# A Bayesian Framework for Generalized Linear Mixed Models in Genome-Wide Association Studies

Xulong Wang, Gregory W. Carter

February 11, 2016

## Methods

### Overview of the statistical models

Statistical models in Bayesian framework are defined by two parts: a likelihood function to describe the data-generating process, and the prior distributions of the likelihood function's parameters. To model binary and categorical variables, likelihood functions in bayes-glmm took the form of logistic regression model (LR) and ordered logistic regression model (OLR), respectively. In LR, the 0/1 response variable $Y_i$ followed a binomial distribution with a scalar parameter $\pi$ representing the probability that $Y_i$ equaled 1. $\pi$ was further transformed by the logit function and modeled in the linear model scheme.

$$\begin{aligned}
\pi &= P(Y_i = 1) && (1)\\
logit(\pi) &= \mathbf{X}\beta + g\beta_0 + u && (2)\\
u &\sim mvN(0, \sigma K) && (3)\\
\beta &\sim N(0, 1) && (4)\\
\beta_0 &\sim N(t * \sigma_0, \sigma_0) && (5)\\
t &\sim N(prior, 1) && (6)\\
\sigma_0 &\sim inv\_gamma(2, 1) && (7)\\
\sigma &\sim inv\_gamma(2, 1) && (8)
\end{aligned}$$

In the above equations, $X$ was a $n$ by $m$ covariate matrix with $n$ the sample size and $m$ the number of conditional variables. $\beta$ was the corresponding parameter vector in length $m$. $g$ was the numerical genotype of a variant with 0 to 2 representing homozygous reference alleletype, heterozygous, and homozygous alternative alleletype, respectively. $\beta_0$ was the variant's effect size. Prior distribution of $\beta_0$ followed $N(\mu/\eta, \sigma/\eta)$. To integrate the reported effect of a given variant, $\mu$ and $\sigma$ took the reported effect size and its standard error, normalized by a defined $\eta$ to match the scales. A standard normal, $N(0, 1)$, was used for variants with no known effects. Further, $\beta$ followed $N(0, 1)$ in prior.

To model the sample relatedness, $u$ was included as a random term that followed a multivariate normal distribution, $mvN(0, K)$ in prior, with expected mean vector 0 and covariance matrix $K$. $K$ was the kinship matrix of the samples. $mvN(0, K)$ was parameterized by the Cholesky factoring of $K$ and $n$ independent standard normal distributions.

$$\begin{aligned} u &= L * z && (9) \\ L &= Chol(K) && (10) \\ z &\sim mvN(0, \sigma I) && (11) \end{aligned}$$

In OLR, the categorical response variable $Y_i$ with $J$ levels followed a multinomial distribution with a vector of parameters $\pi$, where $\pi_{ij}$ represents the probability that the $i$th observation falls in response category $j$. Cumulative distribution of $\pi$ was logit-transformed and modeled in the linear model scheme.

$$\begin{aligned} P(Y_i \leq j) &= \pi_{i1} + ... + \pi_{ij} && (12) \\ logit(P(Y_i \leq j)) &= \theta_j - \mathbf{X}\beta - g\beta_0 + u \quad j = 1, ..., J-1 && (13) \\ \theta &= 10 * cumsum(\theta_0) && (14) \\ \theta_0 &\sim dirichlet(1) && (15) \end{aligned}$$

$\theta_j$ modeled the distances between the categories by providing each category an unique intercept. $\theta$ was defined as ten times the cumulative sum of a multivariate variable $\theta_0$, where $\theta_0$ followed a $J - 1$ dimension Dirichlet distribution in prior.

## Model estimations

Our models were built under Stan, which provides a flexible and efficient programming environment for statistical modeling. Inherited from Stan, bayes-glmm support two methods for parameter estimation, L-BFGS maximal likelihood estimation (MLE) and Hamilton Markov chain Monte Carlo (HMC) sampling. MLE method made a point estimation for each parameter that maximized the joint posterior of model parameters, whereas MCMC sampling method captured a full posterior distribution for each parameter by iterative sampling. Significance of the estimated effect size $\beta_0$ can be accessed by combing $\beta_0$ and its standard error $SE(\beta_0)$. Standard errors of MLE were computed as the inverse of the square root of the diagonal elements of the observed Fisher Information matrix (Pawitan, 2001). In MCMC sampling, $SE(\beta_0)$ was computed directly from the samples. A standardized $z$ value was computed as $\beta_0/SE(\beta_0)$, which led to a P-value that quantified the probability of obtaining the $\beta_0$ by chance.

$$SE(\hat{\theta}_{ML}) = \frac{1}{\sqrt{I(\hat{\theta}_{ML})}}) \tag{16}$$

$$I(\theta) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\theta) \quad 1 \leq i, j \leq p \tag{17}$$

$\hat{\theta}_{ML}$ was MLE of model parameters. $I(\theta)$ was the Fisher Information matrix. $p$ was the number of parameters.

In genetic association studies, comparing the two nested null and full models was a widely used method to estimate the significance of a variant. The full models were the same as described above whereas the null models ignored the variant genotypes, $g$, as a linear predictor. In MLE, the null to full model improvements was quantified by LRT, which equals two times the log likelihood difference between the full and null models using the MLE estimation of model parameters. In MCMC sampling, WAIC was quantified as two times the expected log point wise predictive density (elpd) using the MCMC samples.

$$LRT = -2 \cdot (log(P(data|\theta_p^n)) - log(P(data|\theta_p^f))) \tag{18}$$

$$WAIC = -2 \cdot (\sum_{i=1}^{n} log(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s))) \tag{19}$$

$\theta_p^n$ and $\theta_p^f$ were the MLE of the parameter spaces under the null and full models, respectively. $S$ is the sampling iteration number. $\theta^s$ is the actual sample of model parameters in the $s$-th sampling iteration.

To estimate the significance of LRT in bayes-glmm, a P-value was computed by approximating LRT to a $\chi^2$ distribution with degree of freedom 1. Further, standard error of WAIC was computed by a method proposed by Vehtari et al.

z-score and the corresponding P-value was quantified to describe the significance of WAIC.

To control the false discovery rate, P-values of relevant statistics, such as $\beta_0$, LRT, and WAIC, could be corrected by Bonferonni or FDR in bayes-glmm. Apart from the P-values and corrected P-values by parametric approximations, bayes-gwas also provides a permutation strategy to determine the significance thresholds of these statistics. Taking LRT for example, the permutation test followed: (1) Randomly shuffle the sample genotypes; (2) Test the association with the shuffled genotypes; (3) record the maximal LRT; (4) repeat 1-3 $N$ times. The N-sample maximal LRT was fitted by a extreme value distribution which stand for null distribution of the maximal LRT. Random terms of the models were dropped in the permutation test because they are irrelevant of the permutation results (Cheng and Palmer, 2013).

Given the approximation nature of the P-values, we recommend the permutation strategy when computing resource is without concern. Further, while P-values are informative in searching for the causal variants, other variant characteristics such as effect sizes, MAF, evolutionary conservation, and genetic consequences are equally important and should be thoroughly considered in prioritizing follow up variants.

## Kinship matrix

We used $u$ as a random term to account for the sample relatedness. $u$ follows $mvNormal(0, K)$, where K was the kinship matrix of the samples. For each $K$ entry, genotype-based relatedness for the sample pair, or IBS coefficient, was computed using the full spectrum of genomic variants in the ADSP samples.

$$k_{i,j} = \frac{1}{M} \sum_{m=1}^{M} (g_{m,i} \cdot g_{m,j} + (1 - g_{m,i}) \cdot (1 - g_{m,j})) \tag{20}$$

$k_{i,j}$ is the IBS relatedness between sample $i$ and $j$. $M$ is the variant number. $g_{m,i}$ and $g_{m,j}$ is the genotype of variant $m$ in sample $i$ and $j$, respectively.

To avoid over-correction, the kinship matrices were computed by the taking-one-off strategy, in that for any given variant of one chromosome, the corresponding kinship matrix was computed by taking off all variants of the given chromosome. KING was used for fast kinship estimation on the massive genotype data (Manichaikul et al., 2010).

## Linear mixed models

To compare the performances of our method to that of a LMM when response variables are categorical, a LMM was constructed as follow:

$$\begin{aligned} y_i &= \mathbf{X}_i\beta + u + e && (21) \\ u &\sim mvN(0, \delta_g^2 K) && (22) \\ e &\sim N(0, \delta_e^2 I) && (23) \end{aligned}$$

$y_i$ was the numerical transformation of the AD categories: no/0, possible/0.25, probable/0.5, definite/1. $X$ was the covariate matrix including age and sex. $u$ was the random term. $e$ was the model residual. The LMM model was estimated with QTLRel in R (Cheng et al., 2011).

## Code availability