Project Title: Recreating Numerical Results of "The R2D2 Prior for Generalized Linear Mixed Models"

Project Members: Thad Creech , Isaac Gohn , Divija Balasankula , Sanith Rao

The goal of this project is to recreate the numerical results presented in the paper: “The R2D2 Prior for Generalized Linear Mixed Methods.” This paper introduces a novel Bayesian prior designed for Generalized Linear Mixed Models (GLMM). Traditionally, when wanting to fit a GLLM, under Bayesian assumptions, priors are placed on the individual parameters of the model. This requires the modeler to either have domain knowledge helpful in selecting the priors for each parameter, or forces the modeler to use uninformative priors that leave the model prone to overfitting. To overcome these shortcomings, the authors propose that a beta prior is placed on a Bayesian coefficient of determination (R^2), inducing a prior distribution on the global variance of the model. This in turn, imposes priors on the individual parameter and is significantly easier to generalize and control to prevent overfitting.

We will break this goal down into four parts including: implementing methodology, running simulations, application to real-world datasets, and comparing the results of our simulations to the results presented by the authors. We will work collaboratively on each part in order to gain a better understanding of the subject matter. However, we will break down smaller tasks needed to complete each part and assign those to individual members in an effort to finish the project efficiently. Each member will be assigned tasks based on their individual skill sets and interests. For example, if two members of the team have more experience coding in R, then those members will be responsible for more of the coding portion of a task while the other two focus on gaining a better theoretical understanding of the methods in order to ensure that the code is being implemented correctly.

As for data, the methods that were presented in the paper were applied to two different datasets. One was referred to in the paper as “Malaria Data” (Section 4.1) and the other was referred to as “Genomics Data” (Section 4.2). The authors mention in the paper that the dataset referred to in Section 4.1 was the *gambia* dataset that is accessible in R from the *geoR* package. The data referenced in Section 4.2 can similarly be accessed through the *mixOmics* package in R. We have confirmed that both datasets are accessible through their respective packages and plan to use the same data accordingly.

Similarly, the authors mentioned that the code for the approximations proposed in Section 3.2 and the analyses covered in Sections 4.1 and 4.2 is available on one of the author’s Github accounts (https://github.com/eyanchenko/r2d2glmm) and can be accessed as a package itself in R. While we plan to use the author’s code in the event that we get stuck, our goal is to write all of our code from scratch as to get a better understanding of the implementation.

Once we have replicated the result of the paper, we will then analyze these results by comparing our results to those presented in the paper. Specifically, we will recreate the graphs presented in Figures 1, 2, 3, and 4, and Tables 1, 2, and 3. We will highlight the similarities and

differences between the results of our work and the results obtained by the authors, investigating any discrepancies between the two.

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