The R2D2 Prior for Generalized Linear Mixed Models

Team Members: Thad Creech, Divija Balasankula, Sanith Rao, and Isaac Gohn

April 28, 2025

Introduction

In recent years, Bayesian methods have become a central tool for modeling complex data structures, thanks to advances in computing power and algorithms. A key element in any Bayesian analysis is the selection of prior distributions, which can significantly impact model behavior and inference. Traditionally, priors have been chosen individually for each parameter, often relying on vague or non-informative choices to minimize subjectivity. However, vague priors can introduce serious problems, including phenomena such as Lindley's paradox and poor model performance when dealing with high-dimensional datasets or limited sample sizes.

This challenge motivates a shift toward more structured prior formulations that control the overall complexity of the model rather than focusing solely on individual parameters. A particularly promising approach involves placing a prior directly on the model's global goodness-of-fit, as measured by the Bayesian coefficient of determination, R^2. Rather than attempting to specify priors separately for each coefficient, defining a prior on R^2 provides a natural and interpretable way to regulate how much variance the model is expected to explain.

In this project, we replicate and expand upon a recent method that proposes placing a beta prior on R^2 to inform the global variance in generalized linear mixed models (GLMMs). This global variance component is then hierarchically decomposed into priors for individual regression coefficients. By controlling the model complexity at a higher level, this framework offers practical advantages in settings involving many parameters, hierarchical structures, or limited data.

Additionally, when direct computation of the induced prior is infeasible, the method uses a generalized beta prime approximation to maintain flexibility and ease of implementation. Through this replication study, we aim to better understand how placing priors at the model level, rather than the parameter level, can lead to improved stability, interpret-ability, and predictive performance in Bayesian modeling.

Setting, Objective, and Methods

General Framework and Objective

The general framework of this project is the study of **generalized linear mixed models (GLMMs)**. In a GLMM, the response variable Y_i depends on both fixed effects $\boldsymbol{\beta}$ and random effects u_i. The model for each observation $i = 1, \ldots, n$ takes the form:

$$g(E[Y_i]) = \eta_i = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \sum_{k=1}^{K} u_{ik}, \tag{1}$$

where $g(\cdot)$ is a link function, $\mathbf{X}_i$ is a vector of covariates, and u_{ik} are random effects associated with different sources of variation.

In traditional Bayesian inference, priors are placed separately on each parameter β_j and u_k. However, vague or poorly chosen priors can lead to severe issues such as Lindley's paradox, poor shrinkage behavior, and distorted posterior distributions, especially in high-dimensional settings.

The objective of this project is to develop and replicate a new approach: **placing a prior directly on the overall model fit**, measured by the Bayesian coefficient of determination R^2. Rather than specifying priors individually, we specify a prior on R^2 and induce priors on the total model variance, which is then hierarchically decomposed among the fixed and random effects. This method aims to provide greater interpretability, flexibility, and robustness in Bayesian modeling.

Description of the Methods

The primary methodological innovation is to specify a prior on R^2 using a **Beta distribution**:

$$R^2 \sim \text{Beta}(a, b) \tag{2}$$

Given R^2, the total variance ϕ_T of the model is induced, and this total variance is partitioned among the individual regression parameters through a Dirichlet distribution:

$$\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p) \sim \text{Dirichlet}(\epsilon_1, \ldots, \epsilon_p) \tag{3}$$

The prior on the model parameters then depends on the total variance ϕ_T and its decomposition $\boldsymbol{\phi}$.

When an analytic form for the induced prior distribution is not available, several approximation strategies are employed:

Linear Approximation Method

When closed-form expressions for R^2 are not available, one approach is to use a linear Taylor series approximation around a baseline β_0.

Suppose we have:

$$\eta_i = g(E[Y_i]) = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \sum_{k=1}^{K} u_{ik} \tag{4}$$

and recall that:

$$R^2 = \frac{\text{Var}(\eta)}{\text{Var}(\eta) + \text{Var}(\epsilon)} \tag{5}$$

The paper proposes approximating $\text{Var}(\eta)$ and $E[\eta]$ using a first-order Taylor expansion around β_0. Specifically, they expand the mean and variance:

- Approximate $E[\eta] \approx \mu(\beta_0)$, - Approximate $\text{Var}(\eta) \approx \mu'(\beta_0)^2 \text{Var}(\beta)$.

Thus, the linear approximation for R^2 becomes:

$$R^2 \approx \frac{W}{W + s^2(\beta_0)} \tag{6}$$

where

$$W = \sigma^2(\beta) = \text{Var}(\eta) \quad \text{and} \quad s^2(\beta_0) = \text{Var}(\epsilon) \quad \text{at baseline} \quad \beta_0.$$

This method simplifies computation but may be inaccurate if the variance changes significantly around β_0.

Generalized Beta Prime (GBP) Approximation Method

When exact methods are not feasible, the paper suggests approximating the distribution of W using a **Generalized Beta Prime (GBP)** distribution.

The GBP prior has the following density:

$$\pi(w; a, b, c, d) \propto \frac{(w^c - 1)^{a-1} \left((1 + w)^{c/d} - 1\right)^{b-1}}{w^{1+c}} \tag{7}$$

for parameters $a, b, c, d > 0$.

The derivation involves: - Fitting a GBP distribution to match the approximate R^2 induced distribution, - Solving for the GBP parameters (a^*, b^*, c^*, d^*) by minimizing the Pearson χ^2-divergence between the true and approximate densities.

The optimization problem is:

$$(a^*, b^*, c^*, d^*) = \operatorname*{argmin}_{a,b,c,d} \int \left(\frac{\pi_{\text{exact}}(w)}{\pi_{\text{GBP}}(w)} - 1 \right)^2 \pi(w)\, dw + \lambda \left\{ (a-c)^2 + (b-c)^2 + (c-1)^2 + (d-1)^2 \right\} \quad (8)$$

where $\lambda > 0$ is a regularization parameter to stabilize fitting.

Important special cases: - When $c = d = 1$, the GBP reduces to a Beta Prime (BP) distribution. - For Poisson regression models and others, specific choices of a and b lead to convenient approximations.

This method is highly flexible and often nearly exact when appropriately tuned.

Analysis Methodology

In an effort to recreate the methods of the authors, we followed their described approach. In order to obtain the R2D2 prior, we followed the authors' description of the model and wrote our own native R functions to recompute the necessary quantities. This included writing our own functions to compute the exact prior distributions when an exact closed-form solution was available, as well as implementing the linear approximation method for the prior distributions and the Generalized Beta Prime (GBP) approximation method. To confirm that our implementation was correct, we recreated Figure 1, Figure 2, and Figure 3 from the original paper and compared our results to the original figures. Once we were confident that our implementation was correct, we proceeded to recreate the results for the two analyses that were presented by the authors in Section 4 of the paper.

The first of these analyzes was the analysis of the gambia dataset in the `geoR` package. The authors used this data set to demonstrate the performance of their method in a real-world scenario. We followed their approach and used the same dataset to recreate the results. Originally, we attempted to write our own native Metropolis-Hastings algorithm to sample from the posterior distribution; however, we quickly realized that this was not feasible due to the complexity of the model. When attempting to implement our own Metropolis-Hastings algorithm, we encountered issues with the convergence of the Markov-Chain, causing our values for W to quickly increase towards infinity. In contrast, this resulted in extremely high R^2 values, which were not consistent with the original paper.

After several attempts to debug our implementation, we decided to use the `rJAGS` package in R, to sample the posterior distribution. This package allowed us to use the JAGS (Just Another Gibbs Sampler) software to perform the sampling, which is a well-established method for Bayesian inference. Since our team was not familiar with JAGS, we decided that it was best to use the code that was implemented by the authors in the original paper, which we found in a Github repository belonging to one of the authors. We were able to use this code to sample from the posterior distribution and obtain the necessary results for the gambia dataset. However, we did decrease the number of iterations for our MCMC algorithm from 105,000 with 5,000 burn-in to a much less computationally intense number of samples, 40,000 iterations with 2,500 burn-in. We then compared our results with the original paper and found that they were consistent with the original results.

After running the analysis on the gambia dataset under the same conditions as the authors, we decided to change a couple of assumptions in the model to see how it would affect our results. We first noticed that, under the authors' implementation, the coordinates of the dataset were scaled before distances were calculated. We decided to remove this scaling step and instead use the original coordinates of the dataset. Additionally, we noticed that the authors had hard-coded the maximum distance between the coordinates to be 0.96 in the code. We decided to remove this hard-coded value and instead use the maximum distance between the non-scaled coordinates. Finally, we noticed that the precision on the fixed effects of our β_j were not specified correctly in the authors' code. JAGS that when defining the dnorm distribution, the precision is specified instead of the variance. In the authors' code, they specified the variance instead of the precision, so we simply took the inverse to ensure the precision was correctly specified. After addressing these three items, we re-ran the analysis on the gambia dataset and recreated the figures presented by the authors. Considering the prior of interest was the R2D2 priors, we did not include the priors that the authors used as the baseline priors in this section.

Finally, we ran the analysis on the second dataset presented by the authors. The second dataset presented by the authors was high-dimensional genomic data from human breast tumors, available in the `mixOmixs` R package. The authors used this dataset to demonstrate the performance of their method in a high-dimensional setting. We followed their approach and used the same dataset to recreate the results. Again, we attempted to create our own Metropolis-Hastings algorithm to sample from the posterior distributions and determine the mean and variance for our β values. However, we encountered extremely high run times due to the setup of the analysis which included running the data 50 times under multiple priors for 10,000 samples. Instead, we decided to attempt to implement the model using the `rSTAN` package in R. This package allowed us to use the STAN software to perform the sampling, which is a well-established method for Bayesian inference. We were able to use this package to sample from the posterior distribution and obtain the necessary results for the breast cancer dataset. However, due to the complexity of the model and our unfamiliarity with the STAN software, we were unable to obtain results for this model after significant trial and error.

Results

Figures 1-2

The primary purpose of recreating Figures 1 and 2 of the paper was to ensure that our function for specifying the R2D2 prior were behaving correctly. The first Figure presented in the paper was the calculated exact prior density of W under a Normal (location and scale) and Poisson model. Since there is no randomness being introduced here, we expect for our results to match the results of the paper exactly, which is consistent for our calculations.

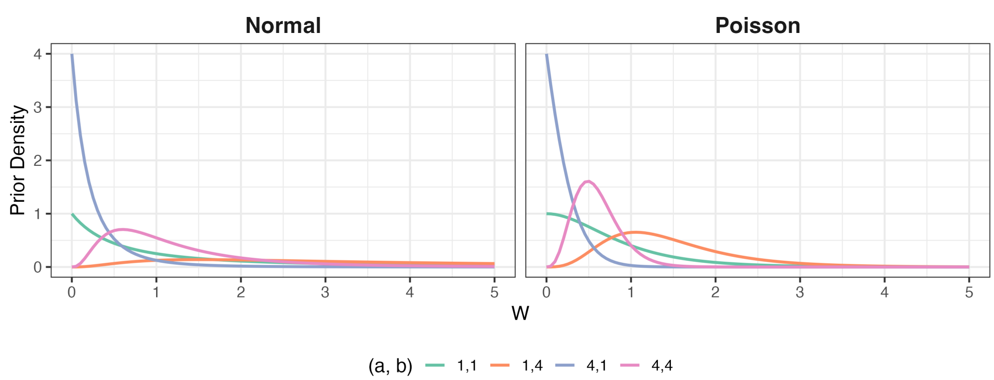

Figure 1: Exact prior density of W for distributions with closed form solutions

Figure 2 of the paper aims to demonstrate the power of the Generalized Beta Prime approximation in mimicking the exact density function of the R2D2 prior on W. In the figure, the density of W is shown for a Poisson regression and the lines show the exact density, the density approximated by the linear approximation presented previously, and the Generalized Beta Prime approximation. Our results are presented in the Figure 2. As expected, and explained by the authors, the Generalized Beta Prime density approximation closely mimic the exact density over the range of a and b values explored. Again, this figure almost exactly matches Figure 2 in the paper.

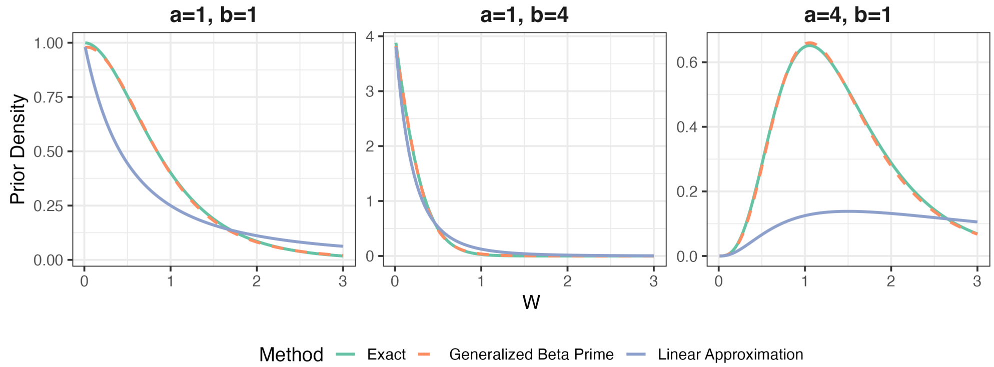

Figure 2: GBP and Linear approximations of prior on W in comparison to the exact density for a Poisson Regression for given induced priors on R^2

Gambia Results

Next, the authors of the paper presented priors for R^2 and global variance W, for the Gambia dataset. The GLMM used to model this data was a logistic regression where the model was specified as:

$$\operatorname{logit}\big(P(Y_i = 1 \mid \eta_i)\big) = \eta_i = \beta_0 + X_i\beta + u_{g_i}.$$

In this model:

$g_i \in \{1, \ldots, L\}$ Index of the village for response i.

u_i Village-level random effect for response i, with

$$E(u_i) = 0, \quad \operatorname{Var}(u_i) = \sigma_u^2.$$

σ_u^2 Marginal variance of the spatial random effects.

$C_{ij} = \operatorname{cor}(u_i, u_j) = e^{-d_{ij}/\rho}$ Exponential spatial correlation between villages i and j.

d_{ij} Euclidean distance between village i and village j.

$\rho > 0$ Spatial range parameter, controlling how rapidly correlation decays with distance.

Additionally, the priors placed on each model parameter are:

$\beta_0 : \beta_0 \sim \mathcal{N}(\mu_0, \tau_0^2)$

$\beta \mid \phi_1, W : \beta \mid \phi_1, W \sim \mathcal{N}\big(0, \frac{1}{5}\,\phi_1\,W\,I_5\big)$

$u \mid \phi_2, W, \rho : u \mid \phi_2, W, \rho \sim \mathcal{N}\big(0, \phi_2\,W\,C\big)$

$\rho : \rho \sim \operatorname{Uniform}\big(0, 2r\big)$

$W : W \sim \operatorname{GBP}\big(a^*, b^*, c^*, d^*\big)$

$\phi : \phi \sim \operatorname{Dirichlet}(\xi_1, \xi_2)$

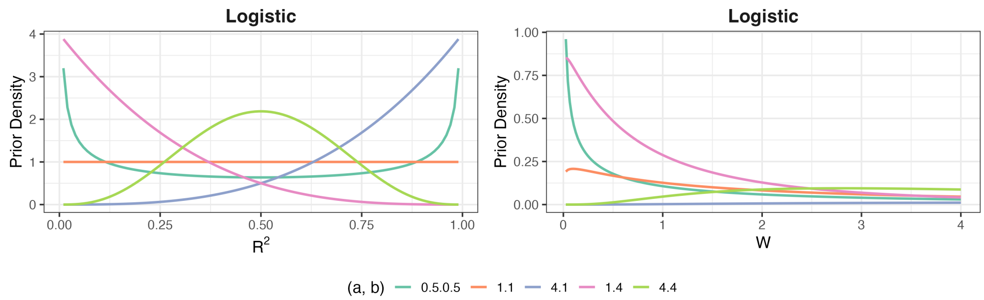

Figure 3: Priors for R^2 and W for the Gambia data over a range of a and b parameters.

For hyperparameter of the priors, these were set to: $\mu_0 = 0$, $\tau_0^2 = 3$, $\xi_1 = \xi_2 = 1$, and r is the maximum Euclidean distance between villages, which was originally hard coded in the authors' code as 0.96. Once, the model was specific, we could determine $\hat{\beta}_0$ by finding the $\bar{Y}$ in our data and performing a logit transformation. This yielded a $\hat{\beta}_0 = -0.59$. In order to induce the R2D2 prior, we first specified our a and b parameters that we would like to induce on R^2 and found the GBP approximation for out parameter based on the value for $\hat{\beta}_0$ and the specific family of GLMM. The parameter results are included in Table 1. In comparison to the parameters that were obtained by the authors, the parameter we obtained closely aligned, any deviations are assumed to be caused by the randomness in the optimization that is being performed to solve for the parameters.

a	b	a^*	b^*	c^*	d^*
1.00	4.00	1.15	2.08	0.91	2.09
0.50	0.50	0.59	0.30	0.89	1.46
1.00	1.00	1.47	0.65	0.79	1.67
4.00	4.00	7.44	2.71	0.73	1.63
4.00	1.00	7.57	0.71	0.69	1.51

Table 1: GBP prior parameters rounded to two decimal places.

Once the parameters for the GBP prior were obtained, we might be interested in understanding and confirming the distributions of the priors on R^2 and W. The resulting distributions from the parameters are shown in Figure 3. The distributions we obtained make sense with what we would expect. The first plot shows the prior distribution of R^2 which is a beta distribution with parameters a and b. Many of the parameters used are common configurations of the beta distribution, so those results are expected. Additionally, if we compare where the peaks exist in the R^2 distribution to where they exist in the prior for W, we would expect to see beta distributions with more mass towards one tail or the other result in W with mass in a similar position. This is observed in both our Figure 3 plot and Figure 3 of the authors, which are very similar.

MCMC sampling was performed to sample the posterior distributions of W, (ϕ_1, ϕ_2), and ρ. The MCMC sampling was performed by the `rJAGS` package, following the code presented by the authors. The only alterations that were made were the number of iterations performed, as described previously. The posteriors distributions for the parameters of interest are shown in Figure 4.

This is the first figure where we start to observe some differences in our results and the results presented by the authors. The primary difference is seen in the distribution for R^2 that is shown to be considerably more symmetric and "normal" in the paper. Alternatively, our densities had a fair bit more left skewness, especially for the lower values of a and b. Additionally, our distribution for R^2 is much wider than that of the Figure 4 in the paper. The bulk of the density in the paper lies between 0.15 and 0.2, where the bulk of the density in our result is between 0.15 and 0.45. Other than that, the remaining densities look overall more similar to the one's proposed in the paper. Even the unique curvature of the ρ density is captured.

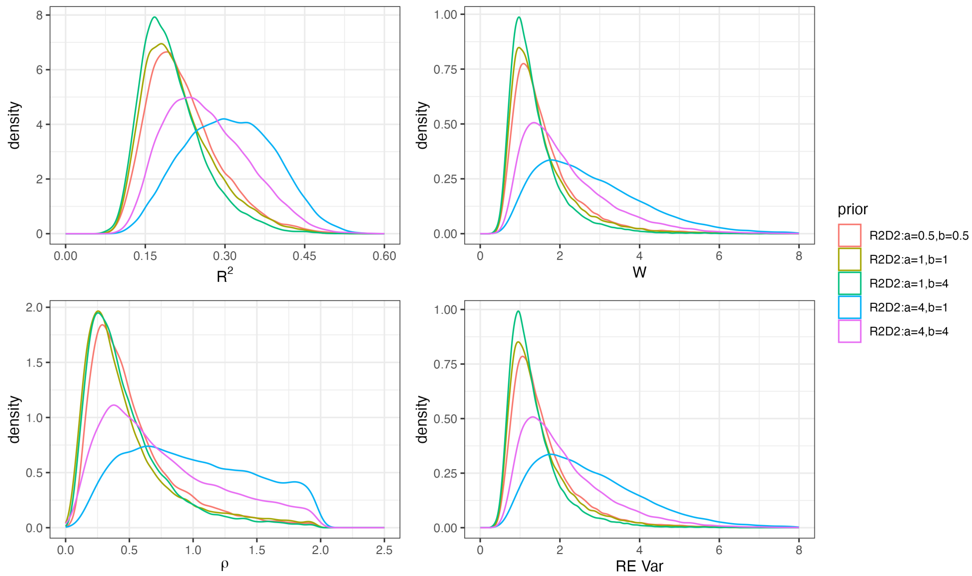

Figure 4: Posteriors for Gambia data under R2D2 priors

The densities that we created were a little less smooth than those in paper. This is expected as we used considerably less iteration in of MCMC algorithm which would cause our samples to not benefit as much from law of large numbers. We also attribute this decreased number of iterations to the distinctions we are seeing in the posterior for R^2. We noticed that the transformation from W to R^2 is very sensitive to changes in W, so we are assuming that with these lower number of iterations, some of our MCMC samples for W had not fully converged and were still a little farther off in the right tail of the distribution than what the authors showed. This is the likely cause of the wider and more skewed density.

Finally, Table 2 shows the overall means and standard deviations observed in the posterior densities for each parameter of interest. Compared to the results of the paper, our results are slightly different, which is expected based on the differing results of our MCMC sampling, as shown in Figure 3. However, the statistics are largely in the same range, so we believe that these difference are also caused by the decreased number of iterations we utilized.

Prior	W_{mean}	W_{sd}	$\text{RE\_Var}_{\text{mean}}$	$\text{RE\_Var}_{\text{sd}}$	R^2_{mean}	R^2_{sd}	ρ_{mean}	ρ_{sd}
R2D2: $a = 0.5, b = 0.5$	1.62	0.88	1.59	0.87	0.22	0.07	0.53	0.35
R2D2: $a = 1, b = 1$	1.51	0.81	1.48	0.81	0.21	0.07	0.49	0.37
R2D2: $a = 1, b = 4$	1.38	0.70	1.35	0.70	0.20	0.06	0.48	0.33
R2D2: $a = 4, b = 1$	2.79	1.38	2.76	1.38	0.31	0.08	1.00	0.49
R2D2: $a = 4, b = 4$	2.15	1.14	2.12	1.14	0.27	0.08	0.75	0.47

Table 2: Posterior summaries of global scale W, random-effect variance, R^2, and spatial range ρ, for Gambia data rounded to two decimal places.

Supplemental Gambia Results

As mentioned previously in the methods, we found three distinct places in code available on Github that we were interested in altering, the first was calculating the distances using the unscaled coordinates rather than the scaled coordinates, the second was the hard coded maximum distance that would no longer be true for the unscaled coordinates, and the final was the small distinction between the variance and precision that was expected by JAGS. After making these alterations to the algorithm, we reran our MCMC under the same conditions specified previously. The results of these updated simulations are shown in Figure 5 and Table 3.

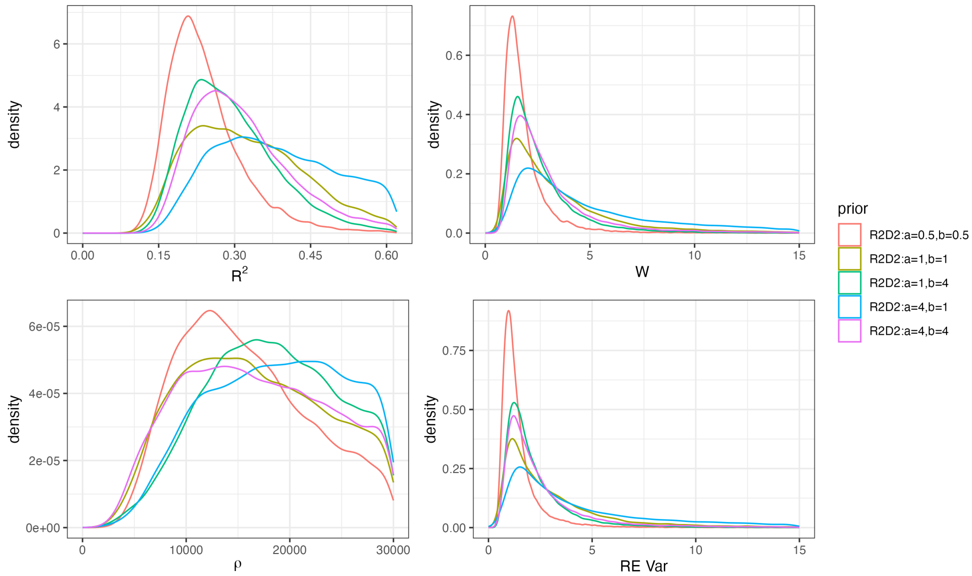

Figure 5: Posteriors for Gambia data under R2D2 priors with Supplemental Alterations

Prior	W_{mean}	W_{sd}	RE_Var$_{\text{mean}}$	RE_Var$_{\text{sd}}$	R^2_{mean}	R^2_{sd}	ρ_{mean}	ρ_{sd}
R2D2: $a = 0.5, b = 0.5$	1.89	1.30	1.51	1.20	0.24	0.08	22759.42	20832.76
R2D2: $a = 1, b = 1$	3.89	4.02	3.46	3.95	0.34	0.12	60080.75	75379.42
R2D2: $a = 1, b = 4$	2.73	2.05	2.32	1.96	0.30	0.09	39076.71	38356.90
R2D2: $a = 4, b = 1$	6.48	6.47	5.82	6.37	0.41	0.14	98689.83	111051.77
R2D2: $a = 4, b = 4$	3.29	3.01	2.79	2.95	0.32	0.10	44068.41	52329.21

Table 3: Posterior summaries of global scale W, random-effect variance, R^2, and spatial range ρ, rounded to two decimal places for Gambia with Supplemental Alterations.

Immediately, you can see that these edits cause some very large differences in the posteriors when compared to the Figure 4 we presented prior and Figure 4 of the paper. While the marginal shapes of R^2, W, and RE Var are mostly the same, the posterior for ρ has been dramatically spread to the right. Additionally, it is worth noting that while the shapes remained mostly unchanged, the X-axis for all of the plots has been extended to include more values, showing the stretching of all of the distribution caused by the aforementioned alterations. This shift is entirely consistent with the fact that, on the unscaled coordinate grid, inter-village distances are numerically much larger; consequently, the MCMC must explore a wider range of values for ρ. This can then bleed into the remaining parameters in the model, causing them to expand as well. These differences are also seen in Table 3. Compared to the results of Table 2. We are seeing all of our values have higher means and significantly higher standard deviations. The significant increase in ρ can be further seen in this table, with ρ values in the tens of thousands. If we think back to the prior on ρ, we specified it to be $U(0, 2r)$. When we scaled the coordinates prior to finding distances, r was set to be around a 0.96, now that we are working in meters, we are seeing much higher values.

Conclusions

The team successfully demonstrated that applying the R2D2 prior within a generalized linear mixed model framework yields stable shrinkage behavior across a variety of settings. Through the Generalized Beta Prime (GBP) approximation, we showed that the prior densities for the global variance parameter W were largely accurately replicated. In the Gambia malaria analysis, the R2D2 prior proved highly effective, highlighting its versatility across different scenarios.

Using `rJAGS`, we recovered posterior distributions for W, the random-effect variance, and the spatial range ρ that were consistent with the original authors' results (see Figure 4, and Table 2). The team also examined the sensitivity of the posterior dispersion to key modeling choices, such as hard-coded distance bounds and other structural assumptions. Additionally, we attempted to implement the method natively using Metropolis–Hastings, but encountered convergence issues that prevented the team from progressing further.

Overall, our replication confirms that the R2D2 prior provides a practical approach to global shrinkage in GLMMs. It offers a flexible yet strong tool for professionals tackling complex inferential problems, particularly in settings with limited data or a large amount of parameters.

References

[1] Yanchenko, E., Bondell, H. D. & Reich, B. J. The R2D2 Prior for Generalized Linear Mixed Models. *The American Statistician* **79**(1), 40–49 (2025). doi: 10.1080/00031305.2024.2352010.