On `predict.MCMCglmm`

We fit the following logistic mixed model below using `MCMCglmm` where $\mathbf{X}$ denote the covariate matrix and $\mathbf{Y}$ the binary trait:

$$\begin{aligned} Y_i|\,\mathbf{X}_i, \boldsymbol{\beta}, u_i &\sim \text{Ber}(\pi_i), \\ \text{logit}(\pi_i) &= \mathbf{X}_i\boldsymbol{\beta} + u_i, \\ \boldsymbol{\beta} &\sim N(0, 10^6\text{I}), \\ \mathbf{u}|\,\sigma_a^2 &\sim N(0, \sigma_a^2\,\Phi), \\ \sigma_a &\sim \text{Half-Cauchy(scale } = 10), \\ \mathbf{u},\, \boldsymbol{\beta} &\text{ and } \mathbf{X} \text{ are independent} \end{aligned} \qquad \text{(Equation 1)}$$

Samples from the joint posterior of $(\boldsymbol{\beta}, \mathbf{u}, \sigma_a^2)$ are obtained through the MCMC-based algorithm.
Note:

$$\mathbb{E}(Y_i|\,\mathbf{X}_i, \boldsymbol{\beta}, \sigma_a^2) = \int \mathbb{E}(Y_i|\,\mathbf{X}_i, \boldsymbol{\beta}, u_i)\, p(u_i|\,\sigma_a^2) du_i$$

and $u_i|\,\sigma_a^2 \sim N(0, \sigma_a^2)$

The numerator of the test statistic is $(\mathbf{y} - \tilde{\boldsymbol{\pi}})^T\mathbf{G}$, and we aim to get prediction for $\tilde{\pi}_i = \mathbb{E}(Y_i|\,\mathbf{X}_i, \boldsymbol{\beta}, \sigma_a^2, \tilde{\mathbf{u}}_i)$ where

$$\tilde{\mathbf{u}} = \arg\max_{\mathbf{u}} p(\boldsymbol{\beta}, \sigma_a^2, \mathbf{u}; \mathbf{Y})$$

. There are three ways to get predictions for $\tilde{\pi}_i$:

1. Estimate $\mathbb{E}(Y_i|\,\mathbf{X}_i, \boldsymbol{\beta}, \sigma_a^2)$ using the fact that

$$\int \text{logit}^{-1}(x)\, \phi(x; \mu, \sigma^2)\, dx \approx \text{logit}^{-1}(\mu/\sqrt{1 + k^2\sigma^2})$$

   where $k = \frac{16\sqrt{3}}{15\pi}$ and $\phi(.\,; \mu, \sigma^2)$ denotes a normal distribution pdf with mean μ and variance σ^2.

2. Get the posterior mode estimates for $(\boldsymbol{\beta}, u_i)$ denoted as $(\boldsymbol{\beta}^*, u_i^*)$ and estimate $\mathbb{E}(Y_i|\mathbf{X}_i, \boldsymbol{\beta}, u_i)$ by $\text{logit}^{-1}(\mathbf{X}_i\boldsymbol{\beta}^* + u_i^*)$

   $\rightarrow$ This approach doesn't directly depend on the estimate of σ_a^2.

3. Get samples from the posterior distribution of $\mathbb{E}(Y_i|\mathbf{X}_i, \boldsymbol{\beta}, u_i) = \text{logit}^{-1}(\mathbf{X}_i\boldsymbol{\beta} + u_i)$ and take the average of these samples.

   $\rightarrow$ We note again that this approach doesn't directly depend on the estimate of σ_a^2.

Note: Because the log(.) is a convex function then logit(.) = log(.) is also convex so its inverse is concave. By Jensen's inequality, we expect the predicted values for $\boldsymbol{\pi}$ obtained in (2) to be lower than those obtained in (3) (assuming fairly symmetric posterior distribution estimate so mode and mean coincide). Thus the residuals $(\mathbf{Y} - \boldsymbol{\pi})$ should be bigger in (2) than in (3).

Questions to think about:

- What are additive polygenic effects $\mathbf{g}$ in the following model?

$$\mathbf{Y} = \mu + \mathbf{g} + \mathbf{e}$$

Consider

$$\mathbf{g} = \sum_k \mathbf{W}_k \gamma_k$$

where $\mathbf{W}_k$ denote the centered vector of minor allele counts for the k-th variant and γ_k denote the genetic effect of the variant on the response (assumed to be small as there are many markers overall – i.e. no one gene has a 'dominating' effect on the trait). So here we have the effect of each marker combined together in an additive fashion thereby the name of polygenic (multiple genes) additive effect.

Note that

$$\mathbb{E}(\mathbf{W}_k) = 0$$

$$\text{Cov}(W_{ik}, W_{jk}) \propto p_k(1-p_k)\Phi_{ij} \Rightarrow \text{Var}(\mathbf{W}_k) \propto p_k(1-p_k)\Phi$$

where p_k is the m.a.f. of k-th variant and Φ_{ij} is twice the kinship coefficient between individuals i and j. We assume that the markers are independent. (i.e. $\text{Cov}(\mathbf{W}_k, \mathbf{W}_l) = 0$ for $i \neq j$)

We then have

$$\mathbb{E}(\mathbf{g}) = \sum_k \mathbb{E}(\mathbf{W}_k)\,\gamma_k = 0,$$

$$\text{Var}(\mathbf{g}) = \sum_k \text{Var}(\mathbf{W}_k)\,\gamma_k^2 \propto \sum_k p_k(1-p_k)\gamma_k^2 \cdot\ \Phi = \sigma^2 \cdot\ \Phi$$

So that σ^2 represents the variance accumulated across the markers and is referred to as the polygenic additive variance. Now $\mathbf{g}$ is the sum of many small independent marker effects so using CLT we get that $\mathbf{g} \sim N(0, \sigma^2\,\Phi)$.

- Identifiability of the random effects $\mathbf{u}$ in a logistic mixed model

$$\textbf{True model: } \text{logit}(\pi_i) = \mathbf{X}_i\,\boldsymbol{\beta} + \underbrace{G_i\gamma + u_i}, \qquad \text{(Equation 2)}$$

$$\textbf{Fitted model: } \text{logit}(\pi_i) = \mathbf{X}_i\,\boldsymbol{\beta} + \quad u_i^* \quad , \qquad \text{(Equation 3)}$$

- Under the alternative, the true model is the one specified in Eqn. 2, where the effect of marker $\mathbf{G}$ is included as a predictor on the logit scale. In this case, the random effect $\mathbf{u}$ captures all of the effects that are not captured by the covariates and the marker of interest. However, in the model specified in Eqn. 3, the signal γ that we are interested in is captured within the random effect $\mathbf{u}^* = \mathbf{G}\gamma + \mathbf{u}$.

  In other words, since the effect being tested is not included as a predictor in the null model, it will be absorbed in the prediction for the random effect $\mathbf{u}^*$. Hence, when the residual is computed using the mean of $\mathbf{Y}$ conditional on $(\mathbf{X}, \mathbf{u}^*)$, this will remove the signal we want to detect (because it is incorporated in $\widehat{\mathbf{u}^*}$, whose effects we are removing by computing the residuals).

  So the predictions for the random effects $\mathbf{u}^*$ actually contain

$$\text{effect of } \mathbf{G} + \text{additive effect of genes other than } \mathbf{G}$$

  So we could consider the following model for the prediction of $\mathbf{u}^*$

$$\widehat{\mathbf{u}^*} = \mu_0 + \mathbf{G}\alpha + \mathbf{g} \qquad \text{(Equation 4)}$$

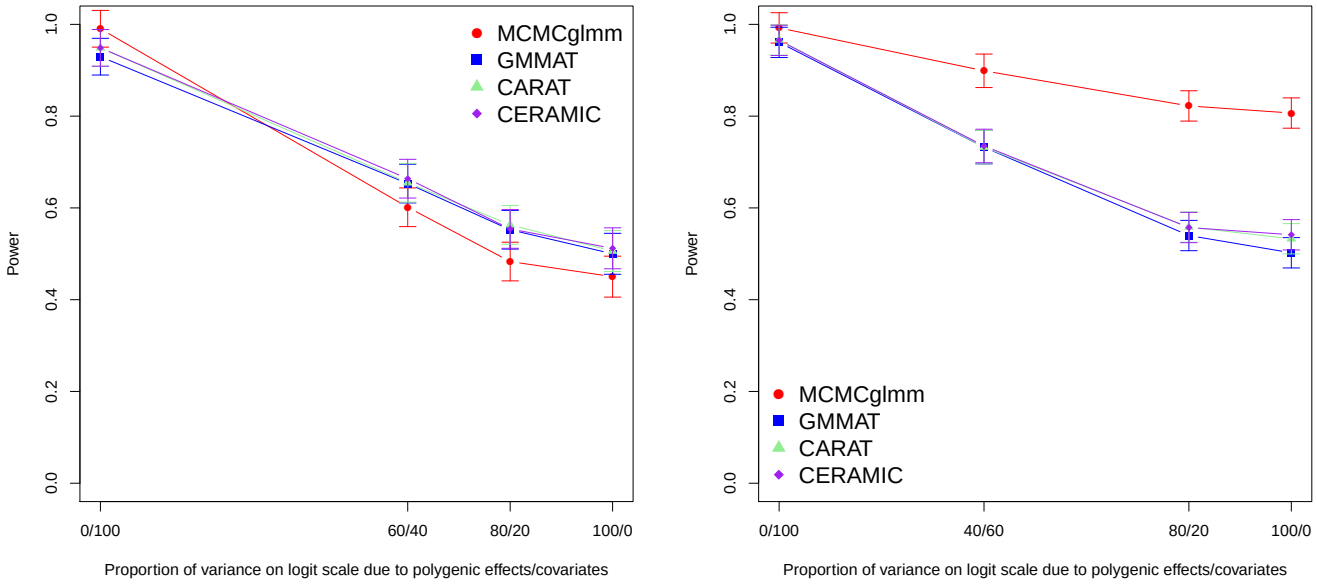
where μ_0 is just an intercept term, α represent the signal that we want to detect (i.e. the effect of $\mathbf{G}$ on the response $\mathbf{Y}$), and $\mathbf{g} \sim N(0, \sigma_g^2 \Phi)$ represents the additive polygenic effect (i.e. the combined effect of genes in the genome). We can use the pedigree relatedness matrix Φ as the correlation matrix for the random effects $\mathbf{g}$ because it captures the expected amount of genetic similarities between related individuals (and so doesn't contain information about the signal between $\mathbf{G}$ and $\mathbf{Y}$). We note that the effects of covariates $\mathbf{X}$ on the binary trait have been removed when obtaining the predictions for $\mathbf{u}^*$ from the model in Eqn. 1, which is why we don't need to include them in Eqn. 4.

→ Am I double counting the effect of $\mathbf{G}$ in Eqn. 4???? Using it both as fixed effect and a small part of it is included in $\mathbf{g}$ through the use of the pedigree relatedness matrix?

→ Can use true values for $(\mathbf{G}\gamma + \mathbf{u})$ and regress that against $\mathbf{G}$ in a linear mixed model with $\mathbf{g}$ and $\mathbf{e}$ included as random effects (use modified MASTOR assuming no i.i.d. errors present).

- Which one of the three estimates makes more sense to use in terms of corresponding to bigger gains in power? Note that only the first approach directly depends on accurate estimation of the additive variance parameter. The idea is that better estimates of σ_a^2 will produce residuals that result in higher power (though no change in type 1 error since well controlled with retrospective variance calculation in denominator of the test statistic)

- For (1) and (2), used the true values for $(\boldsymbol{\beta}, \mathbf{u}, \sigma_a^2)$ as the estimates to compute $\tilde{\boldsymbol{\pi}}$ in order to see if accurate estimation does lead to gains in power.



Plot on left corresponds to using (1) i.e. integrating the conditional mean over the distribution of the random effects. No significant difference in power between 4 methods (type 1 error is well-controlled besides for GMMAT).

Plot on the right corresponds to using (2), i.e. taking predicted random effects and using these to compute the conditional mean of the response. This indicates that power improvement can be obtained through more accurate prediction of the random effects.