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 averages or proportions, possibly together with covariates of each group but without population-level data. The approximate Bayesian tool equipped with the adjustment for density maximization produces point and interval estimates for each random effect, point estimates and standard errors for regression coefficients, and a point estimate and its standard error for a second-level variance component. For the Binomial data, the package provides an option to produce posterior samples of all the model parameters via the acceptance-rejection method. The main goal of **Rgb** is to produce approximate Bayesian interval estimates for the random effects that meet their nominal confidence levels, and the package uses unique improper hyper-prior distributions for that purpose. **Rgbp** provides a quick way to check whether the resultant 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, repeated sampling coverage evaluation, R.

1. Introduction

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 the 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 conjugate prior distribution is non-exchangeable if the model incorporates covariate information of each group via a linear, log-linear, or logistic regression according to the data type, and exchangeable if no covariates are available with only an intercept term for the regression.

With an assumption of homogeneity within each group, the observed data are sufficient statistics for the random effects, such as averages or proportions, possibly together with each group's covariate information. This type of data is common for a biological analysis on litter data, a meta analysis on independent studies, or small area estimation problems. For these data, the two-level model that **Rgbp** assumes can be considered as a conjugate hierarchical generalized linear model (Lee and Nelder 1996; Lee, Nelder, and Pawitan 2006) where each random effect has a conjugate prior distribution.

Rgbp takes a Bayesian approach with our special improper hyper-prior distributions on hyper-parameters, the parameters of the conjugate prior distribution, to produce Bayesian interval estimates for the random effects that achieve nominal confidence levels. The hyper-prior distributions lead to Stein's harmonic prior known to produce good repeated sampling coverage rates of the Bayesian interval estimates for random effects in a two-level Gaussian hierarchical model (Morris and Tang 2011; Morris and Lysy 2012; Kelly 2014). We apply an analog to Stein's harmonic prior to Poisson and Binomial hierarchical models.

When it comes to fitting the model, **Rgbp** adopts the adjustment for density maximization (Morris 1988a; Christiansen and Morris 1997; Morris and Tang 2011) (ADM), a Pearson family approximation via maximization. For example, a Delta method is a special case of ADM that uses a Normal distribution to obtain an approximate distribution of a function of a parameter with its MLE and observed information plugged-in. In this article, the ADM uses Beta distributions to approximate the posterior distributions of shrinkage factors, functions of the second-level variance component, via maximization. For the approximation to the distributions of the shrinkage factors, the Beta distribution is more appropriate than the Normal distribution (Delta method) because the support of each shrinkage factor is between 0 and 1, not $(-\infty, \infty)$. Also, the ADM is free of a boundary effect that a restricted MLE (Patterson and Thompson 1971) (REML) sometimes suffers from, preventing a zero estimate for the second-level variance component. Using the approximate Beta posterior distributions of shrinkage factors, **Rgbp** estimates the first two (three for the Gaussian case) posterior moments of the random effects. Finally, **Rgbp** approximates the posterior distribution of each random effect by a skewed Normal distribution for a Gaussian case, a Gamma distribution for a Poisson case, and a Beta distribution for a Binomial case whose two parameters (three for the skewed Normal distribution) are matched to the previously estimated posterior moments of the random effects.

For the Binomial multilevel model, **Rgbp** provides an option to draw independent posterior samples of all the model parameters via an acceptance-rejection method instead of the approximate tool.

In addition to fitting hierarchical models, **Rgbp** provides a quick way to evaluate 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). It is a unique procedure that distinguishes **Rgbp** from any other R packages for hierarchical modeling such as **hglm** (Rönnegård, Shen, and Alam 2010, 2011) for conjugate hierarchical generalized models and **arm** (Gelman, Su, Yajima, Hill, Pittau, Kerman, and

Zheng 2014) for Bayesian hierarchical regression models. The evaluation procedure which we call "frequency method checking" adopts a parametric bootstrapping that generates mock data sets given the values of the hyper-parameters and estimates the coverage rates based on the generated mock data sets.

The rest of this paper is organized as follows. We specify the Bayesian hierarchical models and discuss their posterior propriety in Section 2. We describe the estimation procedures including the ADM and the acceptance-rejection method in Section 3, and the frequency method checking in Section 5. We explain the usages of main functions in **Rgbp** in Section 6, and apply them to three examples in Section 7.

2. Conjugate hierarchical modeling structure

One of the functions in **Rgbp**, `gbp`, fits a conjugate hierarchical model whose first-level has distributions of observed data and whose second-level has conjugate distributions on the first-level mean parameters (random effects). The `gbp` function allows users to choose one of three types of hierarchical models according to the type of data, namely Normal-Normal, Poisson-Gamma, and Binomial-Beta models.

2.1. Normal-Normal ("g"=Gaussian data)

The following is the general Normal-Normal hierarchical model assumed by `gbp`. For reference, V_j ($\equiv \sigma^2/n_j$) below is assumed to be known, and subscript j indicates the j-th group among k groups in the dataset. For $j = 1, 2, \ldots, k$,

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

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

where $\mu_j^E = E(\mu_j|\boldsymbol{\beta}, A) = \boldsymbol{x}_j^\top \boldsymbol{\beta} = \beta_1 x_{j,1} + \beta_2 x_{j,2} + \cdots + \beta_m x_{j,m}$ is the expected random effect and m is the number of regression coefficients to be estimated. The default of the function `gbp` is to set $x_{j,1}$ to 1 for an intercept term, though `gbp` 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 with an 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 distributions of the random effects. For $j = 1, 2, \ldots, k$,

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

where $B_j \equiv V_j/(V_j + A)$, $j = 1, \ldots, k$, are called shrinkage factors, and $\boldsymbol{y}^T = (y_1, y_2, \ldots, y_k)$. Note that the conditional posterior mean of the random effect, $\mu_j^* \equiv E(\mu_j|\boldsymbol{\beta}, A, \boldsymbol{y}) = (1 - B_j)y_j + B_j\mu_j^E$, is a convex combination of the observed sufficient statistic 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 more accurate second-level conjugate prior distribution.

2.2. Poisson-Gamma ("p"=Poisson data)

The function `gbp` is also capable of estimating a conjugate Poisson-Gamma hierarchical model, though its usage is limited to the case where the expected random effects are known ($m = 0$). For $j = 1, 2, \ldots, k$,

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

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

where n_j is the exposure of group j, which is not necessarily an integer, $\lambda^E = E(\lambda_j|r)$ is the known expected random effect, and r is the unknown second-level variance component. The mean and variance of the conjugate Gamma prior distribution are λ^E and λ^E/r, respectively. We interpret r as the amount of prior information as n_j represents the amount of observed information, which makes intuitive sense because the uncertainty of the conjugate prior distribution increases as r decreases; in the limit of r going to 0, the conjugate prior distribution gets flatter. The conditional posterior distribution of the random effect λ_j for this Poisson-Gamma model is

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

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

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

where $B_j \equiv r/(r + n_j)$ for $j = 1, 2, \ldots, k$. The conditional posterior mean is a convex combination of the observed sufficient statistic $\bar{y}_j = y_j/n_j$ and the expected random effect λ^E weighted by the relative amount of information in the prior compared to the data, called a shrinkage factor $B_j = r/(r+n_j)$. 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%, borrowing more information from the prior distribution.

Note that the conditional posterior variance in Equation 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 in the prior and conditional posterior distributions quadratic functions of their means respectively.

2.3. Binomial-Beta ("b"=Binomial data)

The Binomial-Beta hierarchical model is the last model that `gbp` can fit. The notation y_j is the number of successes (or failures) out of n_j trials. Unlike the Poisson-Gamma model, the expected random effect is either known ($m = 0$) or unknown ($m \geq 1$) a priori.

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

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

where $p_j^E \equiv E(p_j|\boldsymbol{\beta}, r) = \exp(\boldsymbol{x}_j^\top\boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_j^\top\boldsymbol{\beta}))$ is the expected random effect of group j ($j = 1, 2, \ldots, k$). 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 j is

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

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

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

The conditional posterior mean is a convex combination of the observed sufficient statistic $\bar{y}_j = y_j/n_j$ and the expected random effect p_j^E weighted by the relative amount of information in the prior compared to the data (a shrinkage factor) $B_j \equiv r/(r+n_j)$ like the Poisson-Gamma model. If the conjugate prior distribution contains more information than the observed distribution does ($r > n_j$), then the resulting conditional posterior mean shrinks towards the expected random effect by more than 50%.

2.4. Hyper-prior Distribution

Hyper-prior distributions are the distributions assigned to the second-level parameters called hyper-parameters. With the goal of objectivity in mind, our choice for the hyper-prior distributions is

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

where m is the number of the regression coefficients to be estimated. 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 produces good repeated sampling coverage properties of the Bayesian interval estimates for the random effects. 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-Gamma and Binomial-Beta, 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 Uniform$(0,\infty)$ for the Gaussian case. The resultant full posterior distribution of random effects and hyper-parameters for the Binomial case is data-dependent. 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 the entire k groups. Then, 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 (Tak and Morris in preparation).

The Poisson-Gamma model with the hyper-prior distributions in Equation 12 provides posterior propriety if and only 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 ($m = 0$). If the expected random effect is unknown a priori, then we recommend staying with the Binomial-Beta model with the same hyper-prior distributions because the Poisson-Gamma model is actually an approximation to the Binomial-Beta model.

3. Inference via the adjustment for density maximization

Obtaining point and interval estimates of each random effect is our primary inferential interest. The approximate Bayesian tool assumes that the unconditional posterior distribution of each random effect follows a skewed-Normal distribution for the Gaussian case, a Gamma distribution for the Poisson case, or a Beta distribution for the Binomial case whose parameters are matched to the estimated unconditional posterior moments of each random effect. We use these assumed unconditional posterior distributions to make point and interval estimates of each random effect.

We illustrate our estimation procedure equipped with the adjustment for density maximization (hereafter ADM) (Morris 1988a; Christiansen and Morris 1997; Morris and Tang 2011), a way to approximate a distribution of the parameter of interest by one of Pearson family distributions based on derivatives like the Delta method, using the Binomial and Poisson hierarchical models. The ADM procedure for the Gaussian hierarchical model adopted in **Rgbp** has been well-documented in Kelly (2014).

3.1. The inferential model

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

$$L(r, \boldsymbol{\beta}) = \prod_{j=1}^{k} f(y_j | r, \boldsymbol{\beta}) = \prod_{j=1}^{k} \binom{n_j}{y_j} \frac{B(y_j + r p_j^E,\ n_j - y_j + r(1 - p_j^E))}{B(r p_j^E,\ r(1 - p_j^E))}, \tag{13}$$

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

$$L(r) = \prod_{j=1}^{k} f(y_j | r) = \prod_{j=1}^{k} \frac{\Gamma(r\lambda^E + y_j)}{\Gamma(r\lambda^E)(y_j!)} (1 - B_j)^{y_i} B_j^{r\lambda^E}. \tag{14}$$

The joint posterior density function of hyper-parameters $f(r, \boldsymbol{\beta} | \boldsymbol{y})$ for the Binomial hierarchical model is proportional to their likelihood functions in Equation 13 multiplied by the hyper-prior density functions of r and $\boldsymbol{\beta}$ in Equation 12 as follows;

$$f(r, \boldsymbol{\beta} | \boldsymbol{y}) \propto L(r, \boldsymbol{\beta}) d\boldsymbol{\beta} dr / r^2. \tag{15}$$

The posterior density function of r, $f(r|\boldsymbol{y})$, for the Poisson hierarchical model is the likelihood function in Equation 14 times the hyper-prior density function of r, dr/r^2;

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

Our goal is to obtain the point and interval estimates of the random effects from their unconditional posterior distributions; for the Binomial case,

$$f(\boldsymbol{p}|\boldsymbol{y}) = \int f(\boldsymbol{p}|r, \boldsymbol{\beta}, \boldsymbol{y}) \cdot f(r, \boldsymbol{\beta}|\boldsymbol{y}) dr d\boldsymbol{\beta}, \tag{17}$$

and for the Poisson case,

$$f(\boldsymbol{\lambda}|\boldsymbol{y}) = \int f(\boldsymbol{\lambda}|r,\boldsymbol{y}) \cdot f(r|\boldsymbol{y})dr. \tag{18}$$

3.2. Estimation for shrinkage factors and expected random effects

Estimating the unconditional posterior moments of the shrinkage factors, $B_1, B_2, \ldots, B_k \equiv r/(r+n_k)$, and the conditional posterior moments of the expected random effects, $p_1^E, p_2^E, \ldots, p_k^E$ (or λ^E for the Poisson model), is the most important estimation problem for the hierarchical models that `gbp` assumes. This is because these moment estimates are used to approximate the unconditional posterior moments of the random effects, $p_1, p_2, \ldots, p_k$ (or $\lambda_1, \lambda_2, \ldots, \lambda_k$ for the Poisson model). Taking the Binomial-Beta model as an example, with the assumption that hyper-parameters r (or A for the Gaussian model) and $\boldsymbol{\beta}$ are independent a posteriori, the unconditional posterior mean and variance of random effect j are

$$\begin{aligned}
E(p_j|\boldsymbol{y}) &= E(E(p_j|r,\boldsymbol{\beta},\boldsymbol{y})|\boldsymbol{y}) = (1 - E(B_j|\boldsymbol{y}))\bar{y}_j + E(B_j|\boldsymbol{y})E(p_j^E|\boldsymbol{y}) && (19)\\
Var(p_j|\boldsymbol{y}) &= E(Var(p_j|r,\boldsymbol{\beta},\boldsymbol{y})|\boldsymbol{y}) + Var(E(p_j|r,\boldsymbol{\beta},\boldsymbol{y})|\boldsymbol{y}) && (20)\\
&= E\big(p_j^*(1-p_j^*)/(r+n_j+1)|\boldsymbol{y}\big) + Var\big(\mathrm{B}_j(\bar{y}_j - p_j^E)|\boldsymbol{y}\big) && (21)\\
&\approx E\big(p_j^*(1-p_j^*)(1-\mathrm{B}_j)/n_i|\boldsymbol{y}\big) + Var\big(\mathrm{B}_j(\bar{y}_j - p_j^E)|\boldsymbol{y}\big) && (22)\\
&= g(E(B_j|\boldsymbol{y}), E(B_j^2|\boldsymbol{y}), E(B_j^3|\boldsymbol{y}), E(p_j^E|\boldsymbol{y}), E((p_j^E)^2|\boldsymbol{y})). && (23)
\end{aligned}$$

Note that the unconditional posterior mean and approximate variance of random effect j in Equation 19 and 23 are functions of the unconditional posterior moments of shrinkage factors and expected random effects. We specify the function g in Equation 23 in Appendix A.

We assumed that hyper-parameters r and $\boldsymbol{\beta}$ were independent a posteriori, considering that they are independent a posteriori in the limit of k going to infinity because they are asymptotically Normally distributed. Also, Christiansen and Morris (1997) empirically showed that their covariance from the observed information matrix of the Poisson-Gamma model, though with a different parametrization, was close to 0 in a small sample setting.

For the Poisson model, the unconditional posterior mean and variance of random effect j are

$$\begin{aligned}
E(\lambda_j|\boldsymbol{y}) &= E(E(p_j|r,\boldsymbol{y})|\boldsymbol{y}) = (1 - E(B_j|\boldsymbol{y}))\bar{y}_j + E(B_j|\boldsymbol{y})\lambda^E && (24)\\
Var(\lambda_j|\boldsymbol{y}) &= E(Var(p_j|r,\boldsymbol{y})|\boldsymbol{y}) + Var(E(\lambda_j|r,\boldsymbol{y})|\boldsymbol{y}) && (25)\\
&= E\big(\lambda_j^*/(r+n_j)|\boldsymbol{y}\big) + Var\big(\mathrm{B}_j(\bar{y}_j - \lambda_j^E)|\boldsymbol{y}\big) && (26)\\
&= h(E(B_j|\boldsymbol{y}), E(B_j^2|\boldsymbol{y})). && (27)
\end{aligned}$$

The unconditional posterior mean and variance of random effect j under the Poisson model are also functions of the unconditional posterior moments of the shrinkage factors. We specify the function h in Equation 27 in Appendix B.

Next, we estimate the unconditional posterior moments of the shrinkage factors and expected random effects after approximating their unconditional posterior distributions by Beta distributions via the ADM.

Unconditional posterior moments of shrinkage factors.

It is noted that the shrinkage factors $(B_1, \ldots B_k)$ are a function of r, i.e., $B_j = r/(r+n_j) = B_j(r)$. One way to approximate the distribution of B_j is to find the maximum likelihood

estimate of r, $\hat{r}_{MLE}$, with its Hessian value and to use a Delta method for an asymptotic Normal distribution of $B_j(\hat{r}_{MLE})$. 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 this approximation can be quite flawed and can even 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 the ADM (Morris 1988a; Christiansen and Morris 1997; Morris and Tang 2011). 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. For our purposes we approximate the posterior distribution of a shrinkage factor with a Beta distribution, which allows us to finally obtain estimates of the posterior moments, i.e., of $E(B_j^c|\boldsymbol{y})$ for $c \geq 0$. Christiansen and Morris (1997) showed that the Beta approximation for the shrinkage factors worked better than Normal approximation using the Poisson hierarchical model when the sample size was small, and Morris and Tang (2011) showed that using the Normal hierarchical model.

The ADM assumes that the shrinkage factors follow Beta distributions a posteriori as

$$B_j|\boldsymbol{y} \sim Beta(a_{1j}, a_{0j}), \text{ for } j = 1, 2, \ldots, k, \tag{28}$$

and the ADM estimates the two parameters of the Beta distribution, i.e., a_{1j} and a_{0j}.

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

$$E(B_j|\boldsymbol{y}) = \frac{a_{1j}}{a_{1j} + a_{0j}} = \arg\max_{B_j} A(B_j|\boldsymbol{y}) \equiv B_j^*, \tag{29}$$

$$Var(B_j|\boldsymbol{y}) = \frac{B_j^*(1 - B_j^*)}{a_{1j} + a_{0j} + 1} = \frac{B_j^*(1 - B_j^*)}{B_j^*(1 - B_j^*)[-\frac{d^2}{dB_j^2} \log(A(B_j|\boldsymbol{y}))|_{B_j = B_j^*}] + 1}. \tag{30}$$

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 Laplace approximation with a Lebesque measure on $\boldsymbol{\beta}$ and that for the Poisson model is specified in Equation 14; see Berger, Liseo, Wolpert *et al.* (1999) for the integrated likelihood in detail. Considering that Equation 29 and 30 involve the maximization and Hessian calculation, we work on a logarithmic scale of r, i.e., $\alpha = -log(r)$, because the distribution of α is more symmetric than that of r and α is defined on a real line without any boundary issues. Since $A(B_j|\boldsymbol{y})$ is proportional to the marginal posterior density $f(\alpha|\boldsymbol{y}) \propto e^{\alpha} L(\alpha)$, the estimated posterior mean in Equation 29 is

$$\hat{B}_j^* = \frac{e^{-\hat{\alpha}}}{n_j + e^{-\hat{\alpha}}}, \tag{31}$$

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

We need the invariance information (Morris and Tang 2011) to estimate the variance in Equation 30, which is defined as

$$\begin{aligned} \text{inv.info} &\equiv -\frac{d^2 \log(A(B_j|\boldsymbol{y}))}{d[logit(B_j)]^2}\bigg|_{B_j=\hat{B}_j^*} = -\frac{d^2 \log(A(B_j(r)|\boldsymbol{y}))}{d[\log(r)]^2}\bigg|_{r=\hat{r}} \\ &= -\frac{d^2 \log(A(B_j(r(\alpha))|\boldsymbol{y}))}{d\alpha^2}\bigg|_{\alpha=\hat{\alpha}} \end{aligned} \tag{32}$$

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 posterior variance in Equation 30 as

$$\widehat{Var}(B_j|\boldsymbol{y}) = \frac{\hat{B}_j^{*2}(1-\hat{B}_j^*)^2}{inv.info + \hat{B}_j^*(1-\hat{B}_j^*)}. \tag{33}$$

After matching the estimated unconditional posterior mean and variance of shrinkage factor j in Equation 31 and 33 to the two parameters of the Beta distribution in Equation 28, i.e., a_{1j} and a_{0j}, we get their estimates as

$$\hat{a}_{1j} = \frac{\text{inv.info}}{1-\hat{B}_j^*} \quad \text{and} \quad \hat{a}_{0j} = \frac{\text{inv.info}}{\hat{B}_j^*}. \tag{34}$$

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

$$\widehat{E}(B_j^c|\boldsymbol{y}) = \frac{B(\hat{a}_{1j}+c, \hat{a}_{0j})}{B(\hat{a}_{1j}, \hat{a}_{0j})}, \tag{35}$$

and the approximation gets better if the true posterior distribution of the shrinkage factor is closer to the Beta distribution (Christiansen and Morris 1997; Morris and Tang 2011; Morris and Lysy 2012).

Unconditional posterior moments of expected random effects for the Binomial model.

We assume that the conditional posterior distribution of each expected random effect given $\hat{\alpha}$ follows a Beta distribution, and estimate the assumed Beta distribution using the first and second derivatives of the conditional posterior distribution of β given $\hat{\alpha}$. We estimate the conditional posterior moments of the expected random effects and treat them as their estimated unconditional posterior moments obtained by the (Laplace) approximate marginal likelihood function of $\boldsymbol{\beta}$. We found that the conditional and unconditional posterior moment estimates were almost identical due to their small covariance between r and $\boldsymbol{\beta}$ in the observed information matrix. In addition, calculating the conditional posterior moments was computationally less burdensome.

The moments of the expected random effects $(p_1^E, p_2^E, \ldots p_k^E)$ involve an intractable integration. For example, the first conditional posterior moment is

$$E(p_j^E|\hat{\alpha}, \boldsymbol{y}) = E\left(\frac{e^{x_j^\top \boldsymbol{\beta}}}{1+e^{x_j^\top \boldsymbol{\beta}}}\bigg|\hat{\alpha}, \boldsymbol{y}\right) = \int_{\mathbf{R}^m} \frac{e^{x_j^\top \boldsymbol{\beta}}}{1+e^{x_j^\top \boldsymbol{\beta}}} f(\beta|\hat{\alpha}, \boldsymbol{y}) d\boldsymbol{\beta}. \tag{36}$$

Considering that the expected random effects are a function of logistic regression coefficients $\boldsymbol{\beta}$, we can use the Delta method for the asymptotic Normal distribution of $p_j^E(\hat{\boldsymbol{\beta}}_{MLE})$. However, the Normal approximation goes through the same support and boundary issues as the shrinkage factors do. Instead, we approximate the conditional posterior distribution of $p_j^E(\boldsymbol{\beta})$ by a Beta distribution using $\hat{\beta}_{MLE}$ and the Hessian value at the MLE.

We assume the conditional posterior distribution of expected random effect j is a Beta distribution as follows;

$$p_j^E|\hat{\alpha}, \boldsymbol{y} = \frac{e^{x_j^\top \boldsymbol{\beta}}}{1+e^{x_j^\top \boldsymbol{\beta}}}\Bigg|\hat{\alpha}, \boldsymbol{y} \sim \text{Beta}(b_{1j}, b_{0j}) \sim \frac{G(b_{1j})}{G(b_{1j})+G(b_{0j})}, \tag{37}$$

where $G(b_{1j})$ is a random variable following a Gamma$(b_{1j}, 1)$ distribution with a unit scale and independently $G(b_{0j})$ has a Gamma$(b_{0j}, 1)$ distribution. Note that the representation in Equation (37) is equivalent to saying $e^{x_j^\top \boldsymbol{\beta}}|\hat{\alpha}, \boldsymbol{y} \sim G(b_{1j})/G(b_{0j})$, a ratio of two independent Gamma random variables. Its mean and variance are

$$E(e^{x_j^\top \boldsymbol{\beta}}|\hat{\alpha}, \boldsymbol{y}) = E\left(\frac{G(b_{1j})}{G(b_{0j})}\right) = \frac{b_{1j}}{b_{0j}-1} \equiv \eta_j, \tag{38}$$

$$Var(e^{x_j^\top \boldsymbol{\beta}}|\hat{\alpha}, \boldsymbol{y}) = Var\left(\frac{G(b_{1j})}{G(b_{0j})}\right) = \frac{\eta_j(1+\eta_j)}{b_{0j}-2}. \tag{39}$$

In order to estimate b_{1j} and b_{0j}, we assume that the conditional posterior distribution of β follows $N[\hat{\beta}, \hat{\Sigma}]$, where $\hat{\boldsymbol{\beta}}$ is the mode of $p(\boldsymbol{\beta}|\hat{\alpha}, \boldsymbol{y})$ and $\hat{\Sigma}$ is a negative Hessian matrix at the mode. Then 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, we get the estimates of posterior mean and variance in Equation (38) and (39) as

$$\hat{\eta}_j = e^{x_j^\top \hat{\boldsymbol{\beta}} + x_j^\top \hat{\Sigma} x_j/2}, \tag{40}$$

$$\widehat{Var}(e^{x_j^\top \boldsymbol{\beta}}|\boldsymbol{y}) = \hat{\eta}_j^2\big(e^{x_j^T \hat{\Sigma} x_j} - 1\big). \tag{41}$$

By matching the estimated mean and variance in Equation (40) and (41) to b_{1j} and b_{0j} in Equation (38) and (39), we obtain the estimates of b_{1j} and b_{0j} as follows;

$$\hat{b}_{1j} = \hat{\eta}_j + \frac{\hat{\eta}_j + 1}{e^{x_j^T \hat{\Sigma} x_j} - 1} \quad \text{and} \quad \hat{b}_{0j} = \frac{\hat{\eta}_j + 1}{\hat{\eta}_j(e^{x_j^T \hat{\Sigma} x_j} - 1)} + 2. \tag{42}$$

The conditional posterior moments of expected random effects can be estimated similarly to those of shrinkage factors; $E((p_j^E)^d|\hat{\alpha}, \boldsymbol{y}) = B(b_{1j}+d, b_{0j})/B(b_{1j}, b_{0j})$ for $d \geq 0$. Finally, we estimate the unconditional posterior moments of expected random effects by their estimated conditional posterior moments as follows.

$$\widehat{E}((p_j^E)^d|\boldsymbol{y}) = \frac{B(\hat{b}_{1j}+d, \hat{b}_{0j})}{B(\hat{b}_{1j}, \hat{b}_{0j})}. \tag{43}$$

3.3. Estimation for random effects

It is intractable to derive the unconditional posterior distribution of each random effect analytically for both Binomial and Poisson models. Instead, we again assume that each random effect of the Binomial model has a Beta distribution as

$$p_j|\boldsymbol{y} \sim \text{Beta}(t_{1j}, t_{0j}), \text{ for } j = 1, 2, \ldots, k, \tag{44}$$

and each random effect of the Poisson model has a Gamma distribution as

$$\lambda_j|\boldsymbol{y} \sim \text{Gamma}(s_{1j}, s_{0j}), \text{ for } j = 1, 2, \ldots, k. \tag{45}$$

The unconditional posterior mean and variance of each random effect for the Binomial model specified in Equation 19 and 23 and those for the Poisson model in Equation 24 and 27 are functions of the unconditional posterior moments of shrinkage factors and expected random effect. Once we plug-in the estimated unconditional posterior moments of shrinkage factors in Equation 35 and those of expected random effect in Equation 43, we get the estimates of the unconditional posterior mean and variance of each random effect in Equation 19 and 20 denoted by $\hat{\mu}_{p_j}$ and $\hat{\sigma}^2_{p_j}$, respectively. For the Poisson model, let $\hat{\mu}_{\lambda_j}$ and $\hat{\sigma}^2_{\lambda_j}$ denote the estimates of the unconditional posterior mean and variance in Equation 24 and 27. The estimates of two parameters t_{1j} and t_{0j} in Equation 44 come as follows;

$$\hat{t}_{1j} = \left(\frac{\hat{\mu}_{p_j}(1-\hat{\mu}_{p_j})}{\hat{\sigma}^2_{p_j}} - 1\right)\hat{\mu}_{p_j}, \text{ and } \hat{t}_{0j} = \left(\frac{\hat{\mu}_{p_j}(1-\hat{\mu}_{p_j})}{\hat{\sigma}^2_{p_j}} - 1\right)(1-\hat{\mu}_{p_j}). \tag{46}$$

The estimates of two parameters s_{1j} and s_{0j} in Equation 45 are

$$\hat{s}_{1j} = \frac{\hat{\mu}^2_{\lambda_j}}{\hat{\sigma}^2_{\lambda_j}}, \text{ and } \hat{s}_{0j} = \frac{\hat{\mu}_{\lambda_j}}{\hat{\sigma}^2_{\lambda_j}}. \tag{47}$$

Finally, the approximate unconditional posterior distribution of random effect j for the Binomial model is

$$p_j|\boldsymbol{y} \dot{\sim} \text{Beta}(\hat{t}_{1j}, \hat{t}_{0j}), \tag{48}$$

and that for the Poisson model is

$$\lambda_j|\boldsymbol{y} \dot{\sim} \text{Gamma}(\hat{s}_{1j}, \hat{s}_{0j}). \tag{49}$$

The ADM approximation is accurate if the true distribution of p_j is close to the the Beta distribution (Christiansen and Morris 1997; Morris and Tang 2011; Morris and Lysy 2012).

We use the posterior mean and (2.5%, 97.5%) quantiles (if we assign 95% confidence level) of the posterior distribution in Equation 48 for the Binomial model or in Equation 49 for the Poisson model as the point and interval estimates of the random effect.

4. The acceptance-rejection method for the Binomial model

As for the Binomial model, the package **Rgbp** also provides a way to independently draw exact posterior samples of random effects and hyper-parameters via the acceptance-rejection

method. We continue working on a logarithmic scale of r, $\alpha = \log(1/r) = -\log(r)$. The joint posterior density function of α and $\boldsymbol{\beta}$ based on their joint hyper-prior density function in Equation 12 is

$$f(\alpha, \boldsymbol{\beta}|\boldsymbol{y}) \propto f(\alpha, \boldsymbol{\beta}) L(\alpha, \boldsymbol{\beta}) \propto e^{\alpha} L(\alpha, \boldsymbol{\beta}) d\alpha d\boldsymbol{\beta}. \tag{50}$$

The Acceptance-Rejection (A-R) method (Everson and Morris 2000; Tang 2002) 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 independently 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 $e^{-\alpha(k-1)}$ when α goes to ∞ and as e^{α} 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 it 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, since $f(\boldsymbol{\beta}|\alpha, \boldsymbol{y})$ of the approximate Gaussian counterpart 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

$$g_1(\alpha) = g_1(\alpha; \mu, \sigma, a, b) \equiv \text{Skewed-}t(\alpha|\mu, \sigma, a, b), \tag{51}$$

$$g_2(\boldsymbol{\beta}) = g_2(\boldsymbol{\beta}; \boldsymbol{\mu}^*, S_{(m\times m)}) \equiv t_4(\boldsymbol{\beta}|\boldsymbol{\mu}^*, S), \tag{52}$$

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

$$\mu + \frac{(a-b)\sqrt{a+b}}{\sqrt{(2a+1)(2b+1)}}. \tag{53}$$

It shows that the tails follow a power law as $\alpha^{-(2a+1)}$ 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

$$\alpha \sim \mu + \sigma \frac{\sqrt{a+b}(2T-1)}{2\sqrt{T(1-T)}}, \text{ where } T \sim \text{Beta}(a, b). \tag{54}$$

The notation $t_4(\boldsymbol{\beta}|\boldsymbol{\mu}^*, S)$ in Equation 52 indicates a density function of a multivariate t-distribution at $\boldsymbol{\beta}$ with 4 degrees of freedom, a location vector $\boldsymbol{\mu}^*$, and a $m \times m$ scale matrix S that leads to the variance-covariance matrix $2S$.

We set the parameters of $g_1(\alpha)$ and $g_2(\boldsymbol{\beta})$, i.e., μ, σ, a, b, $\boldsymbol{\mu}^*$, 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 Equation 53 to $\hat{\alpha}$ by fixing the location parameter μ at $\hat{\alpha} - (a - b)\sqrt{a+b}/\sqrt{(2a+1)(2b+1)}$. We set the scale parameter σ to $(-H^{-1}_{\hat{\alpha}})^{0.5}\eta$, where η is a tuning parameter; $\eta = 1.3$ is the default. When the A-R method produces extreme values of weights defined in Equation 56 below, we increase the value of η.

As for $g_2(\boldsymbol{\beta})$, we matches the location vector $\boldsymbol{\mu}^*$ 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 $-H^{-1}_{\hat{\boldsymbol{\beta}}}$;

$$g_2(\boldsymbol{\beta}) \equiv t_4(\boldsymbol{\beta}|\boldsymbol{\mu}^* = \hat{\boldsymbol{\beta}}, S = -H^{-1}_{\hat{\boldsymbol{\beta}}}/2). \tag{55}$$

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

$$w_i \equiv w(\alpha^{(i)}, \boldsymbol{\beta}^{(i)}) = \frac{f(\alpha^{(i)}, \boldsymbol{\beta}^{(i)}|\boldsymbol{y})}{g_1(\alpha^{(i)})g_2(\boldsymbol{\beta}^{(i)})}, \text{ for } i = 1, 2, \ldots, 4N. \tag{56}$$

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. The usual acceptance rates from our data examples are around 25%. 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 sample additional pairs (6 times more 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.

$$f(\boldsymbol{p}|\boldsymbol{y}) = \int f(\boldsymbol{p}|\alpha, \boldsymbol{\beta}, \boldsymbol{y}) \cdot f(\alpha, \boldsymbol{\beta}|\boldsymbol{y}) d\alpha d\boldsymbol{\beta}. \tag{57}$$

The integration can be done by sampling $(p_1, p_2, \ldots, p_k)$ from the independent Beta conditional posterior distributions $f(p_j|\boldsymbol{\beta}, r, \boldsymbol{y})$ in Equation 10 given r $(= e^{-\alpha})$ and $\boldsymbol{\beta}$ already sampled from $f(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ via the A-R method.

5. Frequency method checking

Whether the 95% interval estimates of random effects obtained by a specific model achieve the nominal 95% confidence level for any true parameter values is one of the key model evaluation criteria. A frequency method checking is a procedure to evaluate it, which is different from a model checking that tests whether a two-level model is appropriate for data (overdispersion exists in data) (Dean 1992; Christiansen and Morris 1996). Conditioning that the two-level

model is appropriate, the frequency method checking generates pseudo-data sets given specific values of hyper-parameters (a parametric bootstrapping) and estimates unknown coverage probabilities based on these mock data sets.

From now on, the explanation will be based on the Normal-Normal model because the idea can be easily applied to the other two models.

5.1. 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 Equation 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., $\{\boldsymbol{\mu}^{(i)},\ i = 1, \ldots, N_{sim}\}$, where the superscript (i) indicates the i-th simulation. Next, using the distribution of observed data in Equation 1, we generate N_{sim} sets of observed data sets $\{\boldsymbol{y}^{(i)},\ i = 1, \ldots, N_{sim}\}$ given each $\boldsymbol{\mu}^{(i)}$. Note that we generate one observed data set per one set of random effects.

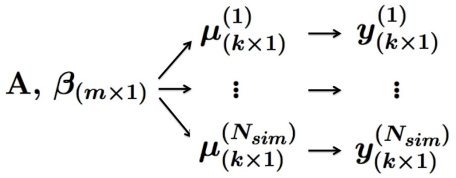

Figure 1: Pseudo-data generating process

5.2. Coverage probability estimation

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

$$I_{A,\boldsymbol{\beta}}(\mu^{(i)}_j) = \begin{cases} 1, & \text{if } \mu^{(i)}_j \in (\hat{\mu}^{(i)}_{j,\ low},\ \hat{\mu}^{(i)}_{j,\ upp}) \\ 0, & \text{otherwise} \end{cases} \tag{58}$$

We consider the coverage indicators as functions of A and $\boldsymbol{\beta}$ because outcomes of indicators depend on the simulated random effects and mock data generated by these hyper-parameters.

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 of 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 Equation 58 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 $C_{A,\boldsymbol{\beta}}(\mu_j)$.

$$\bar{I}_{A,\boldsymbol{\beta}}(\mu_j) = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} I_{A,\boldsymbol{\beta}}(\mu_j^{(i)}),\ j = 1, 2, \ldots, k. \tag{59}$$

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 $\boldsymbol{\beta}$.

The unbiased variance estimator of $Var(\bar{I}_{A,\boldsymbol{\beta}}(\mu_j))$ is

$$\widehat{Var}(\bar{I}_{A,\boldsymbol{\beta}}(\mu_j)) = \frac{1}{N_{sim}(N_{sim}-1)} \sum_{i=1}^{N_{sim}} (I_{A,\boldsymbol{\beta}}(\mu_j^{(i)}) - \bar{I}_{A,\boldsymbol{\beta}}(\mu_j))^2,\ j = 1, 2, \ldots, k. \tag{60}$$

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 effects, which is more efficient than the simple indicator-based coverage estimator. For $j = 1, 2, \ldots, k$,

$$C_{A,\boldsymbol{\beta}}(\mu_j) = E(\bar{I}_{A,\boldsymbol{\beta}}(\mu_j)|A, \boldsymbol{\beta}) = E\left[\frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} E(I_{A,\boldsymbol{\beta}}(\mu_j^{(i)})|A, \boldsymbol{\beta}, \boldsymbol{y}^{(i)}) \middle| A, \boldsymbol{\beta}\right], \tag{61}$$

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

$$\begin{aligned} \bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j) &= \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} E(I_{A,\boldsymbol{\beta}}(\mu_j^{(i)})|A, \boldsymbol{\beta}, \boldsymbol{y}^{(i)}) \\ &= \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} Pr(\mu_j^{(i)} \in (\hat{\mu}^{(i)}_{j,\ low},\ \hat{\mu}^{(i)}_{j,\ upp})|A, \boldsymbol{\beta}, \boldsymbol{y}^{(i)}). \end{aligned} \tag{62}$$

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 Equation 3. The variance of $\bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j)$ is smaller than or equal to the variance of a simple coverage estimator $\bar{I}_{A,\boldsymbol{\beta}}(\mu_j)$ (Rao 1945; Blackwell 1947).

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

$$\widehat{Var}(\bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j)) \equiv \frac{1}{N_{sim}(N_{sim}-1)} \sum_{i=1}^{N_{sim}} \left(E(I_{A,\boldsymbol{\beta}}(\mu_j^{(i)})|A, \boldsymbol{\beta}, \boldsymbol{y}^{(i)}) - \bar{I}^{RB}_{A,\boldsymbol{\beta}}(\mu_j)\right)^2. \tag{63}$$

6. 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.

The basic usage of fitting a Normal-Normal model via the function `gbp` is simply

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

The argument `y` is a vector of k observed sufficient statistics, the argument `se` for the Normal model is a vector of k standard errors of the corresponding sufficient statistics, and the argument `model` indicates that `gbp` fits a Normal-Normal model.

There are many optional arguments for `gbp`.

7. Examples

7.1. Data of 31 hospitals with a known second-level mean

In this example we adopt the perspective of a person living in the state of New York (NY) who has been suffering from severe coronary heart disease. If this person must receive coronary artery bypass graft (CABG) surgery soon, he or she might want to find the most reliable hospital for such a procedure.

For this purpose, data were gathered from 31 hospitals in NY composed of the number of deaths ($\mathbf{z}_{(31\times1)}$) for a specified period after CABG surgeries and the total number of patients ($\mathbf{n}_{(31\times1)}$) receiving CABG surgeries in each hospital. For reference, caseloads (n_j) can be interpreted as exposures. These data can be loaded into R using the following code where the symbol '`R>`' represents a command prompt and is not to be typed into R.

```
R> z <- c( 3,   2,   5,  11,   9,  12,  12,   4,  10,  13,  14,   7,  12,
          11,  13,  22,  15,  11,  14,  11,  16,  14,   9,  15,  13,  35,
          26,  25,  20,  35,  27)
R> n <- c(67,  68, 210, 256, 269, 274, 278, 295, 347, 349, 358, 396, 431,
         441, 477, 484, 494, 501, 505, 540, 563, 593, 602, 629, 636, 729,
         849, 914, 940, 1193, 1340)
```

or

```
R> data(``hospital'')
R> z <- hospital$d
R> n <- hospital$n
```

In addition, suppose one knows that the state-level death rate per exposure of this surgery is 0.030 (λ_0). Using these data and the known second-level mean (λ_0), **Rgbp** provides point and interval estimates of the true death rate (λ_j) so that one can evaluate each hospital's reliability.

The independent Poisson distribution, i.e., $z_j|\lambda_j \overset{ind}{\sim} \text{Poisson}(n_j\lambda_j)$, $j = 1, \ldots, 31$, would be an ideal choice to describe these data based on the Poisson approximation when λ is small and n is relatively large.

Next, `gbp` fits the Poisson hierarchical model with the Gamma conjugate prior distribution as the NY state-level population distribution of the death rates whose mean is 0.030 ($\lambda_0 = 0.030$). For reference, the number of regression coefficients (m) is 0 because we do not need to estimate the prior mean (the second-level mean) via regression for this Poisson-Gamma model.

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

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
colMeans             517       0.03     0.600   0.0201    0.0293   0.0403 0.00517
```

For reference, we need to type '`R> print(p, sort = FALSE)`' instead of '`R> p`' in order to list hospitals by the order of data input in the above output. '`R> p`' automatically sorts the output by the increasing order of caseload (n_j), as shown above.

The output contains information about observed death rates (y_j), caseloads (n_j), known prior mean (λ_0), shrinkage estimates ($\hat{B}_j$), lower bounds of interval estimates ($\hat{\lambda}_{j,low}$, 2.5% percentiles of the approximate posterior distributions if 95% confidence level is given), approximate posterior means ($\hat{\lambda}_j = E(\lambda_j|\boldsymbol{y})$), upper bounds of interval estimates ($\hat{\lambda}_{j,upp}$, 97.5% percentiles of the approximate posterior distributions), and standard deviations of the approximate posterior distributions ($sd(\lambda_j|\boldsymbol{y})$).

As we can see in (6), the posterior mean, $(1 - B_j)y_j + B_j\lambda_0$, is a convex combination of the sample mean and prior mean ($\lambda_0 = 0.030$) with the shrinkage factor, $B_j \equiv r/(r + n_j)$, determining the weight. This makes intuitive sense because r and n_j can be interpreted as the degree (sample sizes) of prior and observed information respectively. If the second level has more information than the first level, i.e., ensemble sample size r exceeds individual sample size n_j, then the estimator will shrink towards the prior mean more than 50%. This is clear because, as caseload increases, shrinkage decreases, depending less on the NY state-level (second-level) information.

A function `summary` shows selective information on hospitals and more detailed estimation results, as below. To be specific, it displays some hospitals (not all as above) with minimum, median, and maximum caseloads (n_j). On top of that, more specific estimation results, such as the posterior mode and standard deviation of $\alpha \equiv \log(1/r)$, follow.

```
R> summary(p)

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
1           -6.53         0.576         684
```

The output of `summary` also provides $\hat{r} = \exp(6.53) = 684$, which is an indicator of how valuable and informative the hypothetical second-level hierarchy is. It means that observed sample means of hospitals whose caseloads are less than 684 will 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 prior distribution is an appropriate choice because the observed amount of information (67) is far less than the amount of state-level information (684).

To obtain a graphical summary the function `plot` can be used, as seen in Figure 2.

```
R> plot(p)
```

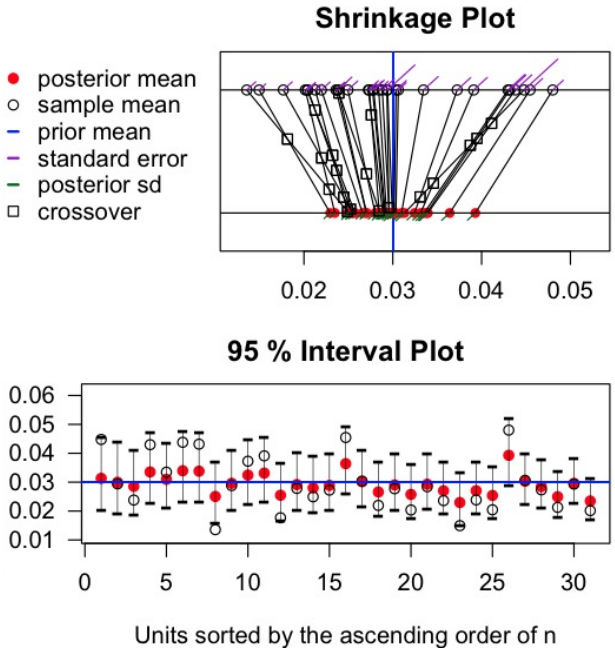

Figure 2: Shrinkage plot and 95% interval plot for 31 hospitals

In Figure 2 the regression towards the mean (RTTM) is obvious in the first graph; the observed sample means, empty dots on the upper horizontal line, are shrinking towards the known second-level mean (a blue vertical line at 0.030) to the different extents. Note that some hospitals' ranks have changed by shrinking much harder towards 0.030 than others. For example, the empty square symbol at the crossing point of the two left-most lines (8th and 23rd hospitals on the list above) indicates that the seemingly safest hospital among 31 hospitals in terms of the observed death rate is probably not the safest in terms of the estimated posterior means.

Intuitively, switching ranks for these two hospitals makes sense. 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 each. 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 death rate estimate for the former hospital shrink toward the state-level mean (λ_0=0.030) 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 bigger uncertainty (variance) than the other. In the second plot of Figure 2, the estimated 95% intervals are displayed. We see that each

posterior mean (red dot) is between the sample mean (empty dot) and the second-level mean (a blue horizontal line). For reference, we could plot this 95% interval plot by the order of data input via `plot(p, sort = FALSE)`.

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 bigger 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 NY state-level hierarchy (or shrinking less toward the state-level mean, 0.030) with smaller variance. Based on the point and interval estimates, the 31st hospital seems the most reliable one among all candidates.

When fitting a model it is always a good idea to question how reliable the estimation procedure is. For example, does our procedure generate interval estimates that have good repeated sampling properties? To answer this question the `coverage` function generates pseudo-datasets assuming the estimated r (= 683.53) is a true value. For reference, we can designate any other value of r, for example $r = 600$, by adding another argument, `A.or.r = 600`, into the code below; `R> pcv <- coverage(p, A.or.r = 600, nsim = 1000)`.

In addition, `gbp` also provides interval estimates with different confidence levels, for example 90%. For this, we need to go back to the code for fitting the model, adding another argument, `Alpha = 0.9`; `R> p <- gbp(z, n, Alpha = 0.9, mean.PriorDist = 0.03, model = "poisson")`. Then, the code below will evaluate whether interval estimates achieve the 90% confidence level.

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

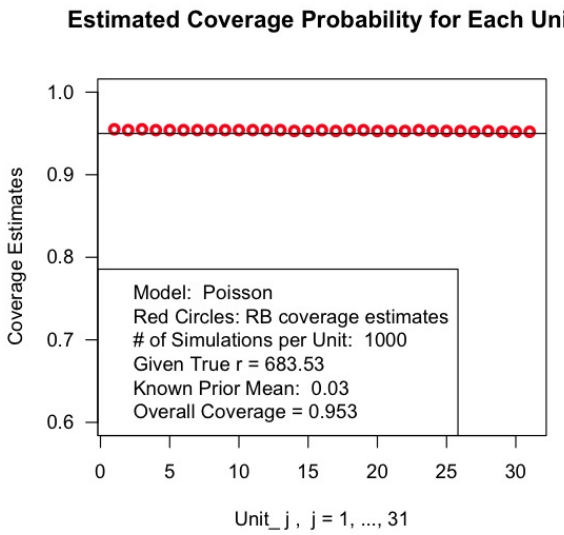

Figure 3: Coverage plot via frequency method checking for 31 hospitals

In Figure 3, the black horizontal line at 0.95 represents the nominal confidence level and the red circles indicate Rao-Blackwellized (RB) unbiased coverage estimates for 31 hospitals. The overall RB unbiased coverage estimate across all the hospitals is 0.953. And none of RB

unbiased coverage estimates for 31 hospitals are less than 0.95 regardless of their caseloads (n_j). This result shows that the interval estimates for this particular dataset accurately achieves 95% confidence under repeated sampling. Note that these estimates depend on the given true value of r, the known prior mean, and the assumption that the model is true.

The following code provides 31 RB unbiased coverage estimates for each hospital.

```
R> pcv$coverageRB
```

```
 [1] 0.955 0.954 0.955 0.954 0.954 0.954 0.954 0.954 0.954 0.954 0.954 0.954
[13] 0.954 0.953 0.953 0.954 0.953 0.954 0.954 0.953 0.953 0.953 0.954 0.953
[25] 0.953 0.953 0.952 0.953 0.952 0.952 0.952
```

And the code below shows 31 simple unbiased coverage estimates for each hospital.

```
R> pcv$coverageS
```

```
 [1] 0.949 0.960 0.958 0.958 0.950 0.945 0.960 0.953 0.960 0.956 0.955 0.946
[13] 0.955 0.954 0.960 0.965 0.955 0.952 0.960 0.956 0.959 0.955 0.964 0.960
[25] 0.945 0.942 0.947 0.960 0.940 0.946 0.956
```

The function `coverage` also calculates the standard errors for each hospital's RB unbiased coverage estimate defined in (63). The following code provides 31 standard errors for RB estimates.

```
R> pcv$se.coverageRB
```

```
 [1] 0.0016 0.0016 0.0014 0.0014 0.0013 0.0013 0.0013 0.0013 0.0012 0.0013 0.0012
[12] 0.0012 0.0011 0.0011 0.00110.0011 0.0011 0.0011 0.0011 0.0011 0.0010 0.0010
[23] 0.0010 0.0010 0.0010 0.0009 0.0009 0.0008 0.0008 0.0007 0.0007
```

Similarly, 31 standard errors for each simple unbiased coverage estimate defined in (60) are

```
R> pcv$se.coverageS
```

```
 [1] 0.0070 0.0062 0.0063 0.0063 0.0069 0.0072 0.0062 0.0067 0.0062 0.0065 0.0066
[12] 0.0072 0.0066 0.0066 0.0062 0.0058 0.0066 0.0068 0.0062 0.0065 0.0063 0.0066
[23] 0.0059 0.0062 0.0072 0.0074 0.0071 0.0062 0.0075 0.0072 0.0065
```

Taking the first hospital as an example, the variance estimate of RB unbiased coverage estimate is about 19 times smaller than that of simple one. It means that 1,000 RB unbiased coverage estimates are as precise as 19,000 simple unbiased coverage estimates in terms of estimating true coverage probability for the first hospital, $p_{cov,1}$.

For reference, two $31 \times 1,000$ matrices `raw.resultRB` and `raw.resultS`, each row of which is about each hospital, in `pcv` contain all the individual estimates, $I_j^{(i)}$ and $E(I_j^{(i)}|y_j^{(i)}, A, \boldsymbol{\beta})$.

Morris and Christiansen (1995) also investigated a similar ranking problem in hierarchical modeling, taking shrinkage into account.

7.2. Data of 8 schools with unknown second-level mean with 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).

```
R> y  <- c(12, -3, 28,  7,  1,  8, 18, -1)
R> se <- c(18, 16, 15, 11, 11, 10, 10,  9)
```

or

```
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 assumed to be 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 Normal-Normal hierarchical model.

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

Summary for each unit (sorted by se):

          obs.mean   se prior.mean shrinkage low.intv post.mean upp.intv post.sd
5            -1.00  9.0      8.168     0.408  -13.297     2.737   16.692   7.634
2             8.00 10.0      8.168     0.459   -7.255     8.077   23.361   7.810
7            18.00 10.0      8.168     0.459   -1.289    13.484   30.821   8.176
4             7.00 11.0      8.168     0.507   -8.780     7.592   23.602   8.257
6             1.00 11.0      8.168     0.507  -13.027     4.633   20.131   8.441
1            28.00 15.0      8.168     0.657   -2.315    14.979   38.763  10.560
3            -3.00 16.0      8.168     0.685  -17.130     4.650   22.477  10.096
8            12.00 18.0      8.168     0.734  -10.208     9.189   29.939  10.227
colMeans          12.5      8.168     0.552   -9.163     8.168   25.723   8.900
```

This output from `gbp` summarizes the results. In this Normal-Normal 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)
```

```
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
1            4.77          1.14         118


Regression Summary:

      estimate   se z.val p.val
beta0    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 second level mean, `beta0`, 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)
```

The frequency method checking, assuming the model is correct, generates new pseudo-data from our assumed model. Unless otherwise specified, the procedure fixes the hyper-parameter values at their estimates ($\hat{A}$ and $\hat{\boldsymbol{\beta}}_0$ in this example) and then simulates "true" θ_j for each group j. The model is then estimated and this is repeated an `nsim` number of times to estimate the coverage probabilities of the procedure.

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

As seen in Figure 5 the desired 95% confidence (black horizontal line at 0.95) is achieved (actually, exceeded) for each school in this example. Note that all the coverage estimates depend on the chosen true values of A and β_0, and the assumption that the model is valid.

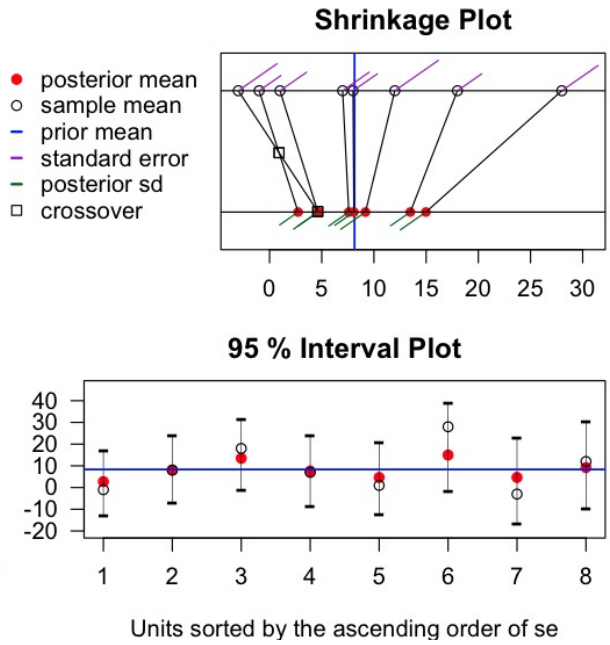

Figure 4: Shrinkage plot and 95% interval plot for 8 schools

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

```
R> gcv$coverageRB

 [1] 0.966 0.959 0.967 0.960 0.959 0.962 0.960 0.966

R> gcv$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, `gcv$raw.resultRB`, each row of which is about each school.

7.3. Data of 18 baseball players with unknown second-level mean and one covariate

The following dataset from the New York Times published on 26 April 1970 contains information on the batting averages and positions (outfielder=1,otherwise=0) of 18 major league baseball players through their first 45 official at-bats of the 1970 season (Efron and Morris 1975).

```
R> z <- c(18, 17, 16, 15, 14, 14, 13, 12, 11, 11, 10, 10, 10, 10, 10,  9,  8,  7)
R> n <- c(45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45)
R> x <- c( 1,  1,  1,  1,  1,  0,  0,  0,  0,  1,  0,  0,  0,  1,  1,  0,  0,  0)
```

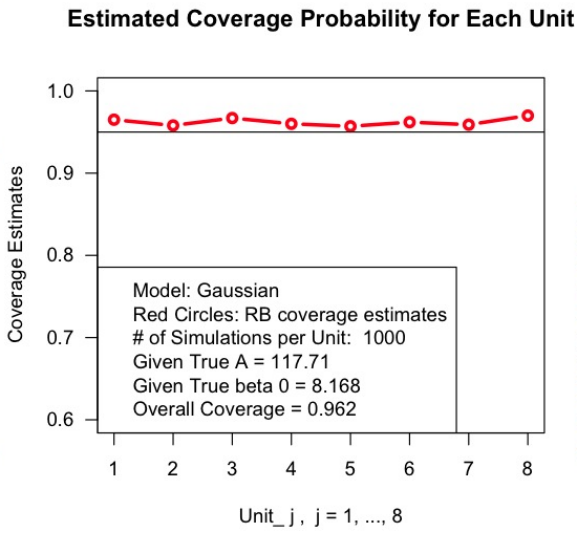

Figure 5: Coverage plot via frequency method checking for 8 schools

or

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

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

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

```
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 1.00      0.310     0.715    0.248     0.335    0.429  0.0462
2           0.378 45 1.00      0.310     0.715    0.244     0.329    0.420  0.0448
3           0.356 45 1.00      0.310     0.715    0.240     0.323    0.411  0.0437
4           0.333 45 1.00      0.310     0.715    0.236     0.316    0.403  0.0429
5           0.311 45 1.00      0.310     0.715    0.230     0.310    0.396  0.0424
6           0.311 45 0.00      0.233     0.715    0.179     0.256    0.341  0.0415
7           0.289 45 0.00      0.233     0.715    0.175     0.249    0.331  0.0400
```

```
8           0.267 45 0.00      0.233     0.715    0.171     0.243    0.323  0.0388
9           0.244 45 0.00      0.233     0.715    0.166     0.237    0.315  0.0380
10          0.244 45 1.00      0.310     0.715    0.210     0.291    0.379  0.0432
11          0.222 45 0.00      0.233     0.715    0.161     0.230    0.308  0.0377
12          0.222 45 0.00      0.233     0.715    0.161     0.230    0.308  0.0377
13          0.222 45 0.00      0.233     0.715    0.161     0.230    0.308  0.0377
14          0.222 45 1.00      0.310     0.715    0.202     0.285    0.375  0.0441
15          0.222 45 1.00      0.310     0.715    0.202     0.285    0.375  0.0441
16          0.200 45 0.00      0.233     0.715    0.155     0.224    0.302  0.0377
17          0.178 45 0.00      0.233     0.715    0.148     0.218    0.297  0.0381
18          0.156 45 0.00      0.233     0.715    0.140     0.211    0.292  0.0389
colMeans          45 0.44      0.267     0.715    0.191     0.267    0.351  0.0410
```

Note that the shrinkage estimates are the same for all players due to the fact that they are determined solely by the relative amount of information between the first-level and the second-level hierarchies, ($\hat{B}_j \equiv \hat{r}/(\hat{r}+45) = 113/(113+45) = 0.715$).

```
R> summary(b)

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
1           -4.73         0.957         113


Regression Summary:

      estimate    se  z.val p.val
```

```
beta0   -1.194 0.131 -9.129 0.000
beta1    0.389 0.187  2.074 0.038
```

From the `Regression Summary`, one of the outputs of `summary`, we see that the two prior means distinguishing outfielders from other positions are significantly different (p-value for $\hat{\beta}_1$= 0.038). Also, the positive sign of $\hat{\beta}_1$ indicates that the population mean batting average for all outfielders tends to be higher than that for those in the other positions (estimated odds ratio = exp(0.389)=1.48).

```
R> plot(b)
```

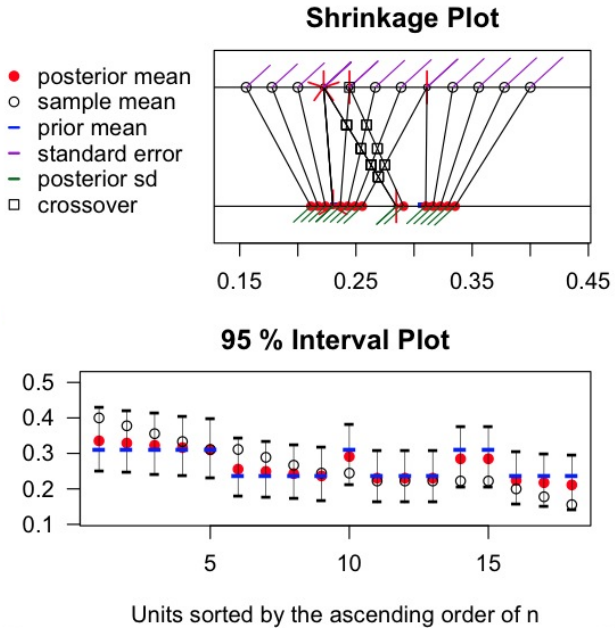

Figure 6: Shrinkage plot and 95% interval plot for 18 baseball players

It is evident in the shrinkage plot in Figure 6 that shrinkage occurs from the sample means (empty dots) on the upper horizontal line towards the two prior means, 0.233 and 0.310. For reference, the short red line symbols on 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 sample mean (0.222) and at this point on the upper horizontal line, there are short red lines toward five directions.

The 95% interval plot 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 likely be attributed to their bad luck because their observed batting averages are close to the lower bounds of their interval estimates. RTTM suggests that their batting averages will shrink towards the expected prior mean of outfielders (0.310) in the long run.

As in the previous examples in Section 7.1 and 7.2, in order to check the level of trust in these interval estimates, we can proceed to frequency method checking by assuming the estimates, 112.95 for $\hat{r}$ and (-1.194, 0.389) for $\hat{\boldsymbol{\beta}}$, are given values.

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

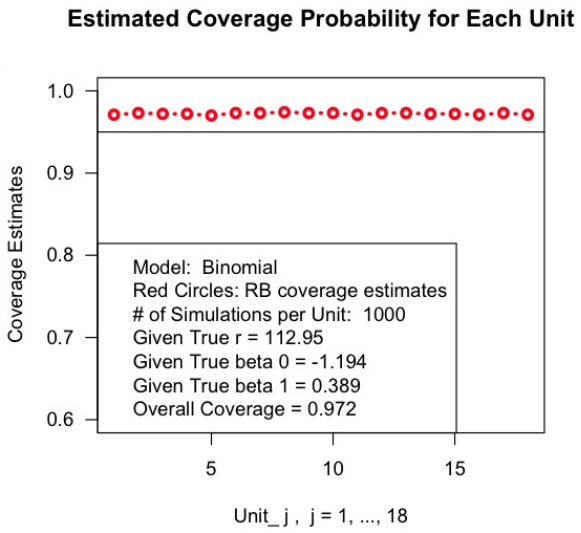

Figure 7: Coverage plot via frequency method checking for 18 players

For reference, to do the frequency method checking at different true values, we need to specify additional arguments in the coverage function. For example, if we want to try different true values, either 100 for r or (-1, 0.2) for (β_0, β_1), the additional arguments are `A.or.r = 100` and `reg.coef = c(-1, 0.2)`; `coverage(b, A.or.r = 100, reg.coef = c(-1, 0.2), nsim = 1000)`.

Finally, in Figure 7, we see that the overall Rao-Blackwellized unbiased coverage estimate is 0.972 (across all the players), conservatively satisfying the definition of the 95% confidence interval. Note that each coverage estimate depends on given true values of r and $\beta_{(2\times1)}$, and the assumption that the model is valid.

The Rao-Blackwellized unbiased coverage estimates and their standard errors for each player follow.

```
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
```

All the simulation results are saved in the $18 \times 1,000$ matrix, `bcv$raw.resultRB`, each row of which is for each player.

8. Discussion and summary

Rgbp is an R package for estimating and validating two-level Gaussian, Binomial and Poisson 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 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 from (Morris 1988b).

A benefit of **Rgbp** is that it produces non-random output 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) 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.

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.

In addition to the built in "frequency method checking" procedure the package can be used to undergo "model checking". For example, in the Gaussian hierarchical model, the assumed marginal distribution of the data is given in (**??**). By substituting the point estimates of A and $\boldsymbol{\beta}$ from the package into this marginal distribution a test can be constructed to see whether the data follow the marginal distribution suggested by the hierarchical model.

9. 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.

A. Unconditional posterior variance of the Binomial model

B. Unconditional posterior variance of the Poisson model

References

Berger JO, Liseo B, Wolpert RL, *et al.* (1999). "Integrated likelihood methods for eliminating nuisance parameters." *Statistical Science*, **14**(1), 1–28.

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, Su YS, Yajima M, Hill J, Pittau MG, Kerman J, Zheng T (2014). "arm: data analysis using regression and multilevel/hierarchical models, 2010." *URL http://CRAN. R-project. org/package= arm. R package version*, pp. 1–3.

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.

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.

Patterson HD, Thompson R (1971). "Recovery of inter-block information when block sizes are unequal." *Biometrika*, **58**(3), 545–554.

Rao CR (1945). "Information and accuracy attainable in the estimation of statistical parameters." *Bulletin of the Calcutta Mathematical Society*, **37**(3), 81–91.

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." *R package version*, **1**.

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." *in preparation*.

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