



## The R2D2 Prior for Generalized Linear Mixed Models

Eric Yanchenko, Howard D. Bondell & Brian J. Reich

To cite this article: Eric Yanchenko, Howard D. Bondell & Brian J. Reich (2025) The R2D2 Prior for Generalized Linear Mixed Models, The American Statistician, 79:1, 40-49, DOI: [10.1080/00031305.2024.2352010](https://doi.org/10.1080/00031305.2024.2352010)

To link to this article: <https://doi.org/10.1080/00031305.2024.2352010>



View supplementary material [↗](#)



Published online: 24 May 2024.



Submit your article to this journal [↗](#)



Article views: 347



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 3 View citing articles [↗](#)



# The R<sup>2</sup>D<sup>2</sup> Prior for Generalized Linear Mixed Models

Eric Yanchenko<sup>a</sup>, Howard D. Bondell<sup>b</sup>, and Brian J. Reich<sup>c</sup>

<sup>a</sup>Akita International University, Akita, Japan; <sup>b</sup>University of Melbourne, Melbourne, Australia; <sup>c</sup>North Carolina State University, Raleigh, NC

## ABSTRACT

In Bayesian analysis, the selection of a prior distribution is typically done by considering each parameter in the model. While this can be convenient, in many scenarios it may be desirable to place a prior on a summary measure of the model instead. In this work, we propose a prior on the model fit, as measured by a Bayesian coefficient of determination ( $R^2$ ), which then induces a prior on the individual parameters. We achieve this by placing a beta prior on  $R^2$  and then deriving the induced prior on the global variance parameter for generalized linear mixed models. We derive closed-form expressions in many scenarios and present several approximation strategies when an analytic form is not possible and/or to allow for easier computation. In these situations, we suggest approximating the prior by using a generalized beta prime distribution and provide a simple default prior construction scheme. This approach is quite flexible and can be easily implemented in standard Bayesian software. Lastly, we demonstrate the performance of the method on simulated and real-world data, where the method particularly shines in high-dimensional settings, as well as modeling random effects.

## ARTICLE HISTORY

Received July 2023  
Accepted April 2024

## KEYWORDS

Bayesian modeling;  
Coefficient of determination;  
Generalized beta prime  
distribution; Goodness-of-fit

## 1. Introduction

As computing power has increased and become more accessible, Bayesian inference has risen to prominence. Researchers are now free to consider complex models with many parameters. An advantage of the Bayesian approach is that it can incorporate prior domain knowledge about some parameters to reduce uncertainty. In the absence of such information, we might select *vague* prior distributions, that is, prior distributions with large variance. This, however, can lead to some unintended consequences such as Lindley's paradox (Lindley 1957). Vague prior distributions can also lead to poor estimates when the number of parameters is large relative to the sample size. This has led to the recent development of shrinkage prior distributions (George and McCulloch 1993; Ročková and George 2018; Park and Casella 2008; Hans 2009; Carvalho, Polson, and Scott 2010; Bhadra et al. 2017; Bhattacharya et al. 2015; Zhang et al. 2022).

Typically, prior distributions are selected for individual parameters based on domain expertise and/or using a general paradigm, for example, shrinkage priors. There are situations, however, where researchers may have meaningful prior information on the model in general as opposed to specific regression coefficients. For example, consider genetic association studies (e.g., Lewis and Knight 2012) where scientists search for the genes that contribute to a specific disease. There may be good understanding of how much genes affect the disease, but little information about which genes are relevant. In this case, it may make more sense to pick a prior for the overall model fit that then induces prior distributions on the parameters. There has

been some previous work toward this end. Hodges and Sargent (2001) use a flat prior distribution on the degrees of freedom in a Gaussian mixed effects model. Simpson et al. (2017) introduce a paradigm that penalizes the complexity of the model as measured by the Kullback-Liebler (KL) divergence between the null and fitted model. This method places a prior on this KL divergence, thus, shrinking the entire model instead of the individual parameters. Fuglstad et al. (2020) present a user-friendly approach to prior construction by using prior beliefs to apportion the overall variance between different random effect components. The authors construct a joint prior distribution which considers the entire model structure. For Gaussian linear regression, Zhang et al. (2022) place a prior on the model fit as measured by the coefficient of determination,  $R^2$ . The authors first derive a Bayesian  $R^2$  and show that the prior  $R^2 \sim \text{Beta}(a, b)$  yields a beta prime prior on the total variance of the regression parameters which is then distributed to each individual parameter through a Dirichlet Decomposition. For sparse high-dimensional regression problems, certain  $R^2$  prior choices and a Dirichlet decomposition give posterior consistency. This method is advantageous because  $R^2$  is an intuitive measure of model fit and it has excellent shrinkage properties.

In this work, we consider a prior on a summary of model fit by proposing a beta prior on Zhang et al.'s (2022) definition of  $R^2$  for generalized linear mixed models. This extends Zhang et al. (2022) beyond linear regression to allow for non-Gaussian responses and random effects. We derive closed-form expressions in multiple scenarios for the prior of the global variance

parameter that induces a beta prior on  $R^2$ . We also present several approximation strategies when an analytic prior distribution is not possible. The main approach we suggest approximates the prior by a *generalized beta prime* (GBP) distribution. This distribution is quite flexible as it can achieve boundedness at the origin as well as a heavy tail (Perez, Pericchi, and Ramirez 2017). The scaled beta prime distribution, a special case of the GBP, has also previously been used as a prior for the variance of the regression coefficients (Klein et al. 2021; Bai and Ghosh 2021). Our method, like Zhang et al. (2022), differs from these previous approaches in that we place a GBP prior on the global variance which is then further decomposed in the hierarchy to the individual regression parameters. Our approach also provides an intuitive way to construct informative prior distributions as well as an automatic approach. The proposed methods can be applied using the `r2d2glmm` package available on GitHub at <https://github.com/eyanchenko/r2d2glmm>.

The remainder of the article proceeds as follows. In Section 2, we describe the generalized linear mixed model framework and present several specific examples. In Section 3, we precisely define a Bayesian  $R^2$  and show how the model-level prior induces prior distributions for the individual model prior parameters. We also present the prior distributions for several specific regression models as well as approximation techniques when a closed-form solution cannot be found. Section 4 applies the proposed method to real-world data and Section 5 concludes with recommendations for default use and next steps.

## 2. Generalized Linear Mixed Models

For notational simplicity, we follow Simpson et al. (2017) and specify our model for a generalized linear mixed model (GLMM), although the ideas presented here can generalize to other settings. For observations  $i \in \{1, \dots, n\}$ , let  $Y_i$  be the response,  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$  be the explanatory variables and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  be the corresponding fixed effects. We standardize the explanatory variables such that each column of  $\mathbf{X}$  has mean zero and variance one. We also assume that there are  $q$  types of random effects,  $\mathbf{u}_k$ ,  $k \in \{1, \dots, q\}$  where  $\mathbf{u}_k = (u_{k1}, \dots, u_{kL_k})^T$  has  $L_k$  levels. We let  $\mathbf{g}_i = (g_{i1}, \dots, g_{iq})^T$  for  $i \in \{1, \dots, n\}$  be membership vectors such that  $g_{ik}$  is the level of random effect  $k$  for observation  $i$  and where mixed-membership is excluded. The fixed and random effects prior distributions are assumed to be independent and are related to the response via the linear predictor

$$\eta_i = \beta_0 + \mathbf{X}_i \boldsymbol{\beta} + \sum_{k=1}^q u_{kg_{ik}} \quad (1)$$

where  $\beta_0$  is the intercept. The responses are assumed to be conditionally independent given the linear predictor and follow density function  $Y_i | \eta_i, \theta \sim f(y | \eta_i, \theta)$ , where  $\theta$  is an additional parameter in the likelihood function (see examples below).

The model for the fixed and random effects is  $\beta_j | \phi_j, W \stackrel{\text{indep}}{\sim} \text{Normal}(0, \phi_j W)$  and  $\mathbf{u}_k | \phi_{p+k}, W \stackrel{\text{indep}}{\sim} \text{Normal}(0, \phi_{p+k} W \mathbf{I}_{L_k})$  where  $W > 0$  controls the overall variance of the linear predictor (not the response) and  $\phi_j \geq 0$  satisfy  $\sum_{j=1}^{p+q} \phi_j = 1$  and

apportion the variance to the different model components. Thus,  $W$  may be interpreted as the total amount of variation in the fixed and random effects, or as a transformation of the total variation of the mean function. In the latter case, the interpretation depends on the link function. Moreover, large values of  $W$  encode a model with greater flexibility since large variance in the mean function means that the model can capture more trends in the data. In the limit as  $W \rightarrow 0$ , conversely, we are reduced to the intercept-only model. This interpretation will be important later in this work when we treat the placement of a large prior mass on  $W$  near zero as “penalizing” toward the null (intercept-only) model. Additionally, notice that the fixed and random effects are modeled similarly, that is, with a random variance. Even so, we maintain their differing interpretations. Specifically, if we are interested in effect estimates themselves, then we treat this effect as “fixed,” but if our interest lies in the underlying population of the effect, then it is treated as “random” (Searle, Casella, and McCulloch 2009). Following this interpretation, we are most interested in the estimates of  $\boldsymbol{\beta}$  and  $\phi_j W$  for  $j = p+1, \dots, p+q$ .

The prior distribution of  $R^2$  relies on the distribution of  $\eta_i$ . For the majority of this work, we assume

$$\eta_i | \beta_0, W \sim \text{Normal}(\beta_0, W). \quad (2)$$

We derive this result in the supplemental materials whether  $\mathbf{X}_i$  is treated as fixed or random. If we treat  $\mathbf{X}_i$  as random, then  $\eta_i$  will be approximately normal for moderate  $p$  by the Central Limit Theorem. On the other hand, if we consider  $\eta_i$  conditional on  $\mathbf{X}_i$ , then the distribution of  $\eta_i$  is exactly normal where the variance is different for each  $i$  but the average variance is  $W$  due to  $\mathbf{X}$ ’s standardization. For either case, we stress that the prior distribution of  $\eta_i$  is independent of the explanatory variables, resulting in a prior that does not depend on  $\mathbf{X}$ , similar to the PC prior (Simpson et al. 2017). Alternative distributions are discussed in Sections 3.2.2–3.2.3 but the normality of  $\eta_i$  is assumed for all experiments.

### 2.1. Variance Decomposition of the Linear Predictor

The variance parameters  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_{p+q})$  determine the relative variance of each component of the model and are restricted to sum to one. These parameters could be fixed, or given prior distributions to add flexibility to the variance decomposition. In the most general case we can assign these parameters a Dirichlet distribution,  $\boldsymbol{\phi} \sim \text{Dirichlet}(\xi_1, \dots, \xi_{p+q})$ . Often times we will take  $\xi_1 = \dots = \xi_{p+q} \equiv \xi_0$ . The concentration parameter  $\xi_0 > 0$  controls the variation of the prior distribution with large  $\xi_0$  encouraging all the variance components to be roughly equal to  $1/(p+q)$ , and small  $\xi_0$  reflecting prior uncertainty in the variance components. In some cases, the effects will be grouped and the variance across groups will be decomposed using a Dirichlet prior, for example, all fixed effects assumed to have the same variance. These ideas are illustrated through examples below.

### 2.2. Examples

To help fix ideas, we present a few specific examples of this prior construction.

**Example 1 (Gaussian linear regression model).** In the linear regression setting with no random effects, the linear predictor is simply

$$\eta_i = \beta_0 + \mathbf{X}_i \boldsymbol{\beta}$$

and we have  $Y_i | \eta_i, \sigma^2 \sim \text{Normal}(\eta_i, \sigma^2)$  so that  $\theta = \sigma^2$  is the error variance. We then take  $\beta_j | \phi_j, W \sim \text{Normal}(0, \phi_j W)$  for  $j = 1, \dots, p$ . Zhang et al. (2022) study the theoretical properties of this approach for various prior distributions on  $W$  and  $\boldsymbol{\phi}$ . In general, this is a global-local shrinkage prior which has been studied in various contexts (e.g., Carvalho, Polson, and Scott 2010; Polson and Scott 2012; Polson et al. 2012; Bhattacharya et al. 2015; Zhang and Bondell 2018).

**Example 2 (Poisson regression with two-way random effects).** For a mixed effects model with two-way (noninteracting) random effects, the linear predictor is

$$\eta_i = \beta_0 + \mathbf{X}_i \boldsymbol{\beta} + u_{1g_{i1}} + u_{2g_{i2}},$$

and  $Y_i | \eta_i \sim \text{Poisson}(\exp(\eta_i))$ . The membership vectors  $g_{i1}$  and  $g_{i2}$  indicate the level assigned to observation  $i$  for random effects type one and two, respectively. The variance weights given to the fixed and random effects are determined by the Dirichlet parameter  $\boldsymbol{\phi}$ . For example, to allow each fixed effect to have a different variance, we might take  $\boldsymbol{\phi} \sim \text{Dirichlet}(\xi_1, \dots, \xi_{p+2})$  where  $\xi_k$  are fixed hyperparameters; on the other hand, for each fixed effect to have the same variance, we might take  $\boldsymbol{\phi} \sim \text{Dirichlet}(\xi_1, \xi_2, \xi_3)$  and then let  $\beta_j | \phi_1, W \sim \text{Normal}(0, \frac{1}{p} \phi_1 W)$  for  $j = 1, \dots, p$  and  $\mathbf{u}_k \sim \text{Normal}(0, \phi_k W \mathbf{I}_{L_k})$  for  $k = 2, 3$ .

**Example 3 (Weibull model).** Survival analysis often uses a Weibull model. For simplicity, we consider uncensored data but this could be extended to censored data. Let there be a single random effect so that the linear predictor is

$$\eta_i = \beta_0 + \mathbf{X}_i \boldsymbol{\beta} + u_{g_i}$$

for membership vector  $g_i \in \{1, \dots, L\}$ . If  $Y_i$  is the survival time, then the model is  $Y_i | \eta_i, \theta \sim \text{Weibull}(e^{\eta_i}, \theta)$  for shape parameter  $\theta$ . If we assume that the fixed effects have equal variance, then  $\boldsymbol{\beta} | \phi_1, W \sim \text{Normal}(0, \frac{1}{p} \phi_1 W \mathbf{I}_p)$  and  $\mathbf{u} | \phi_2, W \sim \text{Normal}(0, \phi_2 W \mathbf{I}_L)$  where  $\boldsymbol{\phi} \sim \text{Dirichlet}(\xi_0, \xi_0)$ .

**Example 4 (Generalized linear regression with spatial random effects).** Consider the scenario where we observe data from  $L$  spatial clusters (e.g., cities or villages) at spatial locations  $\mathbf{s}_1, \dots, \mathbf{s}_L \in \mathcal{R}^2$ . Then let  $Y_i$  be the response from location  $\mathbf{s}_{g_i} \in \mathcal{R}^2$  where  $g_i \in \{1, \dots, L\}$  is the cluster indicator. Spatial generalized linear models account for correlation between observations at nearby locations by adding spatially-correlated random effects (e.g., Diggle, Tawn, and Moyeed 1998). Let  $u_{g_i}$  be the Gaussian random effect for cluster  $g_i$ . The linear predictor is then  $\eta_i = \beta_0 + \mathbf{X}_i \boldsymbol{\beta} + u_{g_i}$ . A stationary and isotropic model assumes  $E(u_i) = 0$  and  $\text{var}(u_i) = \sigma_u^2$  for all  $i$  and  $\text{cor}(u_i, u_j) = C(d_{ij})$ , where  $C$  is a spatial correlation function such as the exponential function  $C(d) = \exp(-d/\rho)$  and  $d_{ij}$  is the distance between locations  $\mathbf{s}_i$  and  $\mathbf{s}_j$ . The covariance structure of the model is determined by the  $L \times L$  correlation matrix  $\mathbf{C}$  with  $(i, j)$  element  $C(d_{ij})$ . The spatial regression model is then in the form of (1) where

$\mathbf{u} | \phi_{p+1}, W, \rho \sim \text{Normal}(0, \phi_{p+1} W \mathbf{C})$  and  $\sigma_u^2 = \phi_{p+1} W$ . While the covariance matrix of the random effect is no longer diagonal, the derivation of (2) still holds as the different random effect levels have the same variance and the covariance terms do not appear in the derivation.

**Example 5 (Generalized additive model).** Nonlinear regression models can also be written as (1). Assume that  $p$  explanatory variables,  $x_{i1}, \dots, x_{ip}$ , are allowed to have a nonlinear relationship with the response variable. The generalized additive model (e.g., Hastie 2017; Klein et al. 2021) is

$$\eta_i = \beta_0 + \sum_{j=1}^p f_j(x_{ij})$$

for unknown functions  $f_1, \dots, f_p$ . A common approach is to model the  $f_j$ 's using a basis expansion

$$f_j(x) = \sum_{l=1}^{L_j} B_{jl}(x) \beta_l^{(k)}$$

where  $B_{j1}, \dots, B_{jL_j}$  are basis function, for example, spline functions and  $\boldsymbol{\beta}^{(k)}$  are “grouped” fixed effects. This model then fits (1) with  $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_p)$  where  $\tilde{\mathbf{X}}_j \in \mathcal{R}^{n \times L_j}$  is such that  $(\tilde{\mathbf{X}}_j)_{ik} = B_{jk}(x_{ij})$ , and  $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)T}, \dots, \boldsymbol{\beta}^{(p)T})^T$ . Then  $\beta_k^{(j)} \sim \text{Normal}(0, \frac{1}{L_j} \phi_j W)$  for  $j \in \{1, \dots, p\}$  and  $k \in \{1, \dots, L_j\}$  such that  $\phi_j$  determines the proportion of the variance allocated to the nonlinear effect of  $x_{ij}$ .

### 3. Variance Decomposition $R^2$ and the R2D2 Prior

Gelman et al. (2019), Gelman and Hill (2006), and Zhang et al. (2022) propose measures of model complexity that we name the *Variance Decomposition  $R^2$*  (VaDeR). For the GLMM in Section 2, define  $E(Y_i | \eta_i) = \mu(\eta_i)$  and  $\text{Var}(Y_i | \eta_i) = \sigma^2(\eta_i)$  which relates the linear predictor to the response distribution. Gelman et al. (2019) use the empirical definition of  $R^2$ , defined as

$$R_n^2 = \frac{V\{\mu(\eta_1), \dots, \mu(\eta_n) | \mathbf{X}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}\}}{V\{\mu(\eta_1), \dots, \mu(\eta_n) | \mathbf{X}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}\} + M\{\sigma^2(\eta_1), \dots, \sigma^2(\eta_n) | \mathbf{X}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}\}} \quad (3)$$

where  $M$  and  $V$  are the sample mean and variance operators, respectively.

In (3),  $V\{\mu(\eta_1), \dots, \mu(\eta_n) | \mathbf{X}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}\}$  is the variance of the expectation of future data and  $M\{\sigma^2(\eta_1), \dots, \sigma^2(\eta_n) | \mathbf{X}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{u}\}$  is the expected variance of future residuals, both conditioned on the explanatory variables, membership vectors and fixed and random effects. Because of this conditioning, Gelman et al. (2019) propose  $R_n^2$  as an a posteriori measure of model fit. In principle, however, if the values of  $\mathbf{X}_i$  and  $\mathbf{g}_i$  are known but we had yet to observe the responses  $Y_i$ , then the prior distributions of the fixed and random effects would induce a prior distribution on  $R_n^2$ . Then  $R_n^2$  is the proportion of variance explained by the model for future data, conditioned on these variables and our prior information for  $\boldsymbol{\beta}$  and  $\mathbf{u}_k$ .



While  $R_n^2$  is an intuitive measure of the fit of the model to a particular dataset, for the purpose of setting prior distributions we follow Zhang et al. (2022). We measure complexity at the population level and use the marginal version of  $R^2$  that averages over variation in both the explanatory variables and random effect levels ( $\mathbf{X}$  and  $\mathbf{g}$ ) as well as parameters ( $\boldsymbol{\beta}$  and  $\mathbf{u}_k$ ). The marginal distribution does not depend on  $\mathbf{X}_i$  or  $\mathbf{g}_i$  so the observations are exchangeable. We can then drop the subscript distinguishing them and consider the model for an arbitrary observation  $Y$  with  $E(Y|\eta) = \mu(\eta)$ ,  $\text{var}(Y|\eta) = \sigma^2(\eta)$  and  $\eta|\beta_0, W \sim \text{Normal}(\beta_0, W)$  as in (2). Then  $R^2$  becomes

$$R^2(\beta_0, W) = \frac{\text{var}\{\mu(\eta)\}}{\text{var}(Y)} = \frac{\text{var}\{\mu(\eta)\}}{\text{var}\{\mu(\eta)\} + E\{\sigma^2(\eta)\}} \quad (4)$$

where  $E\{\sigma^2(\eta)\}$  and  $\text{var}\{\mu(\eta)\}$  are summaries of the distribution of  $\eta$  and thus depend on parameters  $\beta_0$  and  $W$ . For the sake of simplicity, we suppress the dependence on  $(\beta_0, W)$  and write  $R^2(\beta_0, W) = R^2$  for the remainder of the article. The supplemental materials discusses the relationship between  $R_n^2$  and  $R^2$  and shows that under general conditions,  $R_n^2$  will converge to  $R^2$  when both the sample size and number of effective parameters increase. We also include a brief discussion comparing  $R^2$  and  $R_n^2$  with other measures of model fit for GLMMs (e.g., McFadden 1973; Cox and Snell 1989). Moreover, we note that the coefficient of determination is not commonly used for GLMMs as a measure of model fit. This could be for several reasons, not least of which being that it is difficult to define and interpret a principled  $R^2$  for logistic regression, poisson regression, etc. A major advantage of  $R^2$  and  $R_n^2$  is that they can easily be extended to GLMMs, while also having intuitive interpretations.

As denoted in (4), the prior distribution of  $R^2$  is determined by the joint prior  $(\beta_0, W)$ . For Gaussian responses the distribution of  $R^2$  is invariant to  $\beta_0$ , and so to reduce the problem to matching univariate distributions, we parameterize the prior for  $(\beta_0, W)$  as the conditional prior for  $W|\beta_0$  and marginal prior for  $\beta_0 \sim \pi_0$ . We then select a prior for  $W|\beta_0$  so that  $R^2 \sim \text{Beta}(a, b)$ . By construction, since  $R^2 \sim \text{Beta}(a, b)$  conditioned on any  $\beta_0$ ,  $R^2$  also follows a  $\text{Beta}(a, b)$  marginally over the joint prior for  $(\beta_0, W)$  for any marginal prior  $\pi_0$ . Combined with the Dirichlet prior distribution on the variance proportions, this defines the  $R^2$  Dirichlet decomposition prior (R2D2).

The  $\text{Beta}(a, b)$  prior for  $R^2$  is our default choice, but in some cases the support of  $R^2$  can be restricted to a subspace of  $[0, 1]$  and a modification is required. Typically, when  $W = 0$  we also have  $\text{var}\{\mu(\eta)\} = 0$  and thus  $R^2 = 0$  assuming the distribution of  $Y|\eta$  is not degenerate, that is,  $\sigma^2(\eta) > 0$ . If, however,  $\text{var}\{\mu(\eta)\} > 0$  when  $W = 0$ , then the lower bound of  $R^2$ ,  $R_{\min}^2$ , is strictly greater than zero (e.g., Poisson regression with offsets in supplemental material). Conversely, for some link functions,  $R^2 < 1$  for all  $W$  (e.g., the zero-inflated Poisson model in supplemental materials). In general, the upper bound of  $R^2$ ,  $R_{\max}^2$ , is 1 if and only if  $E\{\sigma^2(\eta)\} = o(\text{var}\{\mu(\eta)\})$  as  $W \rightarrow \infty$ . In cases where  $R_{\min}^2 > 0$  and/or  $R_{\max}^2 < 1$ , we use a  $\text{Beta}(a, b)$  prior distribution for the shifted and scaled  $R^2$ , denoted  $\tilde{R}^2 = (R^2 - R_{\min}^2)/(R_{\max}^2 - R_{\min}^2)$ . This is equivalent to using a four-parameter beta distribution for the prior where  $R^2 \sim \text{Beta}(a, b, R_{\min}^2, R_{\max}^2)$  has density function

$$\pi(r^2) = \frac{(r^2 - R_{\min}^2)^{a-1} (R_{\max}^2 - r^2)^{b-1}}{(R_{\max}^2 - R_{\min}^2)^{a+b-1} B(a, b)}, \quad R_{\min}^2 \leq r^2 \leq R_{\max}^2.$$

In most cases,  $R_{\min}^2 = 0$  and  $R_{\max}^2 = 1$  so unless otherwise noted we simply denote the prior as  $R^2 \sim \text{Beta}(a, b)$ .

### 3.1. Special Cases with Exact Expressions

Below we derive the expressions for the prior distribution for  $W$  in several special cases where the exact prior distribution is available.

**Location-scale models.** The location-scale model is  $Y_i = \eta_i + \sigma\epsilon_i$ , where the errors  $\epsilon_i$  have mean zero and variance one. Then  $\mu(\eta) = \eta$  and  $\sigma^2(\eta) = \sigma^2$  and thus  $R^2 = W/(W + \sigma^2)$ . Assuming  $R^2$  follows a  $\text{Beta}(a, b)$  and  $\sigma = 1$  (or more generally that  $\sigma^2$  appears in the prior variance,  $\beta_j|\sigma^2, \phi_j, W \sim \text{Normal}(0, \sigma^2\phi_j W)$ ), Zhang et al. (2022) show that the induced prior on  $W$  is a Beta Prime distribution, denoted  $W \sim \text{BP}(a, b)$  with density function

$$\pi(w) = \frac{1}{B(a, b)} \frac{w^{a-1}}{(1+w)^{a+b}}, \quad w \geq 0, \quad (5)$$

where  $B(\cdot, \cdot)$  denotes the Beta function. In the left panel of Figure 1, we plot  $\pi(w)$  for various values of  $a$  and  $b$ , and we can see that the BP prior distribution for  $W$  has heavier tails when the expected  $R^2$  is large ( $a > b$ ) versus small ( $a < b$ ).

For  $\sigma^2 \neq 1$ , and not included in the prior variance, that is,  $\beta_j|\phi_j, W \sim \text{Normal}(0, \phi_j W)$ , the induced prior distribution for  $W$  is a Generalized Beta Prime (GBP) distribution,  $W|\sigma^2 \sim \text{GBP}(a, b, 1, \sigma^2)$ . The GBP distribution can be obtained via a transformation of a BP random variable, that is, if  $V \sim \text{BP}(a, b)$  then  $W = dV^{1/c} \sim \text{GBP}(a, b, c, d)$  and has density function

$$\pi(w; a, b, c, d) = \frac{c \left(\frac{w}{d}\right)^{ac-1} \left(1 + \left(\frac{w}{d}\right)^c\right)^{-a-b}}{dB(a, b)}, \quad w \geq 0 \quad (6)$$

for  $a, b, c, d > 0$ . The GBP reduces to the BP if  $c = d = 1$ .

We note a few properties of the GBP distribution. The behavior at the origin is controlled by the value of  $ac$ , with

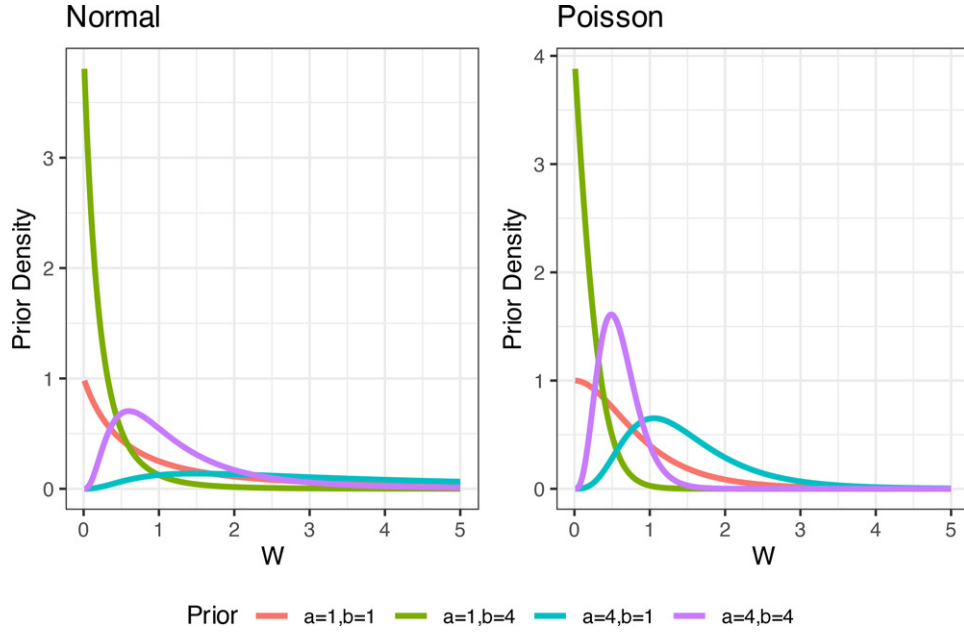
$$\lim_{w \rightarrow 0} \pi(w; a, b, c, d) = \begin{cases} \infty & ac < 1 \\ \frac{c}{B(a, b)d} & ac = 1 \\ 0 & ac > 1 \end{cases}.$$

The tail behavior is controlled by  $bc$  with valid mean if only if  $bc > 1$ . Also, for any model with  $W \sim \text{GBP}(a, b, c, d)$  for the overall variance, then the standard deviation has prior distribution  $W^{1/2} \sim \text{GBP}(a, b, 2c, d^{1/2})$ . As another special case of the GBP, if  $a = 1/2$ ,  $b = \nu/2$ ,  $c = 2$  and  $d = \sqrt{\nu\sigma^2}$ , then  $W$  is distributed as a half- $t$  distribution with  $\nu$  degrees of freedom and scale  $\sigma^2$ . Specifically, if  $W \sim \text{GBP}(\frac{1}{2}, \frac{1}{2}, 1, \sigma^2)$ , then  $\sqrt{W}$  follows a half-Cauchy distribution with scale  $\sigma$  as in Gelman (2006).

**Poisson regression.** The Poisson regression model is  $Y|\eta \sim \text{Poisson}(e^\eta)$  and thus  $\mu(\eta) = \sigma^2(\eta) = e^\eta$ . Since  $\eta|\beta_0, W \sim \text{Normal}(\beta_0, W)$ ,  $e^\eta|\beta_0, W \sim \text{LogNormal}(\beta_0, W)$ , and thus

$$R^2 = \frac{e^W - 1}{e^W - 1 + e^{-\beta_0 - \frac{1}{2}W}}. \quad (7)$$

$R^2 \sim \text{Beta}(a, b)$  induces (see supplemental materials) the prior for  $W$  with density



**Figure 1.** Plot of the prior distribution of  $W$  to induce  $R^2 \sim \text{Beta}(a, b)$  with  $\beta_0 = 0$ . The title of each panel corresponds to the response model. The normal case takes  $\sigma^2 = 1$ .

$$\pi(w|\beta_0; a, b) = \frac{1}{B(a, b)} \frac{(e^w - 1)^{a-1} e^{-b(\beta_0 + w/2)} (3e^w - 1)}{2(e^w - 1 + e^{-\beta_0 - w/2})^{a+b}}, \quad w \geq 0. \quad (8)$$

We plot this distribution in Figure 1 and show that shape of the prior looks very similar to that of the location-scale case. We do note, however, that the prior for  $W$  has exponential-decaying tails on the scale of  $E(Y|\eta) = e^\eta$  as seen in (8). But, on the scale of  $\log\{E(Y|\eta)\} = \eta$ , which is the same scale as  $\beta$  and  $\mathbf{u}$ , the prior has polynomial-decaying tails. The value of the prior at 0 is  $\infty$  if  $a < 1$ ,  $be^{\beta_0}$  if  $a = 1$  and 0 if  $a > 1$ .

In the supplemental materials, we also include the exact prior distributions for: Poisson with offsets, negative binomial, zero-inflated Poisson, and the Weibull model.

### 3.2. Approximate Methods

In some cases, for example, logistic regression, a closed-form expression for VaDeR is not available, so in this section we discuss alternatives.

#### 3.2.1. Linear Approximation

The two components we must compute for VaDeR are  $\text{var}\{\mu(\eta)\}$  and  $E\{\sigma^2(\eta)\}$ . The simplest approach to approximate these is with a linear approximation. Applying a first-order Taylor series approximation of  $\mu(\eta)$  and  $\sigma^2(\eta)$  around  $\beta_0$  gives

$$\text{var}\{\mu(\eta)\} \approx \{\mu'(\beta_0)\}^2 W \quad \text{and} \quad E\{\sigma^2(\eta)\} \approx \sigma^2(\beta_0). \quad (9)$$

Then denoting  $s^2(\beta_0) = \sigma^2(\beta_0)/\{\mu'(\beta_0)\}^2$  we have

$$R^2 \approx \frac{W}{W + s^2(\beta_0)}. \quad (10)$$

If  $R^2 \sim \text{Beta}(a, b)$ , the resulting prior for  $W$  is  $W|\beta_0 \sim \text{GBP}(a, b, 1, s^2(\beta_0))$ . This result does not require any distributional assumptions about  $\eta_i$  other than a finite mean and variance after transformation by  $\mu(\cdot)$  and  $\sigma^2(\cdot)$ . Additionally, the computational burden is essentially zero.

#### 3.2.2. Quasi-Monte Carlo (QMC)

As we will show, in many cases the linear approximation does a poor job approximating the true  $R^2$  distribution. Therefore, we must turn to other methods. Since finding  $R^2$  reduces to computing complicated integrals, we can use integral approximation techniques, like quasi-Monte Carlo (QMC; e.g., Morokoff and Caflisch 1995). In usual Monte Carlo integration, the integral of interest is approximated by summing over a randomly generated sample of points. QMC is similar except that the points are selected deterministically. To construct the R2D2 prior, we approximate

$$E\{\mu(\eta)^m\} \approx \tilde{\mu}_m(W|\beta_0) = \frac{1}{K-1} \sum_{i=1}^{K-1} \mu(\beta_0 + z_i \sqrt{W})^m \quad (11)$$

and

$$E\{\sigma^2(\eta)\} \approx \tilde{\sigma}^2(W|\beta_0) = \frac{1}{K-1} \sum_{i=1}^{K-1} \sigma^2(\beta_0 + z_i \sqrt{W}) \quad (12)$$

where  $z_i$  is the  $i/K$  quantile of a standard normal distribution and  $m = 1, 2$ . This gives an approximation of  $R^2$  for a given  $\beta_0$  and  $W$ , which we denote by

$$\tilde{R}^2(W|\beta_0) \approx \frac{\tilde{\mu}_2(W|\beta_0) - \tilde{\mu}_1^2(W|\beta_0)}{\tilde{\mu}_2(W|\beta_0) - \tilde{\mu}_1^2(W|\beta_0) + \tilde{\sigma}^2(W|\beta_0)}. \quad (13)$$

Assuming  $R^2 \sim \text{Beta}(a, b)$ , then the prior for  $W$  is

$$\begin{aligned} \pi(w|\beta_0; a, b) &= \frac{1}{B(a, b)} \{\tilde{R}^2(w|\beta_0)\}^{a-1} \{1 - \tilde{R}^2(w|\beta_0)\}^{b-1} \left| \frac{d\tilde{R}^2(w|\beta_0)}{dw} \right|, \quad w \geq 0. \end{aligned} \quad (14)$$

Since this cannot be represented with elementary operations, in practice, we take a numerical derivative to evaluate the prior at a given value, and the entire calculation is completed virtually immediately.

The results in (11) and (12) make use of the normality of  $\eta_i$  from (2). The QMC procedure can be modified to account for nonnormal  $\eta_i$ . Let  $\eta \sim F(\eta|\beta_0, W)$  for distribution function  $F(\eta|\beta_0, W)$ . Then we approximate

$$E\{\mu(\eta)^m\} \approx \tilde{\mu}_m(W|\beta_0) = \frac{1}{K-1} \sum_{i=1}^{K-1} \mu\{q_i(\beta_0, W)\}^m$$

where  $q_i(\beta_0, W)$  is the  $i/K$  quantile of  $F(\eta|\beta_0, W)$ . A similar result holds for approximating  $E\{\sigma^2(\eta)\}$  which then leads to an analogous result to (13). In practice,  $F(\eta|\beta_0, W)$  can be derived analytically if the distribution of  $\mathbf{X}_i$  is known. A more general strategy is to average over the empirical distribution of  $\mathbf{X}$  giving a mixture of normal distributions for  $F(\eta|\beta_0, W)$ .

### 3.2.3. Generalized Beta Prime Approximation

The GBP distribution provides an exact solution for the location-scale model in Section 3.1, and an approximate solution for the linear approximation in Section 3.2.1. The prior  $W \sim \text{GBP}(a, b, c, d)$  also induces the exact  $R^2 \sim \text{Beta}(a, b)$  prior distribution for any model with link functions  $\text{var}\{\mu(\eta)\} = W^c$  and  $E\{\sigma^2(\eta)\} = d^c$ . The GBP will not give an exact solution in all cases, but it is a flexible four-parameter model which may often provide a reasonable approximation. Therefore, a general approximation strategy is to find the values of  $(a^*, b^*, c^*, d^*)$  so that the prior  $W \sim \text{GBP}(a^*, b^*, c^*, d^*)$  gives an approximate  $\text{Beta}(a, b)$  distribution for  $R^2$ .

The optimal values of  $(a^*, b^*, c^*, d^*)$  depend on  $\mu(\cdot)$  and  $\sigma^2(\cdot)$  as well as  $\beta_0$ ,  $a$ , and  $b$ . For given link functions and parameters, let  $W \sim \pi(w)$  be the distribution that gives exactly  $R^2 \sim \text{Beta}(a, b)$ . The GBP parameters are then set to minimize the Pearson  $\chi^2$ -divergence (Rényi 1961) between the true and approximated PDFs since this metric enforces a close fit at both the origin and in the tails. We found that minimizing this quantity alone, however, led to unstable solutions, that is, the surface being maximized over is “flat.” This means that vastly different values of  $(a^*, b^*, c^*, d^*)$  may lead to GBP distributions that yield roughly the same approximation of  $\pi(w)$ . Thus, we also add a regularization term to shrink the prior toward a  $\text{GBP}(a, b, 1, 1)$  distribution. We regularize toward this distribution because it gives the exact solution in the location-scale case and can be considered the baseline distribution. This results in the following optimization problem:

$$(a^*, b^*, c^*, d^*) = \underset{\alpha, \beta, c, d}{\text{argmin}} \int_0^\infty \left\{ \frac{f_{\text{GBP}(w; \alpha, \beta, c, d)} - \pi(w)}{\pi(w)} \right\}^2 \pi(w) dw + \lambda\{(\alpha - a)^2 + (\beta - b)^2 + (c - 1)^2 + (d - 1)^2\}, \quad (15)$$

where  $\lambda > 0$  is a tuning parameter. A larger value of  $\lambda$  yields a more stable solution but with a worse fit whereas a smaller value of  $\lambda$  yields a better fit but with more instability. We found that  $\lambda = \frac{1}{4}$  gives a good balance between fit and stability. In practice, the integral is approximated by a sum and  $\pi(w)$  is approximated using QMC as in Section 3.2.2, if necessary. Since the GBP approximation may depend on the QMC procedure which can be modified to allow for non-normality in  $\eta$ , the GBP approach can similarly be adapted to allow for any distribution of  $\eta$ .

A major advantage of the GBP prior is that it can be easily implemented in standard Bayesian software such as JAGS (Plummer et al. 2016) or Stan (Carpenter et al. 2017). Because the exact prior distributions found in Section 3.1, as well as the resulting distributions from the QMC procedure in Section 3.2.2 do not easily allow for Gibbs sampling, these priors would be difficult for a practitioner to implement. By finding the GBP approximation, however, the R2D2 prior can be easily coded in JAGS and Stan. Example code is available as a vignette in our R package on GitHub. To specify the prior in these packages, we use the relationship that if  $R^2 \sim \text{Beta}(a, b)$  and  $W = d\{R^2/(1 - R^2)\}^{1/c}$ , then  $W \sim \text{GBP}(a, b, c, d)$ . Because of these features, we recommend this method for general use even in cases when the exact expression is available, and will be the method we consider in the simulation studies.

Since the GBP approximation depends on  $\beta_0$  (and  $\theta$ ), this approximation should be updated with the unknown parameter  $\beta_0$ . The calculation in (15) takes about half a second, however, so this would be overly time prohibitive to compute during every MCMC iteration. Instead, we find the GBP approximation once at the beginning of the analysis at  $\hat{\beta}_0 = g(\sum_{i=1}^n Y_i/n)$  for link function  $g(\cdot)$ . Thus, to induce  $R^2 \sim \text{Beta}(a, b)$ , the first step is to find  $(a^*, b^*, c^*, d^*)$  as in (15) at  $\hat{\beta}_0$  (and  $\hat{\theta}_{\text{MLE}}$ , if necessary, the maximum likelihood estimate of the dispersion parameter). After determining  $(a^*, b^*, c^*, d^*)$ ,  $\beta_0$  (and  $\theta$ ) are treated as unknown parameters in the subsequent Bayesian analysis. In the supplemental materials, we report the GBP approximations for various values of  $(a, b)$  and different models. In most cases, the best fitting  $a^*$  and  $b^*$  values are not close to  $(a, b)$  which demonstrates the need for this approximation.

Figure 2 compares the linear and GBP approximations with the true distribution for the Poisson model (Section 3.1). The GBP is nearly a perfect match to the true distribution for each prior. The linear approximation is reasonable when  $a = 1, b = 4$ , but very poor when  $a = 4, b = 1$ . This example shows that the GBP is a very good approximation to the true prior distribution of  $W$ .

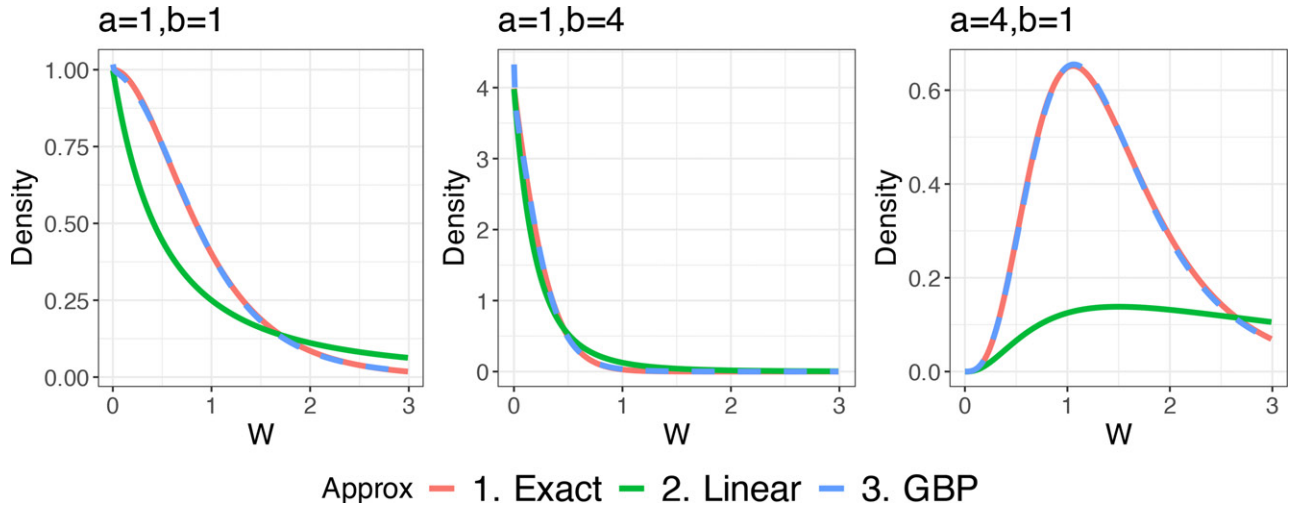
In the supplemental materials, we conduct a simulation study to compare the R2D2 prior with a standard vague prior, the Horseshoe prior (Carvalho, Polson, and Scott 2009), and the PC prior (Simpson et al. 2017) while also comparing different combinations of  $(a, b)$ . The proposed method performs favorably across all settings and particularly well in the high-dimensional regression setting. Indeed, we empirically find that prior distributions with large prior mass near  $R^2 = 0$  yield good shrinkage properties.

## 4. Real Data Analysis

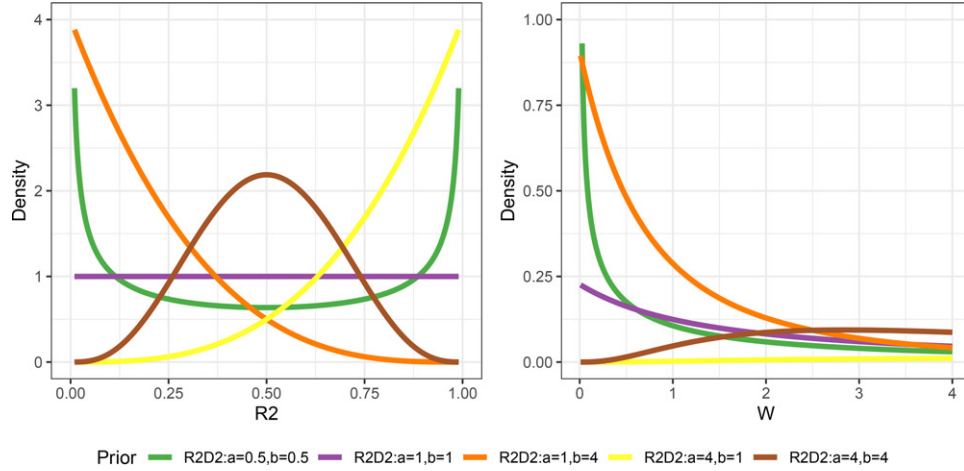
The proposed method is now applied to two real-world datasets. The first analysis is focused on inference, while the second looks at prediction in the high-dimensional setting.

### 4.1. Malaria Data

We now analyze the gambia dataset (Thomson et al. 1999) from the `geOR` package (Ribeiro Jr et al. 2007) in R to demonstrate the use of the R2D2 prior in practice. We also consider



**Figure 2.** Comparison of different approximation methods for the Poisson regression model with  $\beta_0 = 0$ . The title of each pane indicates the induced prior distribution on  $R^2$ .



**Figure 3.** Prior  $R^2$  and global variance parameter for R2D2 prior for Gambia data.

PC and vague prior distributions. There are  $n = 2035$  children in this dataset with binary response variable  $Y_i$  which equals 1 if child  $i$  tested positive for malaria and 0 otherwise. There are  $p = 5$  explanatory variables including age, indicator of using a bed net, indicator of whether the bed net is treated, “greenness” of village and indicator of a health center in the area. These variables are standardized to have mean zero and variance one. There are also the  $L = 65$  villages where each child lived, along with the spatial location of each village.

We model the village effect as a spatial random effect. As in [Example 4](#) from [Section 2.2](#), the linear predictor is

$$\text{logit}\{P(Y_i = 1|\eta_i)\} = \eta_i = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + u_{g_i} \quad (16)$$

where  $g_i \in \{1, \dots, L\}$  is the village of response  $i$ . We also assume that  $E(u_i) = 0$  and  $\text{var}(u_i) = \sigma_u^2$  for all  $i$  and exponential spatial correlation  $C_{ij} = \text{cor}(u_i, u_j) = e^{-d_{ij}/\rho}$  where  $d_{ij}$  is the distance between village  $i$  and  $j$  and  $\rho > 0$  is the spatial range parameter. Then the full prior specification for R2D2 is

$$\begin{aligned} \beta_0 &\sim \text{Normal}(\mu_0, \tau_0^2), \boldsymbol{\beta} | \phi_1, W \sim \text{Normal}(0, \frac{1}{5}\phi_1 W \mathbf{I}_5), \\ \mathbf{u} | \phi_2, W, \rho &\sim \text{Normal}(0, \phi_2 W \mathbf{C}), \end{aligned}$$

**Table 1.** Generalized Beta Prime approximation parameters for Gambia data with  $\hat{\beta}_0 = -0.59$ .

$a$	$b$	$a^*$	$b^*$	$c^*$	$d^*$
1	4	1.15	2.08	0.91	2.09
0.5	0.5	0.57	0.29	0.90	1.54
1	1	1.47	0.65	0.79	1.67
4	4	7.45	2.72	0.73	1.63
4	1	7.77	0.71	0.68	1.45

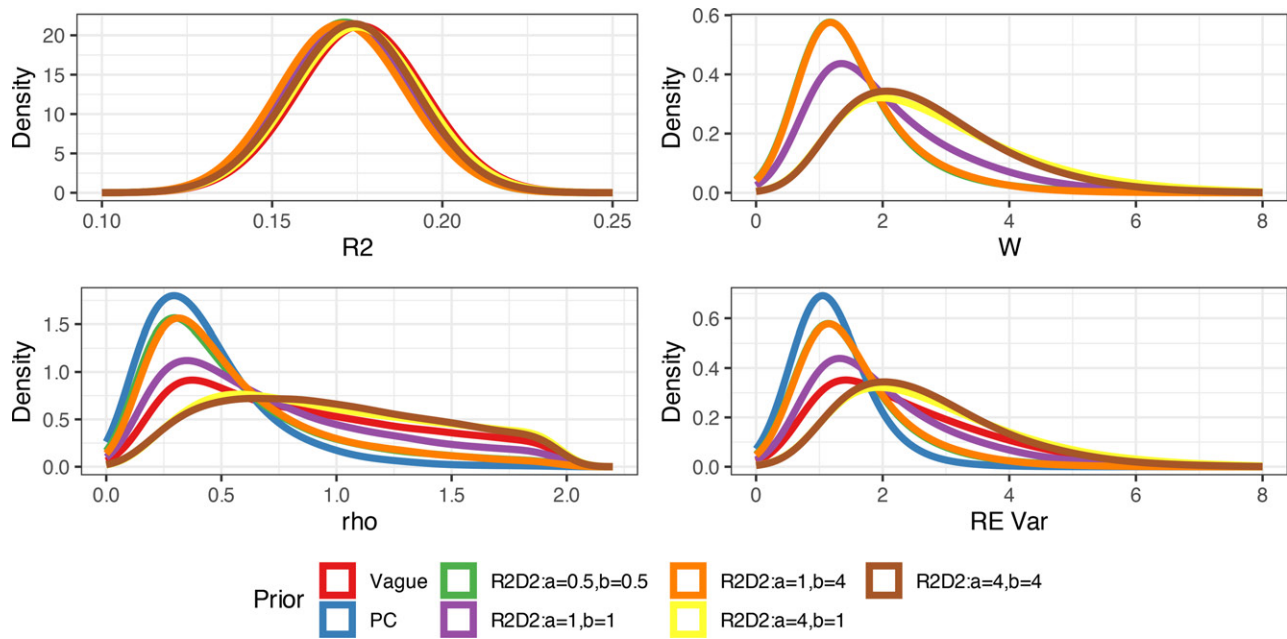
$$\begin{aligned} \rho &\sim \text{Uniform}(0, 2r), W \sim \text{GBP}(a^*, b^*, c^*, d^*), \\ \boldsymbol{\phi} &\sim \text{Dirichlet}(\xi_1, \xi_2) \end{aligned} \quad (17)$$

for hyper-parameters set to  $\mu_0 = 0$ ,  $\tau_0^2 = 3$ ,  $\xi_1 = \xi_2 = 1$  and  $r$  is the maximum distance between pairs of villages. Note that  $\sigma_u^2 = \phi_2 W$  in this model. We find  $\hat{\beta}_0 = -0.59$  and  $(a^*, b^*, c^*, d^*)$  are in [Table 1](#) and the resulting prior distributions are plotted in [Figure 3](#).

For PC prior, the full prior specification is

$$\begin{aligned} \beta_0 &\sim \text{Normal}(\mu_0, \tau_0^2), \boldsymbol{\beta} \sim \text{Normal}(0, \tau_1^2 \mathbf{I}_5), \\ \mathbf{u} | \sigma_u^2 &\sim \text{Normal}(0, \sigma_u^2 \mathbf{C}), \\ \rho &\sim \text{Uniform}(0, 2r), \sigma_u \sim \text{Exp}(\lambda_0). \end{aligned} \quad (18)$$





**Figure 4.** Posterior  $R^2$ , global variance and random effect variance and  $\rho$  for vague (uninformative) prior distributions, PC and R2D2 for Gambia data with village spatial random effect.

**Table 2.** Posterior mean and standard deviation for  $R^2$ , global variance ( $W$ ), random effect variance ( $\sigma_u^2$ ) and spatial range ( $\rho$ ) for each method for Gambia data considering spatial random effect.

Method	$R^2$		$W$		$\sigma_u^2$		$\rho$	
	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev
Vague	0.176	0.016	—	—	2.389	1.397	0.851	0.500
PC	0.173	0.016	—	—	1.210	0.515	0.427	0.270
$R^2 \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$	0.171	0.015	1.538	0.827	1.511	0.824	0.549	0.388
$R^2 \sim \text{Beta}(1, 1)$	0.173	0.016	2.023	1.148	1.992	1.143	0.722	0.465
$R^2 \sim \text{Beta}(1, 4)$	0.171	0.016	1.540	0.796	1.513	0.792	0.556	0.382
$R^2 \sim \text{Beta}(4, 1)$	0.175	0.016	2.813	1.412	2.780	1.409	0.977	0.488
$R^2 \sim \text{Beta}(4, 4)$	0.174	0.016	2.675	1.237	2.637	1.227	0.981	0.479

where  $\mu_0 = 0$ ,  $\tau_0^2 = 3$ ,  $\tau_1^2 = 100$  and  $\lambda_0 = -\log(0.01)/0.968$ . The vague prior has the same form as the PC prior except  $\sigma_u^2 \sim \text{InvGamma}(0.5, 0.0005)$  (Spiegelhalter et al. 2003).

We take 105,000 MCMC samples with the first 5000 discarded as burn-in, where the analysis is performed in JAGS. The results are in Figure 4 and Table 2. We also present trace plots in the supplemental materials to check convergence of the MCMC chain, as well as the effective sample size and computation time. We can see that the posterior distributions of  $R_n^2$  are very similar across the different methods. The posterior of  $W$ , however, varies across the different R2D2 priors with the Beta(4,1) and Beta(4,4) having the greatest mean and Beta(0.5,0.5) and Beta(1,1) having the smallest mean. The posterior distributions of  $W$  and  $\sigma_u^2$  are almost identical for the R2D2 priors which means that the vast majority of the global variance mass is shifted on the random effect variance and away from the fixed effect variance. The posterior for  $\sigma_u^2$  has the smallest mean for the PC prior, which follows from the fact that this prior shrinks the spatial variance toward zero. Lastly, the posterior of  $\rho$  varies across the different priors with the PC prior also yielding the smallest posterior mean.

## 4.2. Genomics Data

For our second data analysis, we apply the proposed method to high-dimensional genomics data from human breast tumors (Perou et al. 2000) in the *mixOmics* R package (Rohart et al. 2017). This data was pre-processed in Pérez-Enciso and Tenenhaus (2003) such that the response,  $Y_i$ , is 1 if tumor specimen  $i$  was analyzed before chemotherapy treatment, and 0 if it was analyzed after chemotherapy, for sample  $i \in \{1, \dots, n\}$  where  $n = 47$ . There are  $p = 1000$  gene expressions as predictor variables, so  $p \gg n$ , and we imputed the appropriate column mean of  $\mathbf{X}$  for any missing values. We consider the R2D2 prior with  $(a, b) \in \{(1, 5), (1, 10), (1, 20), (1, 30)\}$ , and compare with the Horseshoe prior. We choose these hyper-parameter combinations because we hypothesize that prior distributions with large mass near  $R^2 = 0$  are useful in high-dimensional contexts.

Since our focus of this analysis is prediction, we randomly split the data into train and test sets, where 75% of the data is used for training, and 25% for testing. The model is fit on the training data, and we compute the Brier Score (BS) and binary cross entropy (BCE) loss on the hold-out test data. If  $Y_i$  is the true response, and  $\hat{Y}_i$  is the predicted response, then  $\text{BS} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$  and  $\text{BCE} = -\frac{1}{n} \sum_{i=1}^n \{Y_i \log \hat{Y}_i + (1 -$

**Table 3.** Prediction results for genomics data with standard error in parentheses.

Method	BS	BCE
Horseshoe	<b>0.23</b> (0.01)	0.79 (0.07)
$R^2 \sim \text{Beta}(1, 5)$	0.29 (0.01)	0.89 (0.06)
$R^2 \sim \text{Beta}(1, 10)$	0.26 (0.00)	0.71 (0.01)
$R^2 \sim \text{Beta}(1, 20)$	0.26 (0.00)	<b>0.70</b> (0.01)
$R^2 \sim \text{Beta}(1, 30)$	0.26 (0.00)	0.71 (0.01)

NOTE: BS is Brier Score, and BCE is the binary cross entropy loss. Best results are labeled in bold.

$Y_i) \log(1 - \hat{Y}_i)\}$ . We compute  $\hat{Y}_i = [1 + \exp\{-(\hat{\beta}_0 + \mathbf{X}_i' \hat{\boldsymbol{\beta}})\}]^{-1}$  where  $\hat{\beta}_0$  and  $\hat{\boldsymbol{\beta}}$  are the posterior means of  $\beta_0$  and  $\boldsymbol{\beta}$ , respectively, and  $\mathbf{X}_i'$  are the predictor variables for the  $i$ th test data. The splitting procedure is repeated 50 times and we find the average of each metric. A good method will have small BS and BCE. The analysis was performed in `Stan` where we computed 10,000 MCMC samples with an additional 1000 for burn-in.

The results are in Table 3. In the supplemental materials, we also report the computation time, and average effective sample size for  $\boldsymbol{\beta}$  and the global variance parameter. We can see that the R2D2 prior with  $(a, b) = (1, 20)$  and Horseshoe perform the best, with smallest BCE and BS, respectively. Indeed, there is a clear trend that the R2D2 priors with more mass near  $R^2 = 0$  yield better predictions. However, there appears to be diminishing returns beyond  $R^2 \sim \text{Beta}(1, 10)$ .

## 5. Discussion

In this work, we proposed a novel method for choosing informative prior distributions in the generalized linear mixed model setting. The proposed prior is flexible and interpretable in terms of overall model fit as measured by a Bayesian  $R^2$ . There are many cases where the prior  $R^2$  can be induced exactly as well as general approximation strategies when an exact form is not possible. The main approach that we suggest is approximating the global variance prior with a generalized beta prime distribution because of its flexibility and ability to be implemented in standard software. Combined with an initial estimate of the intercept via a method of moments estimator and the GBP approximation in the `r2d2glm` package, we provide a simple and intuitive method for setting prior distributions in GLMMs.

If there is domain knowledge available on how well the model is expected to fit the data, then this could be used to inform prior choice for  $R^2$ . In the absence of any prior information, we suggest  $R^2 \sim \text{Beta}(1, 1)$  as a reasonable default choice. This prior expresses a full range of model fits, but should not be confused with a flat prior for  $\boldsymbol{\beta}$ . Choosing  $R^2 \sim \text{Beta}(1, b)$  for large  $b$ , or another prior with large mass near 0, is also a good choice, especially when working in a high-dimensional setting. Indeed, substantial prior mass near  $R^2 = 0$  prevents the model from overfitting by shrinking toward the “base model” of  $\boldsymbol{\beta} = \mathbf{0}_p$ .

The proposed approach naturally fits within the global-local shrinkage prior framework where  $W$  controls the global shrinkage, and  $\phi_j$  the local shrinkage. Our approach does differ from, for example, Hamura, Irie, and Sugawara (2022), who studies shrinkage priors for count data. Using the notation of our article, Hamura, Irie, and Sugawara (2022) let  $Y_i | \lambda_i \sim \text{Poisson}(\lambda_i e^{\eta_i})$  where  $\lambda_i \sim \text{Gamma}(\alpha, \beta / \nu_i)$ ,  $\nu_i \sim \pi(\cdot)$ , and  $\eta_i$  is some offset

which may depend on covariates, that is,  $\eta_i = \beta_0 + \mathbf{X}_i' \boldsymbol{\beta}$ . This article is primarily interested in the prior distribution for the local component ( $\nu_i$ ) and estimating the rate parameter ( $\lambda_i$ ). The R2D2 prior, on the other hand, is derived for the global component ( $W$ ), and is focused on estimating the regression components ( $\eta_i$ ).

A limitation of the proposed method is that the hierarchical framework only allows for random intercepts and not, for example, random slopes. Additionally, the finite mean and variance requirement precludes applications to some models, for example, extreme value analysis (Coles et al. 2001). We have also not proven concentration or shrinkage properties which is an avenue for future work. We could also extend the method to allow for other survival analysis settings beyond the uncensored Weibull model and models that are not GLMMs such as Bayesian deep learning. Finally, at present, the method cannot easily be written in `INLA` (Lindgren and Rue 2015), so we consider this an important next step.

## Supplemental Materials

In the supplemental materials, we provide: further discussion and comparison of the different coefficient of determination definitions, additional models and derivations for priors with closed-form expressions, a table showing different values of  $(a^*, b^*, c^*, d^*)$  for the GBP approximation, a simulation study, and additional results for the real-data analysis.

## Acknowledgments

The authors thank Brandon Feng for help with the derivation of the Weibull model.

## Disclosure Statement

The authors report there are no competing interests to declare.

## Funding

The authors thank the National Institutes of Health (R01ES031651-01) and King Abdullah University of Science and Technology (3800.2) for financial support.

## References

- Bai, R., and Ghosh, M. (2021), “On the Beta Prime Prior for Scale Parameters in High-Dimensional Bayesian Regression Models,” *Statistica Sinica*, 31, 1–23. [41]
- Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017), “The Horseshoe+ Estimator of Ultra-Sparse Signals,” *Bayesian Analysis*, 12, 1105–1131. [40]
- Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015), “Dirichlet-Laplace Priors for Optimal Shrinkage,” *Journal of the American Statistical Association*, 110, 1479–1490. [40,42]
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017), “Stan: A Probabilistic Programming Language,” *Journal of Statistical Software*, 76, 1–32. [45]
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009), “Handling Sparsity via the Horseshoe,” in *Artificial Intelligence and Statistics*, pp. 73–80, PMLR. [45]
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), “The Horseshoe Estimator for Sparse Signals,” *Biometrika*, 97, 465–480. [40,42]

- Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001), *An Introduction to Statistical Modeling of Extreme Values* (Vol. 208), London: Springer. [48]
- Cox, D. R., and Snell, E. J. (1989), *Analysis of Binary Data*, New York: Chapman & Hall. [43]
- Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based Geostatistics," *Journal of the Royal Statistical Society, Series C*, 47, 299–350. [42]
- Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020), "Intuitive Joint Priors for Variance Parameters," *Bayesian Analysis*, 15, 1109–1137. [40]
- Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," (comment on article by Browne and Draper), *Bayesian Analysis*, 1, 515–534. [43]
- Gelman, A., Goodrich, B., Gabry, J., and Vehtari, A. (2019), "R-squared for Bayesian Regression Models," *The American Statistician*, 73, 307–309. [42]
- Gelman, A., and Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge: Cambridge University Press. [42]
- George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [40]
- Hamura, Y., Irie, K., and Sugawara, S. (2022), "On Global-Local Shrinkage Priors for Count Data," *Bayesian Analysis*, 17, 545–564. [48]
- Hans, C. (2009), "Bayesian Lasso Regression," *Biometrika*, 96, 835–845. [40]
- Hastie, T. J. (2017), *Generalized Additive Models*, Boca Raton, FL: Routledge. [42]
- Hodges, J. S., and Sargent, D. J. (2001), "Counting Degrees of Freedom in Hierarchical and Other Richly-Parameterised Models," *Biometrika*, 88, 367–379. [40]
- Klein, N., Carlan, M., Kneib, T., Lang, S., and Wagner, H. (2021), "Bayesian Effect Selection in Structured Additive Distributional Regression Models," *Bayesian Analysis*, 16, 545–573. [41,42]
- Lewis, C. M., and Knight, J. (2012), "Introduction to Genetic Association Studies," *Cold Spring Harbor Protocols*, 2012, pdb-top068163. [40]
- Lindgren, F., and Rue, H. (2015), "Bayesian Spatial Modelling with R-INLA," *Journal of Statistical Software*, 63, 1–25. [48]
- Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192. [40]
- McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, 105–142. [43]
- Morokoff, W. J., and Caflisch, R. E. (1995), "Quasi-Monte Carlo Integration," *Journal of Computational Physics*, 122, 218–230. [44]
- Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [40]
- Perez, M.-E., Pericchi, L., and Ramirez, I. (2017), "The Scaled Beta2 Distribution as a Robust Prior for Scales," *Bayesian Analysis*, 12, 615–637. [41]
- Pérez-Enciso, M., and Tenenhaus, M. (2003), "Prediction of Clinical Outcome with Microarray Data: A Partial Least Squares Discriminant Analysis (pls-da) Approach," *Human Genetics*, 112, 581–592. [47]
- Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000), "Molecular Portraits of Human Breast Tumours," *Nature*, 406, 747–752. [47]
- Plummer, M., Stukalov, A., Denwood, M., and Plummer, M. M. (2016), Package "rjags", Vienna, Austria. [45]
- Polson, N. G., and Scott, J. G. (2012), "On the Half-Cauchy Prior for a Global Scale Parameter," *Bayesian Analysis*, 7, 887–902. [42]
- Polson, N. G., Scott, J. G., Clarke, B. S., and Severinski, C. (2012), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," in *Bayesian Statistics 9* eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, pp. 501–538, Oxford: Oxford University Press. [42]
- Rényi, A. (1961), "On Measures of Entropy and Information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 547–561, University of California Press. [45]
- Ribeiro Jr, P. J., Diggle, P. J., Ribeiro Jr, M. P. J., and Suggs, M. (2007), "The geoR Package," *R News*, 1, 14–18. [45]
- Ročková, V., and George, E. I. (2018), "The Spike-and-Slab Lasso," *Journal of the American Statistical Association*, 113, 431–444. [40]
- Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017), "mixomics: An R Package for 'Omics Feature Selection and Multiple Data Integration," *PLoS Computational Biology*, 13, e1005752. [47]
- Searle, S. R., Casella, G., and McCulloch, C. E. (2009), *Variance Components*, Hoboken, NJ: Wiley. [41]
- Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017), "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors," *Statistical Science*, 32, 1–28. [40,41,45]
- Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003), "Winbugs User Manual." [47]
- Thomson, M. C., Connor, S. J., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M., and Greenwood, B. (1999), "Predicting Malaria Infection in Gambian Children from Satellite Data and Bed Net Use Surveys: The Importance of Spatial Correlation in the Interpretation of Results," *The American Journal of Tropical Medicine and Hygiene*, 61, 2–8. [45]
- Zhang, Y., and Bondell, H. D. (2018), "Variable Selection via Penalized Credible Regions with Dirichlet-Laplace Global-Local Shrinkage Priors," *Bayesian Analysis*, 13, 823–844. [42]
- Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2022), "Bayesian Regression Using a Prior on the Model Fit: The r2-d2 Shrinkage Prior," *Journal of the American Statistical Association*, 117, 862–874. [40,41,42,43]