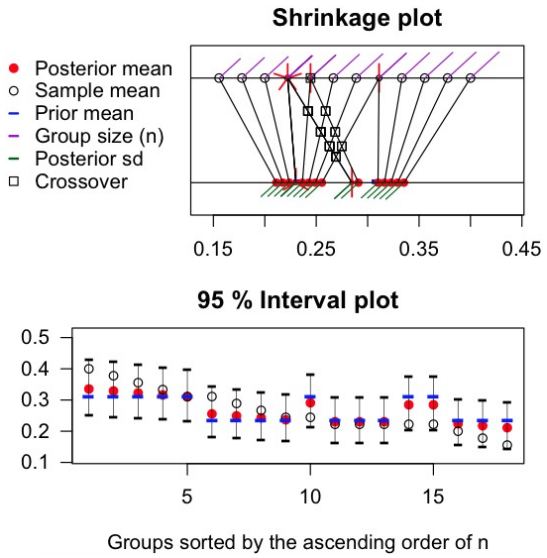

Figure 6: Shrinkage plot and 95% interval plot for 18 baseball players.

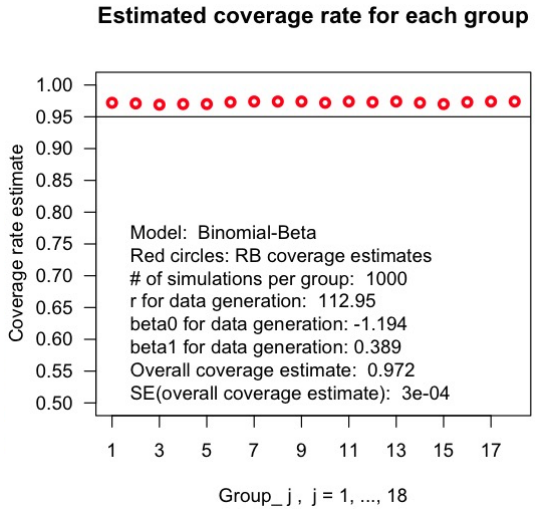

Figure 7: Coverage plot via frequency method checking for 18 players.

In Figure 7, the estimated coverage probabilities for random effects are beyond 0.95, conservatively satisfying the 95% confidence level. The RB overall unbiased coverage estimate across all the players is 0.972. Note that each coverage estimate depends on given true values of r and $\boldsymbol{\beta}_{(2\times1)}$, and the assumption that the model is valid (overdispersion exists).

We can check the RB unbiased coverage estimates and their standard errors for each player.

```
R> bcv$coverageRB
```

```
 [1] 0.971 0.973 0.972 0.972 0.970 0.973 0.973 0.974 0.973 0.973 0.971 0.973
[13] 0.973 0.972 0.972 0.971 0.973 0.971
```

```
R> bcv$se.coverageRB

 [1] 0.0015 0.0012 0.0013 0.0014 0.0016 0.0010 0.0012 0.0010 0.0010 0.0013
[11] 0.0015 0.0013 0.0019 0.0013 0.0014 0.0015 0.0011 0.0014
```

If we want to draw 2,000 posterior samples of random effects and hyper-parameters from their full posterior distribution via the A-R method, we use the following R code.

```
R> b.output <- gbp(y, n, x, model = "binomial", n.AR = 2000)
```

The "gbp" object `b.output` contains 8,000 weights (`b.output$weight`), 2,000 posterior samples of α (`b.output$alpha`), a $2,000 \times 2$ matrix of β (`b.output$beta`) each column of which corresponds to 2,000 posterior samples of each regression coefficient, and a $k \times 2,000$ matrix of random effects (`b.output$p`) each row of which has posterior samples of each random effect.

If we run the frequency method checking using this "gbp" object obtained via the A-R method, the N_{sim} simulations also run the A-R method each time.

9. Discussion

Rgbp is an R package for estimating and validating two-level Gaussian, Poisson, and Binomial hierarchical models. The package aims to provide a procedure that is computationally efficient with good frequency properties and includes "frequency method checking" functionality to examine repeated sampling properties and to test that the method is valid at specified hyper-parameter values.

As an alternative to other maximization based estimation methods such as MLE and REML, **Rgbp** provides approximate point and interval estimates of parameters via ADM. Using the ADM approach, with our specified choice of priors, protects from cases of overshrinkage and undercoverage from which the aforementioned methods suffer (Morris 1988b).

A benefit of **Rgbp** is that it produces non-random output (except the A-R method for the Binomial model) and so results are easily reproduced and compared across studies. In addition to being a standalone analysis tool the package can be used as an aid in a broader estimation procedure. For example, by checking the similarity of output of **Rgbp** and that of another estimation procedure such as MCMC (Markov Chain Monte Carlor), the package can be used as a confirmatory tool to check whether the alternative procedure has been programmed correctly. In addition, the parameter estimates obtained via **Rgbp** can be used to initialize a MCMC thus decreasing time to convergence. Lastly, due to its speed and ease of use, **Rgbp** can be used as a method of preliminary data analysis. Such results may tell statisticians and practitioners alike whether a more intensive method in terms of implementation and computational time, such as MCMC, is needed.

A. Posterior propriety of the Poisson model

If the posterior distribution of r is proper, then the full posterior distribution of random effects and r is also proper because

$$f(\boldsymbol{\lambda}, r|\boldsymbol{y}) = f(\boldsymbol{\lambda}|\boldsymbol{y}) \cdot f(r|\boldsymbol{y}), \tag{72}$$

where $f(\boldsymbol{\lambda}|\boldsymbol{y})$ is a product of k proper conditional posterior density function in (41). Thus, our goal is to show that $\int_0^\infty f(r|\boldsymbol{y})dr < \infty$;

$$f(r|\boldsymbol{y}) \propto \frac{1}{r^2}L(r) \propto \frac{1}{r^2}\prod_{j=1}^{k}\frac{\Gamma(r\lambda^E + y_j)}{\Gamma(r\lambda^E)}(1-B_j)^{y_i}B_j^{r\lambda^E} \tag{73}$$

$$= \frac{1}{r^2}\left[r^{\sum_{j=1}^k y_j} + \cdots + a_k r^k\right]\exp\left(-r\lambda^E\sum_{j=1}^{k}\log(1+n_j/r)\right)\prod_{j=1}^{k}\left(\frac{n_j}{n_j+r}\right)^{y_j}, \tag{74}$$

where the polynomial function of r in the bracket has constant coefficients.

If there are at least two groups whose observed values y_j are non-zero, then $f(r|\boldsymbol{y})$ goes to zero as r goes to zero due to the polynomial function of r in (74); the following two factors in (74) approach one. As r becomes infinite, $f(r|\boldsymbol{y})$ touches zero exponentially fast due to the exponential term in the middle of (74). Thus, the integration of $f(r|\boldsymbol{y})$ must be finite.

Acknowledgments

The authors thank Professor Cindy Christiansen, Professor Phil Everson and the 2012 class of Harvard's Stat 324r: Parametric Statistical Inference and Modeling for their valuable inputs.

References

Albert JH (1988). "Computational methods using a Bayesian hierarchical generalized linear model." *Journal of the American Statistical Association*, **83**(404), 1037–1044.

Azzalini A (1985). "A Class of Distributions which Includes the Normal Ones." *Scandinavian journal of statistics*, pp. 171–178.

Blackwell D (1947). "Conditional Expectation and Unbiased Sequential Estimation." *The Annals of Mathematical Statistics*, pp. 105–110.

Christiansen C, Morris C (1996). "Fitting and Checking a Two-Level Poisson Model: Modeling Patient Mortality Rates in Heart Transplant Patients." In D Berry, D Stangl (eds.), *Bayesian Biostatistics*, pp. 467–501. CRC press.

Christiansen C, Morris C (1997). "Hierarchical Poisson Regression Modeling." *Journal of the American Statistical Association*, **92**(438), pp. 618–632. ISSN 01621459. URL http://www.jstor.org/stable/2965709.

Daniels MJ (1999). "A Prior for the Variance in Hierarchical Models." *Canadian Journal of Statistics*, **27**(3), 567–578.

Dean CB (1992). "Testing for Overdispersion in Poisson and Binomial Regression Models." *Journal of the American Statistical Association*, **87**(418), 451–457.

Efron B, Morris C (1975). "Data Analysis Using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association*, **70**(350), pp. 311–319. ISSN 01621459. URL http://www.jstor.org/stable/2285814.

Everson PJ, Morris CN (2000). "Inference for Multivariate Normal Hierarchical Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(2), 399–412.

Gelman A, Carlin JB, Stern HS, Rubin DB (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Ghosh M, Rao J (1994). "Small area estimation: an appraisal." *Statistical science*, pp. 55–76.

Jones M, Faddy M (2003). "A Skew Extension of the t-Distribution, with Applications." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(1), 159–174.

Kass RE, Steffey D (1989). "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)." *Journal of the American Statistical Association*, **84**(407), pp. 717–726. ISSN 01621459. URL http://www.jstor.org/stable/2289653.

Kelly J (2014). *Advances in the Normal-Normal Hierarchical Model*. Ph.D. thesis, Harvard University.

Lee Y, Nelder JA (1996). "Hierarchical Generalized Linear Models." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 619–678.

Lee Y, Nelder JA, Pawitan Y (2006). *Generalized Linear Models with Random Effects: A Unified Analysis via h-Likelihood*. Chapman & Hall/ CRC, New York.

Morris C (1988a). "Approximating Posterior Distributions and Posterior Moments." In J Bernardo, MH DeGroot, DV Lindley, AFM Smith (eds.), *Bayesian Statistics 3*, pp. 327–344. Oxford University Press.

Morris C (1988b). "Determining the Accuracy of Bayesian Empirical Bayes Estimates in the Familiar Exponential Families." In S Gupta, J Berger (eds.), *Statistical Decision Theory and Related Topics IV*, pp. 251–263. Springer-Verlag.

Morris C, Christiansen C (1995). "Hierarchical Models for Ranking and for Identifying Extremes, with Application." In J Bernardo, J Berger, A Dawid, A Smith (eds.), *Bayesian Statistics 5*, pp. 227–296. New York: Oxford University Press.

Morris C, Lysy M (2012). "Shrinkage Estimation in Multilevel Normal Models." *Statistical Science*, **27**(1), 115–134.

Morris C, Tang R (2011). "Estimating Random Effects via Adjustment for Density Maximization." *Statistical Science*, **26**(2), pp. 271–287. ISSN 08834237. URL http://www.jstor.org/stable/23059992.

Rao CR (1945). "Information and Accuracy Attainable in the Estimation of Statistical Parameters." *Bulletin of the Calcutta Mathematical Society*, **37**(3), 81–91.

Rao JN (2003). *Small area estimation*. Wiley Online Library.

Robert C, Casella G (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Rönnegård L, Shen X, Alam M (2010). "**hglm**: A Package for Fitting Hierarchical Generalized Linear Models." *The R Journal*, **2**(2), 20–28. ISSN 20734859.

Rönnegård L, Shen X, Alam M (2011). *The **hglm** Package*. URL http://CRAN.R-project.org/package=hglm.

Rubin DB (1981). "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics*, **6**(4), pp. 377–401. ISSN 03629791. URL http://www.jstor.org/stable/1164617.

Skellam J (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 (Methodological)*, **10**(2), 257–261.

Tak H, Morris C (in preparation). "Posterior Propriety and Frequency Coverage Evaluation of Bayesian Beta-Binomial Logistic Regression Model."

Tamura RN, Young SS (1987). "A stabilized moment estimator for the beta-binomial distribution." *Biometrics*, pp. 813–824.

Tang R (2002). *Fitting and Evaluating Certain Two-Level Hierarchical Models*. Ph.D. thesis, Harvard University.

Affiliation:

Hyungsuk Tak
Department of Statistics
Harvard University
1 Oxford Street, Cambridge, MA
E-mail: hyungsuk.tak@gmail.com

Joseph Kelly
Google
76 Ninth Avenue, New York, NY
E-mail: josephkelly@google.com

Carl Morris
Department of Statistics
Harvard University
1 Oxford Street, Cambridge, MA
E-mail: morris@fas.harvard.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd