Bayesian Analysis (2025) **TBA**, Number TBA, pp. 1–34

Generalized Decomposition Priors on R2

Javier Enrique Aguilar [*,†,§] and Paul-Christian Bürkner[‡]

Abstract. The adoption of continuous shrinkage priors in high-dimensional linear models has gained widespread attention due to their practical and theoretical advantages. Among them, the R2D2 prior has gained popularity for its intuitive specification of the proportion of explained variance (R^2) and its theoretically grounded properties. The R2D2 prior allocates variance among regression terms through a Dirichlet decomposition. However, this approach inherently limits the dependency structure among variance components to the negative dependence modeled by the Dirichlet distribution, which is fully determined by the mean. This limitation hinders the prior's ability to capture more nuanced or positive dependency patterns that may arise in real-world data.

To address this, we propose the Generalized Decomposition R2 (GDR2) prior, which replaces the Dirichlet decomposition with the more flexible Logistic-Normal distribution and its variants. By allowing richer dependency structures, the GDR2 prior accommodates more realistic and adaptable competition among variance components, enhancing the expressiveness and applicability of R^2-based priors in practice. Through simulations and real-world benchmarks, we demonstrate that the GDR2 prior improves out-of-sample predictive performance and parameter recovery compared to the R2D2 prior. Our framework bridges the gap between flexibility in variance decomposition and practical implementation, advancing the utility of shrinkage priors in complex regression settings.

Keywords: prior specification, shrinkage priors, variance decomposition, regularization.

1 Introduction

Linear regression is one of the most widely used statistical techniques, serving as the foundation for many advanced modeling methods (Gelman and Hill, 2006; Gelman et al., 2013). However, its limitations become apparent in high-dimensional settings or when predictors exhibit multicollinearity (Hoerl and Kennard, 1970; Tibshirani, 1996; Giraud, 2014). In such cases, parameter estimation becomes unstable, and traditional models often fail to deliver reliable or interpretable results. To address these challenges, a variety of priors with exceptional theoretical and empirical properties have been developed over the last decade. These priors often assume that the true underlying model is sparse, meaning that many regression coefficients are exactly zero, and have the purpose of enforcing sparsity (Carvalho et al., 2010; van der Pas et al., 2016; van der Pas, 2021).

[*]Department of Statistics, TU Dortmund University, Germany, javier.aguilarr@icloud.com; url: https://jear2412.github.io

[†]Cluster of Excellence SimTech, University of Stuttgart, Germany

[‡]Department of Statistics, TU Dortmund University, Germany, paul.buerkner@gmail.com; url: https://paul-buerkner.github.io

[§]Corresponding author.

 https://doi.org/10.1214/25-BA1524

Even in cases where the true model is not sparse, sparse approximations have been widely recommended and successfully applied across a range of disciplines (Tibshirani, 1996; Giraud, 2014; van der Pas et al., 2014; Hastie et al., 2015; Zhang and Bondell, 2018; Bhadra et al., 2019). Dense solutions, while theoretically possible, tend to violate the principle of parsimony by increasing model complexity, which can lead to overfitting, reduced interpretability, and higher computational costs (Giraud, 2014; Hastie et al., 2015; Yang et al., 2016). As a result, there has been significant interest in priors that balance regularization and flexibility while remaining computationally efficient.

Continuous global-local shrinkage priors are a powerful class of priors that strike a balance between sparsity and regularization while maintaining flexibility through the incorporation of both global and local shrinkage effects (Tadesse and Vannucci, 2021). Over the past decade, these priors have been extensively studied and widely adopted within the Bayesian community due to their excellent empirical performance, computational efficiency, and the availability of fast implementations (van der Pas et al., 2016; Piironen and Vehtari, 2017; van der Pas et al., 2017; Johndrow et al., 2020).

Furthermore, a broad class of continuous global-local shrinkage priors has been shown to achieve near-minimax recovery rates (Ghosal et al., 2000; van der Pas et al., 2014, 2016; Tadesse and Vannucci, 2021). General conditions over the local scales have also been established to ensure that the posterior distribution concentrates at the minimax estimation rate under these shrinkage priors (van der Pas et al., 2016; Ročková, 2018; Bhadra et al., 2019). These theoretical guarantees provide a foundation for understanding why the priors commonly used in practice perform so well: the observed empirical success is, in essence, a reflection of the strong theoretical properties underpinning these models.

Prominent examples of shrinkage priors include the horseshoe, horseshoe-plus, three-parameter beta, normal gamma, and generalized double Pareto priors (Carvalho et al., 2010; Bhadra et al., 2017; Griffin and Brown, 2010; Armagan et al., 2011, 2013). These priors typically assume the local scales are conditionally independent given the global scale. However, an alternative family of priors introduces dependencies between the scales via joint distributions, albeit in a much smaller subset of the literature. Examples include the Dirichlet-Laplace, R2D2, and its multilevel extension, the R2D2M2 prior (Bhattacharya et al., 2015; Zhang et al., 2020; Aguilar and Bürkner, 2023). These approaches model local scales as proportions using a Dirichlet distribution. Therefore the competition among variance components will only be able to gravitate towards negative dependence structures fully specified by the mean. Yet, in reality, specific coefficients or groups may compete differently for the total variability than the Dirichlet would allow for. Thus, the ability to capture more nuanced or positive dependence structures among variance components, which might better reflect real-world scenarios is severely limited.

The value of moving beyond Dirichlet distributions to more flexible alternatives has long been recognized in fields such as Compositional Data Analysis Aitchison (1986); Greenacre et al. (2023), Categorical Data Analysis (Agresti and Hitchcock, 2005), and Machine Learning in areas such as Correlated Topic Modeling (Blei and Lafferty, 2007; Chen et al., 2013). Despite their success, these methods have not been explored in the context of shrinkage priors, leaving a gap in the literature.

Among existing shrinkage priors, the R2D2 prior is particularly notable for its focus on global quantities of interest—such as the proportion of explained variance (R^2)—and its ability to jointly regularize regression coefficients (Zhang et al., 2020; Aguilar and Bürkner, 2023; Mikkola et al., 2021). By simplifying prior elicitation and enhancing interpretability, the R2D2 framework has become a valuable tool for practitioners. However, its reliance on Dirichlet decompositions limits its flexibility in modeling dependencies among variance components.

The R2D2 prior specifies a prior on R^2 and uses a Dirichlet decomposition to allocate the variance across regression terms. In this work, we address the limitations of the Dirichlet-based decomposition by introducing the Generalized Decomposition R2 (GDR2) prior, which extends the R2D2 framework using Logistic-Normal decompositions and their variants. This extension enables richer, more expressive parameterizations of dependency structures among variance components, capturing complex relationships that are inaccessible with Dirichlet-based approaches. By combining the intuitive interpretability of the R2D2 framework with increased flexibility, the GDR2 prior represents a significant step forward in the development of continuous global-local shrinkage priors.

Our work is organized as follows: we begin by discussing preliminaries and the implied distributions on the proportions of variance and R^2 by shrinkage priors in Sections 2.1, 2.2 respectively. We proceed in Section 2.3 by introducing and describing the GDR2 prior framework for high-dimensional Bayesian linear regression. To circumvent the limitations associated with the Dirichlet decomposition, we propose the use of Logistic-Normal decompositions and its variants as an alternative in the decomposition step of R^2-based shrinkage priors in Section 2.4. Section 2.5 follows with a discussion on hyperparameter specification, offering intuitive explanations for selecting hyperparameters and exploring their effects on shrinkage. We believe this contribution is particularly valuable, as hyperparameter specification for Logistic-Normal distributions is rarely addressed in the literature.

In Section 3, we present the results of simulation studies and real life experiments conducted to evaluate the capabilities of the GDR2 prior. Our findings demonstrate that allowing for richer dependency structures among variance components improves out-of-sample predictive performance and parameter recovery. For these simulations and experiments, we implemented our models in the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2024), leveraging a parameterization optimized for fast convergence and efficient Hamiltonian Monte Carlo sampling. We also provide a Slice-within-Gibbs sampler for alternative usage in the Supplementary Material (Aguilar and Bürkner, 2025). To validate our approach, we tested the GDR2 prior on three real-world benchmarks commonly used in the shrinkage prior literature. Across all scenarios, the GDR2 prior consistently led to significantly improved results. All code and replication materials are openly available on the Open Science Framework (`https://osf.io/ns2cv/`).

2 Methods

2.1 Preliminaries

Consider the linear regression model

$$y_n = x_n' b + \varepsilon_n,\ n = 1, \ldots, N, \tag{1}$$

where y_n is the nth response value, x_n is the K dimensional vector of covariates for the nth observation, $b = (b_1, \ldots, b_K)'$ is the vector of regression coefficients, and $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ is the residual error, with σ being the residual standard deviation. Continuous Global-Local (GL) shrinkage priors (van der Pas et al., 2016; Van Erp et al., 2019; van der Pas, 2021) are a special type of continuous prior distributions that arise from scale mixtures of normals (West, 1987). They take on the form

$$b_k \mid \lambda_k^2, \tau^2, \sigma^2 \sim \mathcal{N}\left(0, \sigma^2 \lambda_k^2 \tau^2\right), \quad \lambda_k \sim \pi(\lambda_k), k = 1, \ldots, K, \quad \tau \sim \pi(\tau), \sigma \sim \pi(\sigma), \tag{2}$$

where λ_k represents local scales unique to each regression coefficient b_k, τ denotes a global scale shared across all coefficients, and $\pi(\cdot)$ represents the corresponding (hyper) prior distributions. Priors on b of the form (2) centered at zero are designed to shrink each coefficient towards zero thus favouring more sparse solutions. These priors have demonstrated empirical success among practitioners, and their theoretical foundations justify their widespread application (Ghosh et al., 2019; Johndrow et al., 2020; Follett and Yu, 2019; Kohns and Szendrei, 2024; Tadesse and Vannucci, 2021).

The choice of $\pi(\lambda_k)$ and $\pi(\tau)$ produces different GL shrinkage priors and will have an important effect on their properties. Popular shrinkage priors include the Horseshoe (Carvalho et al., 2010), the Normal-Gamma (Griffin and Brown, 2010), Generalized Double Pareto (Armagan et al., 2013), the Dirichlet-Laplace (Bhattacharya et al., 2015), the Regularized Horseshoe (Piironen and Vehtari, 2017), and the R2D2 prior (Zhang et al., 2020). The global scale τ controls the overall sparsity, ideally represent the proportion of true signals (Piironen and Vehtari, 2017; Van Erp et al., 2019; van der Pas, 2021). The local scales λ_k can either counteract or enforce the shrinkage towards zero.

The amount of shrinkage exerted on each coefficient towards zero can be quantified by studying the posterior distribution of the coefficients b given the scales λ_k, τ, σ and the observations y. The conditional posterior of the regression coefficients b is $b \mid y, \lambda, \tau, \sigma \sim \mathcal{N}\left(\bar{b}, \Sigma_b\right)$, with mean $\bar{b} = \mathbb{E}\left(b \mid y, \lambda, \tau, \sigma\right) = \left(X'X + 1/\tau^2\Lambda^{-1}\right)^{-1} X'y$ and covariance matrix $\Sigma_b = \sigma^2\left(X'X + 1/\tau^2\Lambda^{-1}\right)^{-1}$, where $\Lambda = \{\lambda_1^2, \ldots, \lambda_K^2\}$ is the diagonal matrix containing the local scales. If X is of full rank, then the conditional posterior mean can be expressed as $\bar{b} = \tau^2\Lambda\left(\tau^2\Lambda + (X'X)^{-1}\right)\hat{b}$ where $\hat{b}$ is the Maximum Likelihood Estimator (MLE). This representation highlights that the conditional posterior mean is a shrunken version of the usual MLE estimate.

For illustration, consider $X = I$, i.e. the normal means problem (Stein, 1981; van der Pas et al., 2016; Bhadra et al., 2017) and let $\sigma = 1$. In this case $\bar{b}_k = \left(1 - \frac{1}{1+\lambda_k^2\tau^2}\right)\hat{b}_k = (1 - \kappa_k)y_k$, where $\kappa_k := \frac{1}{1+\lambda_k^2\tau^2}$. The quantity κ_k serves as a *shrinkage factor* and

allows us to assess how much shrinkage the prior exerts on the maximum likelihood estimator of b_k (Polson et al., 2013; Bai and Ghosh, 2019; Bhadra et al., 2019; Aguilar and Bürkner, 2023). Importantly, by an application of Fubini's theorem, we can show that $\mathbb{E}[b_k \mid y_k] = (1 - \mathbb{E}[\kappa_k \mid y_k])\, y_k$, demonstrating that the posterior mean of b_k is at most y_k (Carvalho et al., 2010; Efron, 2011).

The prior on κ_k is determined by the interplay between the priors on λ_k and τ. Thus, calibrating shrinkage requires careful consideration of these priors. Notably, for a fixed value of τ, the definition of κ_k reveals that the posterior mean of b_k is influenced by the value of the local scale λ_k. Small values of λ_k will shrink b_k towards zero, whereas large values push the posterior mean of b_k towards the observation y_k. If either $\tau \to \infty$ or $\lambda_k \to \infty$ then the MLE is recovered.

In general, a good shrinkage prior should possess the following characteristics (Castillo and van der Vaart, 2012; Piironen and Vehtari, 2017; Van Erp et al., 2019; Bhadra et al., 2019; van der Pas, 2021): 1) *Heavy tails:* Sufficient mass in the tails of the prior is crucial to properly recover signals (i.e., truly nonzero coefficients). 2) *Sufficient mass near zero:* Shrinkage priors should allocate enough prior mass near zero in order to shrink redundant (truly zero) coefficients towards it. 3) *Efficient and stable sampling:* While theoretical properties are important, designing shrinkage priors should also consider efficient sampling from the posterior distribution.

Continuous GL shrinkage priors have gained popularity due to their capability to discern noise from signals, while yielding solutions that avoid a search over the entire space of models. They effectively shrink the influence of covariates considered unimportant towards zero while recovering the values of signals. This holds true even in scenarios where the true vector of regression coefficients b is ultra sparse (Bhadra et al., 2017; Song and Liang, 2017; van der Pas et al., 2017). Since continuous GL shrinkage priors cannot product *exact* zero estimates, additional posterior variable selection is required to induce sparsity in the posterior estimates, leading to a two step procedure: first perform inference of the GL shrunken model and second employ a decision rule to perform variable selection (Bai and Ghosh, 2019; Piironen et al., 2020; van der Pas et al., 2017; Zhang and Bondell, 2018; Pavone et al., 2020).

2.2 Implied Priors by the Local and Global Scales

Consider the linear regression model (1), where the covariates are standardized, such that $\mathbb{E}(x) = 0$ and $\text{var}(x) = \Sigma_X$, with Σ_X having a diagonal of ones. Assume a prior distribution for the regression coefficients b such that $\mathbb{E}[b] = 0$ and $\text{var}(b) = \sigma^2\Lambda$, where Λ is a diagonal matrix with entries $\lambda_1^2, \ldots, \lambda_K^2$. The conditional variance of the linear predictor $x'b$ is given by (Gelman et al., 2020):

$$\text{var}(x'b) = \sigma^2 \sum_{k=1}^{K} \lambda_k^2. \tag{3}$$

We define the quantity $\omega^2 \coloneqq \sum_{k=1}^{K} \lambda_k^2$ as the *total variance*. Equation (3) provides insight into the prior distribution for the proportion of explained variance, R^2 (Gelman

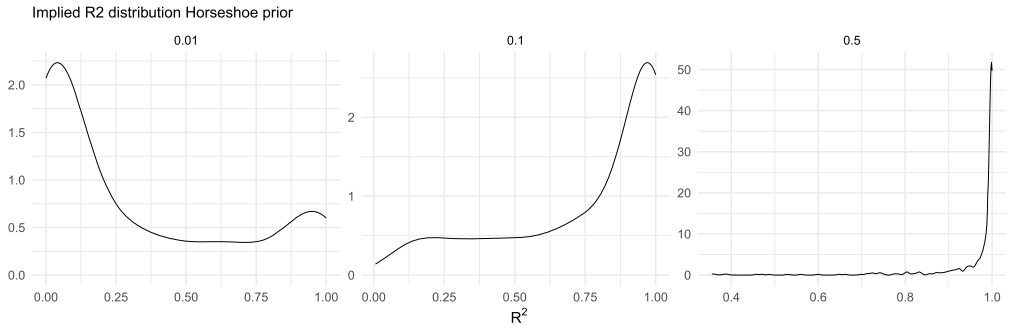

Figure 1: Shrinkage prior implied distribution for R^2 when considering the Horseshoe prior for different values of τ, which is interpreted as the a priori proportion of signals.

et al., 2020), induced by a GL shrinkage prior. The proportion of explained variance is defined as the square of the correlation coefficient between y and the linear predictor $x'b$:

$$R^2 := \text{corr}^2(y, x'b) = \frac{\text{var}(x'b)}{\text{var}(x'b) + \sigma^2} = \frac{\omega^2}{\omega^2 + 1}. \tag{4}$$

Thus, there is a one-to-one relationship between the priors set on R^2 and those implied on ω^2. Specifying a prior for b and σ implicitly defines a distribution for R^2. For instance, when b_k and σ have weakly informative priors (Gelman, 2006a), as shown in Aguilar and Bürkner (2023), the resulting prior for R^2 is highly concentrated near 1, even for moderate values of K. To illustrate the case of shrinkage priors, assume $X = I$ and use the well-known Horseshoe prior (Carvalho et al., 2010), which is specified as:

$$b_k \mid \lambda_k \sim \mathcal{N}(0, \lambda_k^2), \;\; \lambda_k \mid \tau \sim C^+(0, \tau), \tau \mid \sigma \sim C^+(0, \sigma), \tag{5}$$

where $C^+(0, \tau)$ denotes a Half Cauchy distribution (Gelman, 2006a; Polson and Scott, 2012) with scale parameter τ. We assume a fixed value for τ as a user-specified hyper-parameter, rather than assigning it its own prior. As demonstrated by van der Pas et al. (2016), it is crucial for τ to reflect the proportion of signals in order to guarantee signal recovery.

Figure 1 shows the implied distribution of R^2 when using the Horseshoe prior with different values for τ. Interpreting τ as the proportion of true signals, we observe that for very low τ values, the implied distribution of R^2 tends to concentrate near zero. On the other hand, when a user believes that even a small proportion of the elements in b are nonzero, i.e., $\tau \approx 0.1$, the distribution shifts toward one, indicating that the Horseshoe prior is focusing on identifying signals.

A significant issue arises when, despite having a low prior expectation for the proportion of nonzero coefficients, the model predicts high values of R^2. This leads to overestimation of the importance and magnitude of potential signals, which can conflict with user intuition about how the number and strength of signals relate to the

proportion of explained variance. Such an overestimation could mislead the user into thinking the model explains more variance than it realistically does, which could skew model interpretation and decision-making. Moreover, controlling the properties of the resulting distribution over R^2, such as its mean and variance, becomes challenging, as they are implicitly determined by the prior and the data. Given the total variance ω^2 in the model, we can compute the *proportion of explained variability* for each coefficient, defined as

$$\phi_k := \frac{\lambda_k^2}{\sum_{k=1}^{K} \lambda_k^2}. \tag{6}$$

Since $\phi_k \geq 0$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \phi_k = 1$, the vector $\phi := (\phi_1, \ldots, \phi_K)$ lies in the $K - 1$ dimensional simplex. Thus, the joint distribution of the λ_k values induces a distribution on the simplex, represented stochastically by ϕ. Understanding the implied distribution of ϕ provides insights into how the total variance ω^2 is distributed among the coefficients and how they compete for it. This understanding can help determine how much of the total variance is allocated to each coefficient, which is key for assessing model fit and identifying influential predictors. For example, if we assume $\lambda_k \sim \text{Gamma}(\alpha_k, 1)$, we obtain the Dirichlet distribution $\phi \sim \text{Dirichlet}(\alpha)$, where $\alpha := (\alpha_1, \ldots, \alpha_K)$ is the concentration vector (Lin, 2016). In this case, the correlations between the elements of ϕ are negative and determined by the mean of the distribution. As with the implicit distribution for R^2, specifying a prior on λ does not necessarily give us direct control over the properties of the implied distribution for ϕ, except in certain special cases, such as the one just mentioned.

The emphasis on implied quantities that are more intuitive and user-friendly has largely been overlooked in the shrinkage prior literature, with only a few notable exceptions (Zhang and Bondell, 2018; Zhang et al., 2020; Aguilar and Bürkner, 2023). This oversight is understandable, as statisticians have primarily concentrated on developing robust automatic procedures capable of tackling complex tasks. However, we believe that emphasizing the more intuitive perspective could not only enhance user comprehension and the practical applicability of well-established shrinkage priors but also inspire the development of new, innovative methodologies.

2.3 Generalized Decomposition R2 Priors

We have discussed how the specification of a shrinkage prior on the regression coefficients b implies priors for interpretable quantities such as R^2 and ϕ. It is equally insightful to explore the reverse scenario, that is, establishing a prior over the proportion of explained variance R^2 and the proportions of total variance ϕ to derive a prior over the regression coefficients b. This concept has been explored before to define prior distributions in the context of additive regression models of varying complexity (Zhang et al., 2020; Aguilar and Bürkner, 2023). However, they always considered $\phi \sim \text{Dirichlet}(\cdot)$. While this choice is certainly a reasonable one, it raises the question of what happens when ϕ follows a distribution other than Dirichlet.

Employing alternative distributions for ϕ allow us to probe how the proportions of total variance $\phi_k, k = 1, \ldots, K$ compete for the total variability ω^2. Left unattended

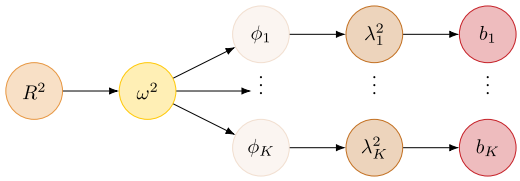

Figure 2: Schematic of the GDR2 prior construction. First, a distribution is assigned to R^2, then its uncertainty is propagated to the regression terms via the simplex distribution of the proportions of explained variance. Finally, a distribution is used to allocate variance to each regression term.

and without any proper control, this competition is naturally guided towards a certain degree of negative dependence structures, i.e., if ϕ_k increases then $\phi_j, j \neq k$ decreases since there is competition for a total quantity. Specifically, the Dirichlet distribution shows this phenomenon (Aitchison, 1986; Kruijer et al., 2010; Lin, 2016). However, in reality, specific (groups of) coefficients may compete differently for the total variability than the Dirichlet would assume and allow for. Indeed, as we show in our simulations and real-world case studies, other prior choices may lead to more favourable results.

To this end, we introduce a prior distribution family that provides greater flexibility in capturing the intricate relationships among the proportions of explained variance. We call them *Generalized Decomposition R2* (GDR2) priors. The main idea is schematically presented in Figure 2. We begin by setting a prior over R^2. According to Equation (4), there exists a one to one relationship between R^2 and the total variability ω^2, immediately establishing a prior distribution on ω^2. Subsequently, we select a distribution for the proportions of explained variance ϕ that can properly represent a desired dependence structure. Afterwards, we shape the variances of the regression coefficients b_k by setting $\lambda_k^2 = \phi_k \omega^2, i = 1, \ldots, K$. Finally, we let b_k follow a normal distribution centered at zero with variance $\sigma^2 \lambda_k^2$.

We opt for a Beta distribution to characterize R^2 and we write $R^2 \sim \text{Beta}(\cdot, \cdot)$. The Beta distribution, offering various parametrizations (Kruschke, 2015), is chosen for its flexibility. We specifically parametrize it in terms of the prior mean μ_{R^2} and prior "precision" φ_{R^2} (also known as mean-sample size parameterization). This choice facilitates an intuitive integration of prior knowledge into the R^2 prior. The hyperparameters μ_{R^2}, φ_{R^2} are interpretable expressions of domain knowledge that represent the existing relationship between the included covariates and the response variable.

If (a_1, a_2) are the canonical shape parameters of the Beta distribution, then $\mu_{R^2} = \frac{a_1}{a_1+a_2}$ and $\varphi_{R^2} = a_1 + a_2$. This selection implies that ω^2 follows a Beta-Prime distribution, with parameters μ_{R^2}, φ_{R^2}, which we denote as $\omega^2 \sim \text{BetaPrime}(\mu_{R^2}, \varphi_{R^2})$. Figure 3 illustrates the flexibility that the Beta distribution can offer in expressing prior knowledge about R^2. The figure also depicts the corresponding Beta Prime prior for ω^2. For instance, for values of $(\mu_{R^2}, \varphi_{R^2}) = (0.5, 1)$, we obtain a bathtub-shaped prior on R^2, concentrating the majority of the mass near the extremes $R^2 = 0$ and $R^2 = 1$. This

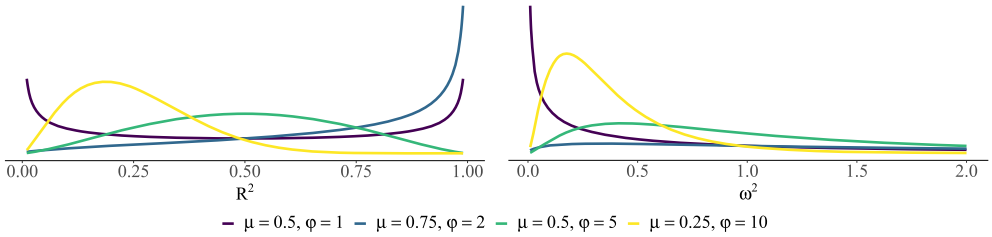

Figure 3: Beta and Beta Prime densities for various values of the prior mean μ_{R^2} and precision φ_{R^2}.

prior behavior signals the user's expectation of a model containing either a substantial amount of either noise or signal, respectively.

Moving forward, the *(probability) simplex* of dimension $K-1$ ($K \in \mathbb{N}$), is defined as $\mathcal{S}^K := \{x \in \mathbb{R}^K | \sum_{k=1}^K x_k = 1, x_k \geq 0, k = 1, \ldots, K\}$. Let $p(\phi \mid \nu)$ represent the chosen distribution for ϕ on the simplex $\mathcal{S}^k$, where ν represents specific hyper-parameters related to the chosen distribution. For example, in the case of the Dirichlet distribution $\nu = \{\alpha\}$. Although the Dirichlet distribution aids in providing analytical results, interpretation, computational efficiency, and alignment with common user knowledge, it exhibits undesirable properties in practice, which we discuss further in Section 2.4.

The next step is to specify a distribution for the coefficients that captures the attributed variance for each regression coefficient b_k, given by $\lambda_k^2 = \phi_k \omega^2$ for $k = 1, \ldots, K$. To do so, we define $b_k | \sigma, \lambda_k \sim \mathcal{N}(0, \sigma^2 \lambda_k^2)$, which integrates our prior into the GL shrinkage prior framework, typically modeled as a scale mixture of normals (West, 1987). In contrast, Bhattacharya et al. (2015) and Zhang et al. (2020) have used double exponential distributions for b_k to express the attributed variances, although the former did not account for the perspective of variance proportions. Additionally, Aguilar and Bürkner (2023) adopted normal distributions to model the attributed variances, offering a different approach to the distribution of the coefficients.

If model (1) includes an intercept b_0, the prior specification is simplified by setting a prior on the centered intercept $\tilde{b}_0$, which is implied when $\mathbb{E}(x_i) = 0$. The original intercept b_0 can then be recovered after model fitting using a simple linear transformation (Bürkner, 2017; Goodrich et al., 2020). Common choices for priors on b_0 include a normal prior with mean $\mathbb{E}(y)$ and a user-chosen scale, which depends on the scale of y, or Jeffrey's prior, which is improper and flat in this case (Good, 1962). Next, we specify a prior for the residual variance σ^2 (or equivalently the residual standard deviation σ). Following the recommendations of Gelman (2006b), we set a Half Student-t prior on σ with ν degrees of freedom and scale η. Consistent with Bürkner (2017) and Aguilar and Bürkner (2023), we set $\eta \approx \text{sd}(y)$, as both the prior's expected mean and variance are

proportional to η. Together, the full GDR2 model is given by

$$\begin{aligned} &y_n \sim \mathcal{N}(\mu_n, \sigma^2), \quad \mu_n = b_0 + \sum_{k=1}^{K} x_{nk} b_k, \quad n = 1, \ldots, N, \\ &b_0 \sim p(\cdot), \quad b_k \sim \mathcal{N}(0, \sigma^2 \phi_k \omega^2), \quad k = 1, \ldots, K, \\ &\omega^2 = \frac{R^2}{1 - R^2}, \quad \phi \sim p(\nu_\phi), \quad R^2 \sim \text{Beta}(\mu_{R^2}, \varphi_{R^2}), \quad \sigma \sim p(.). \end{aligned} \tag{7}$$

While, in this paper, we focus on the linear regression case, the additive structure of the prior readily allows its extension to linear multilevel models or other kinds of additive models if desired.

2.4 Distributions on the Simplex

In this section, we explore potential distributions for ϕ defined on the simplex. While extensions of the Dirichlet distribution are often considered, we do not pursue them here, as they share similar limitations (Connor and Mosimann, 1969; Barndorff-Nielsen and Jørgensen, 1991; Ongaro and Migliorati, 2013; Chow, 2022). Instead, we draw on methods from Compositional Data Analysis (CDA) (Aitchison, 1986; Boogaart and Tolosana-Delgado, 2013; Greenacre et al., 2023), Categorical Data Analysis (CatDA)(Agresti and Hitchcock, 2005) and Correlated Topic Modeling (CTM) (Blei and Lafferty, 2007; Chen et al., 2013), where the logistic normal distribution (and its variants) is the preferred approach for analyzing simplicial variables.

The Dirichlet Distribution

The Dirichlet distribution is widely employed for modeling data existing in the simplex (Gupta and Richards, 2001; Lin, 2016). A vector $\phi = (\phi_1, \ldots, \phi_K)' \in \mathcal{S}^K$ follows a Dirichlet distribution with concentration vector $\alpha \coloneqq (\alpha_1, \ldots, \alpha_K)', \alpha_k > 0, k = 1, \ldots, K$ if its density has the following form:

$$p(\phi|\alpha) = \frac{\Gamma(\alpha_+)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \phi_k^{\alpha_k - 1} \mathbb{1}(\phi \in \mathcal{S}^k), \tag{8}$$

where $\alpha_+ = \sum_{k=1}^{K} \alpha_k$. The mean and covariance of the distribution are respectively expressed as: $\mathbb{E}(\phi_k) = \frac{\alpha_k}{\alpha_+} =: \tilde{\alpha}_k$ and $\text{cov}(\phi_k, \phi_j) = \frac{\delta_{kj}\tilde{\alpha}_k - \tilde{\alpha}_k\tilde{\alpha}_j}{1+\alpha_+}$, where δ_{kj} is the Kronecker delta function. If $k \neq j$, then $\text{cov}(\phi_k, \phi_j) = \frac{-\tilde{\alpha}_k\tilde{\alpha}_j}{\alpha_+ + 1} < 0$ for all k, j. Once the mean vector $\tilde{\alpha}$ is chosen, only the scalar α_+ remains to express the entire variance-covariance structure. In particular, the covariance function is always proportional to the product of the corresponding means.

The Dirichlet distribution gained prominence during the conjugate prior era, as it serves as the conjugate prior for the multinomial likelihood (Ongaro and Migliorati, 2013; Lin, 2016; Wang and Polson, 2024). Its interpretability is another advantage, the

hyperparameters α_k can easily be interpreted in relation to the distribution's behavior. In particular, the quantity $\tilde{\alpha}_k = \mathbb{E}(\phi_k)$ represents the relative a priori importance of the kth element ϕ_k. To illustrate, consider $\alpha = (1, \ldots, 1)'$, which gives a distribution that is flat over all possible simplexes, leading to a suitable choice in the absence of additional prior knowledge. Generally, setting $\alpha = (a_\pi, \ldots, a_\pi)'$ with a concentration parameter $a_\pi > 0$ (i.e., a *symmetric Dirichlet distribution*) drastically reduces the number of hyperparameters to specify and allows us to globally control the shape of the Dirichlet distribution with a single value. A symmetric Dirichlet distribution is useful when there is no prior preference that favors one component over another. Values $a_\pi < 1$ concentrate mass on the edges of the simplex, while $a_\pi > 1$ results in concentration around the center of the simplex. The user can also specify asymmetric Dirichlet distributions by choosing different values for the individual α_k. This represents different a priori expected importance of the corresponding components.

A major limitation of the Dirichlet distribution is its inherent lack of flexibility in modeling dependencies among the elements of ϕ. Its covariance structure is determined entirely by the mean, restricting the ability to specify custom dependency structures for simplex data. Moreover, the sum-to-one constraint induces negative correlations between components, making the Dirichlet unsuitable for scenarios where components may exhibit positive correlations (Connor and Mosimann, 1969; Wong, 1998). Even many complex negative correlation patterns cannot be accurately captured, further highlighting the limitations of the Dirichlet distribution in accommodating diverse dependency structures.

From a technical perspective, the Dirichlet distribution imposes additional constraints that limit its applicability. It requires properties such as closure under marginalization and conditioning, as well as complete subcompositional independence. This last property implies that all renormalized subsets of ϕ are independent of each other (Aitchison, 1986). Subcompositional independence has been widely criticized, as it imposes $1/2K(K-3)$ constraints on the covariance structure, rendering the Dirichlet implausible for most real-world applications (Aitchison and Shen, 1980; Aitchison, 1986; Ongaro and Migliorati, 2013). These limitations make the Dirichlet unsuitable for modeling complex or realistic dependency structures, necessitating more flexible alternatives like the Logistic Normal distribution, which we present below.

The Logistic Normal Distribution

The logistic normal distribution, proposed by Aitchison and Shen (1980) offers a solution to overcoming limited correlation structures. The core concept involves mapping a multivariate normal random variable defined from $\mathbb{R}^K$ into the $K-1$ dimensional simplex $\mathcal{S}^K$ by the use of log ratio transformations, as described below. This approach enables researchers to make use of well known established techniques for multivariate analysis in the unbounded real space and then seamlessly transforming the results into the simplex (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011).

Let $\phi \in \mathcal{S}^K$, the *additive log-ratio* (alr) transformation, relative to the Kth component, is defined as: $\eta = \text{alr}(\phi) := \left(\ln\left(\frac{\phi_1}{\phi_K}\right), \ldots, \ln\left(\frac{\phi_{K-1}}{\phi_K}\right)\right) \in \mathbb{R}^{K-1}$ (Aitchison and

Shen, 1980). This transformation establishes a one-to-one correspondence between the elements of the simplex and the log-ratio vectors $\ln\left(\frac{\phi_i}{\phi_K}\right)$, allowing any statement about the components of the simplex to be equivalently expressed in terms of these log-ratios (Aitchison and Shen, 1980; Aitchison, 1986). By mapping compositional data into the unconstrained real space $\mathbb{R}^{K-1}$, the difficulties associated with working in a constrained space are effectively removed.

The alr transformation is inherently asymmetric since the log-ratios require a reference component. The last component ϕ_K is usually chosen as reference, but any component can be selected based on interpretability or practical considerations (Aitchison, 1986). Guidelines for choosing the reference component have been proposed (Greenacre et al., 2023), but in general, practitioners of compositional data analysis (CDA) and correlated topic models (CTM) recognize that the choice of reference has minimal impact on results or inference (Pawlowsky-Glahn and Buccianti, 2011; Boogaart and Tolosana-Delgado, 2013; Greenacre et al., 2023).

The inverse alr transformation is defined by the *standard logistic* (*softmax*) transformation with a fill-up term for the reference component: $\phi_k = \frac{e^{\eta_k}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}, k = 1, \ldots, K-1, , \phi_K = 1 - \sum_{k=1}^{K-1} \phi_k$. We say that $\phi \in \mathcal{S}^K$ follows an *additive logistic normal* (ALN) distribution if $\eta = \text{alr}(\phi)$ follows a multivariate normal distribution in $\mathbb{R}^{K-1}$. The density of the ALN distribution is given by (Aitchison, 1986):

$$p(\phi|\mu, \Sigma)| = |2\pi\Sigma|^{-1/2} \left(\prod_{k=1}^{K} \phi_k\right)^{-1} \exp\left\{-\frac{1}{2}\left[\text{alr}(\phi) - \mu\right]' \Sigma^{-1} \left[\text{alr}(\phi) - \mu\right]\right\}. \tag{9}$$

We denote this as $\phi \sim \text{ALR}(\mu, \Sigma)$, where μ and Σ depend on the chosen reference component.

A symmetric representation for ϕ, independent of a reference component, can be obtained by starting from the unconstrained space and mapping into the simplex. Specifically, let $\eta \sim \mathcal{N}_K(\mu, \Sigma)$ and apply the *symmetric logistic transformation*:

$$\phi_k = \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}}, \ k = 1, \ldots, K. \tag{10}$$

In this case, ϕ is said to follow a (symmetric) *Logistic Normal* (LN) distribution with parameters μ, Σ, denoted as $\text{LogisticNormal}(\mu, \Sigma)$. For identifiability, either the constraint $\sum_{k=1}^{K} \eta_k = 0$ or $\eta_K = 0$ can be imposed, the latter being equivalent to the ALR (Bishop, 2006; Blei and Lafferty, 2007; Goodfellow et al., 2016). If we consider the sum-to-zero constraint, then $\eta_k = \ln\left(\frac{\phi_k}{g(\phi)}\right)$, where $g(\phi)$ denotes the geometric mean of ϕ.

Importantly, $\Sigma = (\sigma_{ij}), i, j = 1, \ldots, K$, represents the covariance structure of the log-ratios, not of the raw proportions $\text{cov}(\phi_i, \phi_j)$. If one chooses to model ϕ using either the ALR or LN distribution, transitioning between them is entirely feasible via a straightforward linear relationships that we show in the Supplementary Material (Aguilar and

Bürkner, 2025). This flexibility allows researchers to select the representation most suited to the practical requirements of their analysis.

While moments of all non-negative orders, $\mathbb{E}_\phi\left(\prod_{k=1}^{K}\phi_k^{m_k}\right)$, $m_k \geq 0$, $k = 1, \ldots, K$, exist, their expressions are not reducible to simple analytic forms (Aitchison and Shen, 1980; Frederic and Lad, 2008; Holmes and Schofield, 2022). Nonetheless, when analyzing quantities in the simplex, it is customary to study ratios or logarithms of ratios of simplex components rather than raw proportions (Aitchison, 1986; Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn and Buccianti, 2011; Creus-Martí et al., 2021). The primary motivation for this is the so-called *coherence* requirement in compositional data analysis, which ensures that the relationships between components remain consistent and interpretable, whether considering the full composition or its subcompositions Greenacre et al. (2023). Coherence is especially important since the constant-sum constraint in compositional data implies that changes in one component inherently affect all others.

Let $\sigma_{ik,jl} = \operatorname{cov}\left(\ln\left(\phi_i/\phi_k\right), \ln\left(\phi_j/\phi_l\right)\right)$ represent the covariance of log-ratio quantities, then the following key relationships hold:

$$\mathbb{E}\left\{\ln\left(\phi_j/\phi_k\right)\right\} = \mu_j - \mu_k, \quad \sigma_{ik,jl} = \sigma_{ij} + \sigma_{kl} - \sigma_{il} - \sigma_{jk}. \tag{11}$$

Thus, the mean vector μ and covariance matrix Σ fully determine the expectations and pairwise covariances of log-ratio quantities. This implies that no recalculations are necessary when transitioning between different log-ratio representations, as all relevant information is encoded in μ and Σ. Moreover, one can derive μ and Σ directly from knowledge of any pair of log-ratio quantities, and vice versa (Aitchison, 1986). We explore these parameters in greater detail in Section 2.5, where we incorporate the LN distribution as a prior for ϕ within the GDR2 framework.

2.5 Hyperparameter Specification in GDR2 Priors

As explained in Section 2.3, the Beta prior for R^2 can be parametrized in terms of the prior mean μ_{R^2} and precision φ_{R^2}. In cases where the user has limited prior knowledge, a flat distribution over the unit interval can be specified by setting $(\mu_{R^2}, \varphi_{R^2}) = (0.5, 2)$. Alternatively, a bathtub-shaped distribution can be chosen with $(\mu_{R^2}, \varphi_{R^2}) = (0.5, 1)$, which places more mass on extreme values of R^2. This reflects a prior belief that either there is no relationship between the response and the covariates, or that the covariates fully explain the variability in the response. An application of the law of Total Variance shows that $\operatorname{var}(b) = \mathbb{E}[\operatorname{var}(b \mid \phi, \omega^2)] + \operatorname{var}[\mathbb{E}(b|\phi, \omega^2)] = \mathbb{E}(\omega^2)\mathbb{E}(\operatorname{cov}(\phi))$. Thus, the hyperparameters selected for ω^2 can control the tail behavior of b_k, potentially inducing infinite variance and heavy tails in its marginal distribution. This configuration enhances sensitivity to detecting signals.

As discussed in Section 2.4, we let ϕ follow either a Dirichlet or an LN distribution. When setting $\phi \sim \text{Dirichlet}(\alpha)$ the GDR2 prior is equivalent to the R2D2 prior (Zhang et al., 2020) for non-hierarchical models, except that b_k follows a normal distribution rather than a Double Exponential. The former scenario has been explored by Aguilar

and Bürkner (2023), who show that if $\phi \sim \text{Dirichlet}(\alpha)$ with $\alpha = (a_\pi, \ldots, a_\pi)$, then a_π determines the behavior of the marginal distribution of b_k near the origin. Thus, a_π can be theoretically chosen to tailor the properties of the marginal distribution of b_k. As an alternative, the elements of α can be elicited from subject matter experts to represent the a priori relative importance of covariates – although we consider it very hard elicit such knowledge for high-dimensional problems.

If $\phi \sim \text{LogisticNormal}(\mu, \Sigma)$, the prior specification involves defining the mean vector μ and covariance matrix Σ for the log-ratios. Setting $\mu_k = 0 \; \forall k$ (Equation (11)) encodes the belief that all proportions ϕ_k are equally weighted, expressing no preference for any specific component. To assign equal weights to a subset of proportions, $\mu_i = c$ can be specified for $i \in I \subset \{1, \ldots, K-1\}$. If the proportions can be grouped into G mutually exclusive sets $I_g \subset \{1, \ldots, K-1\}$, group-specific weights c_g may reflect their relative importance, enabling the inclusion of prior hierarchical or structural information.

To better understand the effect of assigning values to μ, consider the symmetric logistic mapping in Equation (10) and assume $\Sigma = \sigma_\phi^2 I$. Setting $\mu_k = c \; \forall k$ is equivalent to $\mu_k = 0$, as the logistic mapping is shift-invariant (Bishop, 2006). This choice reflects a lack of prior knowledge about relative importance and by symmetry it can be shown that $\mathbb{E}[\phi_k] = 1/K$. When $\mu_k = c_k$, proportions will differ depending on the sign of c_k; components with $c_k > 0$ will, on average, be larger, while those with $c_k < 0$ smaller. To see this, let $R_k = e^{\eta_k}$ and $S = \sum_j e^{\eta_j}$. Using a first order Taylor expansion around the mean, we can approximate (Stuart and Ord, 2009):

$$\mathbb{E}[\phi_k] = \mathbb{E}\left(\frac{e^{\eta_k}}{\sum_j e^{\eta_j}}\right) \approx \frac{\mathbb{E}[R_k]}{\mathbb{E}[S]} = \frac{e^{c_k + \frac{\sigma^2}{2}}}{\sum_j e^{c_j + \frac{\sigma^2}{2}}} = \frac{e^{c_k}}{\sum_j e^{c_j}}, \tag{12}$$

where we compute the expectations by noting that R_k follows a log normal distribution. This argument extends to a diagonal covariance matrix Σ with different scales σ_j; whose values interact with c_j to redistribute mass in favor or against ϕ_k.

The relationship becomes more complex for a general Σ and μ. While a Taylor expansion provides an approximation,[1] it relies on the term $\text{Cov}(R_k, S)$, which lacks a closed-form solution, requiring Monte Carlo methods for estimation. While this offers insights into the interaction of key quantities that determine the marginal expectation, its added complexity may outweigh its utility. Our experiments in Section 3 show that, setting $\mu_k = 0$ is a robust default choice in the absence of prior knowledge, balancing simplicity and performance. The specification of Σ can be understood by analyzing the implied prior on the norm of η (the log ratios) and how this translates to the simplex. We first consider the case of a diagonal $\mu = 0, \Sigma = \sigma_\phi^2 I$. Let $\eta \sim N(0, \sigma_\phi^2 I)$, since η is a sub-gaussian variable with independent entries (Giraud, 2014; Vershynin, 2018), $\|\eta\|$ concentrates around $\sqrt{K}\sigma_\phi$. Specifically, the following two-sided tail bound holds: $\mathbb{P}\left(\left|\frac{\|\eta\|}{\sqrt{K}\sigma_\phi} - 1\right| \geq t\right) \leq 2\exp\left(-\frac{K\sigma_\phi^2 t^2}{2C}\right)$ for a constant $C > 0$. We illustrate this behavior in Figure 4.

[1] see Supplementary Material (Aguilar and Bürkner, 2025).

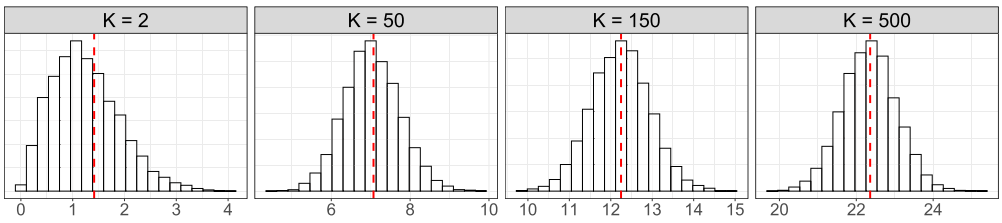

Figure 4: Implied distribution on the norm of a Multivariate Normal as the dimension K increases. The red line shows the value $\sqrt{K}\sigma_\phi$, which is the point of concentration, where K is the number of covariates and $\sigma_\phi = 1$.

The user-specified σ_ϕ defines bounds for the norm of the logits, setting a "budget" that can be distributed across them. For $\sigma_\phi \to 0$, the logits concentrate near zero, resulting in $\phi_k \to 1/K$. Conversely, as σ_ϕ increases, the logits spread further, increasing the quota for $\|\eta\|$. If $\mu \neq 0$, nonsparse elements will be determined by the locations in which $\mu_k \neq 0$, especially if the budget is low.

The choice of σ_ϕ balances the variability in the simplex as K grows. Setting $\sigma_\phi = \sqrt{\gamma/K}, \gamma > 0$ ensures that $\|\eta\|^2 \approx \gamma$, maintaining a consistent magnitude for the logits regardless of K. Alternatively, choosing $\sigma_\phi = \sqrt{\gamma}$ (independent of K) allows the logits' norm to grow proportionally to $\sqrt{K}$, introducing greater variability. This variability interacts with μ to influence how mass is distributed across ϕ_k. For scenarios where sparsity is critical, setting $\sigma_\phi = \sqrt{\log(K)/K}$ slows the growth of the logits' norm, preventing excessively large values and reducing variability. This encourages sparse distributions, with mass concentrated at locations dictated by μ. Such a setup is particularly effective in high-dimensional settings. Users must carefully consider the interaction between μ and σ_ϕ. Lower values of σ_ϕ concentrate the mass of ϕ at positions where μ has higher values, promoting focused distributions. In contrast, higher values of σ_ϕ introduce variability, allowing for more diffuse distributions across the simplex.

When Σ incorporates non-zero correlations, deriving a general inequality for the concentration of measure of $\|\eta\|$ becomes significantly more complex. We do not explore this case further, since we consider the intuition provided by σ_ϕ sufficient for practical specification. However, if users have prior knowledge about how the variables are correlated, they can include this information via a correlation matrix Ω_ϕ, using the decomposition $\Sigma = \mathrm{D}_\phi \Omega_\phi \mathrm{D}_\phi$, where D_ϕ is a diagonal matrix containing the scales, and Ω_ϕ is a correlation matrix. For example, if the coefficients are ordered and proportions are expected to be similar for adjacent variables, one might use an autoregressive correlation structure. Alternatively, for cases where groups of variables exhibit distinct interdependencies, a block correlation matrix may be appropriate. This flexibility allows users to encode meaningful structural relationships directly into the prior.

Prior Matching

An alternative approach to specifying the hyperparameters of the LN distribution is to draw upon theoretical insights and practical experience from the $\phi \sim \text{Dirichlet}(\alpha)$ case. Specifically, we propose using the Dirichlet distribution as an initial guide to determine the hyperparameters of the LN distribution. Our method involves defining a Dirichlet distribution that captures the desired theoretical and practical characteristics of the prior. Then, we minimize a divergence measure between the Dirichlet and LN distributions. This automated process, which we term *prior matching*, provides a systematic way to derive the prior mean and covariance matrix for the LN distribution.

Prior matching simplifies the complex task of specifying LN hyperparameters by using the more analytically tractable Dirichlet distribution as a reference. While the LN distribution offers increased flexibility compared to the Dirichlet, its analytical properties can be challenging to study, particularly in the context of shrinkage priors. Prior matching mitigates these challenges, automating the derivation of the LN's prior mean and covariance matrix for improved efficiency and user convenience. Exact matching is not the ultimate goal, as our objective is to develop priors that outperform the Dirichlet. Prior matching serves as a starting point, enabling users to leverage the unique flexibility of the LN distribution. For instance, the LN distribution addresses a critical limitation of the symmetric Dirichlet distribution: sampling difficulties and numerical instability as $a_\pi \to 0$.

As divergence metric for prior matching, we propose to use the Kullback Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence between two probability distributions f, g is defined as $\text{KL}\left(f||g\right) \coloneqq \mathbb{E}_f\left[\ln\left(\frac{f(x)}{g(x)}\right)\right] = \int_{-\infty}^{\infty} \ln\left(\frac{f(x)}{g(x)}\right) f(x)dx$. To find the LN distribution that best approximates a given Dirichlet distribution, we minimize the KL divergence between them. Letting $f \sim \text{Dirichlet}(\alpha)$ with α fixed and $g \sim \text{LogisticNormal}(\mu, \Sigma)$, the optimal LN parameters are obtained by solving: $\mu^*, \Sigma^* = \arg\min_{\mu,\Sigma} \text{KL}\left(f||g\right)$. The closed form solution is given by

$$\mu_k^* = \delta(\alpha_k) - \delta(\alpha_K), \sigma_{kk}^* = \varepsilon(\alpha_k) + \varepsilon(\alpha_K),\ k = 1, \ldots, K,\quad \sigma_{kj}^* = \varepsilon(\alpha_K)\ (k \neq j), \quad (13)$$

where $\delta(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ and $\varepsilon(x) = \delta'(x)$ are the digamma and trigamma functions respectively (Aitchison and Shen, 1980; Aitchison, 1986; Jeffrey et al., 2007). Given their analytical form, these quantities can be computed instantly, adding no computational overhead to the procedure. While other divergence measures could be considered, the closed-form solution for the KL divergence offers a clear advantage in terms of both practical applicability and computational efficiency.

To exemplify KL matching, consider a symmetric Dirichlet distribution with a single hyperparameter a_π. The value of a_π controls the amount of shrinkage induced by the R2D2 prior (Aguilar and Bürkner, 2023). When $a_\pi \leq 1/2$ the prior will be unbounded near the origin, leading to strong shrinkage towards zero. Users who wish to replicate this shrinkage behavior in the LN distribution can propose a value of $a_\pi \leq 1/2$ and use Equation (13) to calculate the corresponding LN hyperparameters: $\mu_k = 0, \sigma_{kk} =$

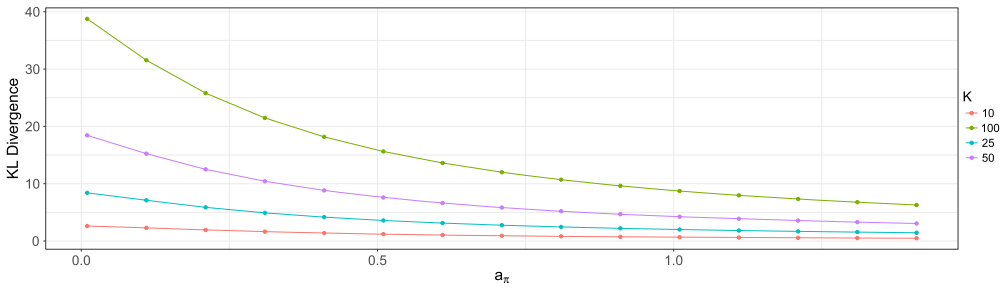

Figure 5: KL divergence between a symmetric Dirichlet distribution (parameter a_π) and the closest logistic normal distribution as a_π and the number of covariates (K) vary. KL divergence decreases monotonically with a_π.

$2\varepsilon(a_\pi)$, and $\sigma_{kj} = \varepsilon(a_\pi)$. As a_π increases, the KL divergence between the Dirichlet and LN distributions decreases (Aitchison, 1986). When $a_\pi \to 0$, the KL divergence increases significantly since the Dirichlet allocates mass along the simplex edges, a challenge for the LN distribution to approximate, given that it is defined within the simplex's interior. This is illustrated in Figure 5 for varying values of a_π and different dimensions K.

A straightforward modification to KL matching involves discarding the correlations in the derived Σ^*, using only its diagonal elements to create a diagonal covariance matrix $\widetilde{\Sigma}^*$. In this case the scales $\sigma^2_{kk} = 2\varepsilon(a_\pi), \forall k$ and $\text{cor}(\eta_i, \eta_j) = 1/2, i \neq j$. In high dimensional settings, it is unrealistic to expect such correlations as a default, as we typically anticipate only a small number of signals (Tosh et al., 2022). Extracting only the scales, while discarding the correlations, is sensible since this type of correlations may be rendered as noise or irrealistic in high dimensions. The deviation between $\widetilde{\Sigma}^*$ and Σ^* can be quantified by the KL divergence in log-ratio space. In the case where Σ^* is obtained through KL matching with a symmetric Dirichlet distribution with hyperparameter a_π, the value of divergence has a simple form: $\frac{1}{2} \ln \left(\frac{2^{K-1}}{K} \right)$, which is independent of a_π.

We conclude this section by noting that discussions on selecting hyperparameter values for the Logistic Normal distribution as a prior are limited, and practical guidance is often scarce. Researchers often prefer to set priors over the hyperparameters (Blei and Lafferty, 2007; Mimno et al., 2008; Xun et al., 2017), either by using priors of the form $\mu \mid \Sigma \sim \mathcal{N}(\mu_0, \gamma\Sigma)$ with $\Sigma \mid W \sim \text{InverseWishart}(\delta, W^{-1})$ (Chen et al., 2013), or by employing a Variational Approximation to the posterior distribution, where the variational family assumes a diagonal structure for Σ (Blei and Lafferty, 2007; Blei et al., 2017). However, these methods often lack sufficient discussion on the choice of hyperpriors for the hyperparameters. In this work, we opted not to explore these approaches due to their added computational complexity, comparable performance with fixed hyperparameters, and the limited understanding of how such specifications interact with shrinkage priors in fixed-sample settings.

3 Experiments and Case Studies

We conducted simulation studies to assess the performance of our generalized R2 priors under various conditions commonly encountered in practical scenarios. These investigations provide insights into how the priors behaves with finite samples and explores different possibilities for the total variance decomposition. Additionally, we present case studies utilizing real-world datasets that are commonly used as benchmarks for shrinkage priors. The shrinkage priors considered in our simulations include the Beta Prime, Dirichlet-Laplace, Horseshoe, and the proposed GDR2 model, which incorporates Dirichlet and Logistic Normal decompositions. While these priors provide a useful context for comparison, our primary focus is on the GDR2 model, particularly its variance decomposition mechanisms. Results for the other priors, evaluated using their default settings, are detailed in Aguilar and Bürkner (2025). Notably, the GDR2 model demonstrates on par performance to these well-established priors and even outperforms them in specific scenarios. This highlights the robustness of GDR2 and underscores the flexibility and potential of decomposition-based approaches.

All models were implemented in the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2024), which provides an extended implementation of the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2011), an adaptive form of Hamiltonian Monte Carlo (HMC) (Brooks et al., 2011). The Stan code for the GDR2 prior is included in the Supplementary Material (Aguilar and Bürkner, 2025), and all associated data and code can be found in `https://osf.io/ns2cv/`.

3.1 Simulations

Generative Models

We use model (1) as a generative model in our simulations, adapting it to accommodate various forms of data encountered in practice, including diverse levels of sparsity and different dependency structures among the covariates. The design matrix X was sampled from a multivariate normal distribution with mean 0 and covariance matrix Σ_x derived from an AR(1) correlation structure with $\rho_x \in \{0, 0.8\}$. The intercept b_0 was drawn from a normal distribution with mean zero and variance $\sigma_I^2 = 4$. We fixed the sample size at $N = 100$, and varied the number of regression coefficients $K \in \{50, 150, 750\}$ to represent both low-dimensional ($K < N$) and high-dimensional ($K \gg N$) scenarios. The residual standard deviation σ was adjusted using Equation (4) to maintain the true explained variance $R_0^2 \in \{0.25, 0.6\}$.

The regression coefficients b_k, were generated in two ways: 1) *Simulated Coefficients:* Coefficients were sampled from a normal distribution with zero mean and covariance Σ_b. Two forms of Σ_b were considered: (A) diagonal with $\sigma_b^2 = 9$, and (B) AR(1) with autocorrelation $\rho_b = 0.8$ and diagonal elements $\sigma_b^2 = 9$. Sparsity was induced by setting each coefficient to zero with probability $v = 0.75$, reflecting real-world scenarios where coefficients arise from random processes (Bhattacharya et al., 2015; Griffin and Brown, 2010, 2013). 2) *Fixed Coefficients*: This setup is akin to examples encountered in the shrinkage prior literature (Carvalho et al., 2010; Griffin and Brown, 2010; Bhattacharya

et al., 2015; Zhang et al., 2020). We place concentrated signals $b^* \in \{3, 7\}$ in specific locations of the coefficient vector b. The first 5 and the last 5 elements are set to b^*, while all others are set to 0.

We have specified Dirichlet and symmetric Logistic Normal distributions for the proportions of total variance ϕ. Two hyperparameter configurations for R^2 were tested: *default* $(\mu_{R^2}, \varphi_{R^2}) = (0.5, 1)$ and *uniform* $(0.5, 2)$. When let ϕ follow a symmetric Dirichlet distribution with concentration parameter $a_\pi \in \{0.5, 1\}$. The *default* case $a_\pi = 0.5$ encourages shrinkage by pushing mass toward the edges of the simplex, whereas the "*uniform*" case $a_\pi = 1$ reduces shrinkage. When $a_\pi \leq 0.5$, the marginal prior distributions of the coefficients b_k become unbounded near the origin, enforcing sparsity by shifting mass to the simplex edges (Aguilar and Bürkner, 2023).

To specify the hyperparameters of $\phi \sim \text{LogisticNormal}(\mu, \Sigma)$, we used KL matching (Section 2.5), referring to the outcome as *Full KL*. Specifically, for $a_\pi = 0.5$, we set $\mu_k = 0$, $\sigma_{kk} = \pi$, and $\sigma_{kj} = \frac{\sqrt{2}}{2}\pi$; for $a_\pi = 1$, we set $\mu_k = 0$, $\sigma_{kk} = \frac{\sqrt{3}}{3}\pi$, and $\sigma_{kj} = \frac{\sqrt{6}}{6}\pi$, both yielding pairwise correlations of 0.5. We also consider a variant restricting Σ^* to a diagonal matrix, termed *Scales KL*. Table 1 summarizes the decompositions.

As discussed in Section 2.5, the marginal distributions of ϕ for the Logistic Normal (LN) priors may exhibit spikes around $1/K$ as K increases and $\mu = 0$. When using diagonal covariance matrices, the LN prior allows for greater variability in the marginals, providing more flexibility, serving as a less informative prior in comparison. If prior knowledge about the location of strong signals exists, this should be encoded in μ. In this study, we mimic complex scenarios with minimal prior knowledge about the coefficients to test the prior under such conditions.

We crossed the default and uniform hyperparameter specifications for R^2 and a_π, resulting in four hyperparameter specifications. We present a subset of the results for the default-default case $(\mu_{R^2}, \varphi_{R^2}, a_\pi) = (0.5, 1, 0.5)$ and show others in the Supplementary Material (Aguilar and Bürkner, 2025). Additionally, we include results for an informative case in the fixed coefficient setup in Supplementary Material (Aguilar and Bürkner, 2025), where α is constructed in a way that incorporates strong user knowledge to emphasize important signals, matching the locations of the signals. This helps assess the extent to which both decompositions benefit from user-provided information.

After fully crossing all conditions, we obtained a total of 192 different simulation configurations. For each configuration, we generated $T = 100$ datasets consisting of $N = 100$ training observations y_n, $n = 1, \ldots, N$. Predictive metrics (discussed below) were computed based on $N_{\text{test}} = 500$ test observations, which were generated independently

Name	Label	Distribution	Details
Dirichlet	D2	Dirichlet	$\alpha_k = a_\pi$
Full KL	LNF	logistic normal	μ^*, Σ^* from KL matching
Scales KL	LNS	logistic normal	$\mu^*, \text{diag}(\Sigma^*)$ from KL matching

Table 1: Different decompositions that are considered in the simulations.

from the training data using the same data-generating mechanism. The simulations were carried out on the Linux HPC Cluster (LiDO3) at TU Dortmund University.

Evaluation Metrics

We evaluated and compared the performance of the models based on two criteria: out-of-sample predictive performance and parameter recovery (Vehtari and Ojanen, 2012; Robert, 2007). Out-of-sample predictive performance was assessed using the expected log-pointwise predictive density (ELPD) (Vehtari and Ojanen, 2012), computed on the test data as $\text{ELPD} = \sum_{i=1}^{N} \ln\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^{(s)})\right)$, where $\theta^{(s)}$ represents the sth posterior draw for $s = 1, \ldots, S$. The ELPD quantifies predictive accuracy across the N data points, evaluating the quality of the overall predictive distribution. A higher ELPD value indicates better out-of-sample predictive performance. Log probability scores, such as ELPD, are often recommended as a general-purpose metric when there is no specific reason to use an alternative (Vehtari and Ojanen, 2012; Vehtari et al., 2016, 2022).

Parameter recovery was evaluated using the posterior Root Mean-Squared Error (RMSE), calculated as $\text{RMSE} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(b_k^{(s)} - b_k\right)^2}$, where $b_k^{(s)}$ represents the sth posterior draw for $k = 1, \ldots, K$, and b_k denotes the true value of the kth regression coefficient. The RMSE serves as a comprehensive measure of estimation error, naturally capturing the tradeoff between bias and variance. In our analysis, we compute three versions of RMSE: (1) Averaged over all coefficients, (2) only over the truly zero coefficients, (3) only over the truly nonzero coefficients. To provide a more comprehensive evaluation of the posterior inference properties, we assess the coverage of 95% marginal credible intervals for each approach based on average interval width, coverage proportion, specificity, and sensitivity. We also present Receiver Operating Characteristic (ROC) curves. These results are presented in the Supplementary Material (Aguilar and Bürkner, 2025).

To compare multiple models based on a metric of interest $\mathcal{F}$ (e.g., ELPD, RMSE), we computed the difference relative to the best-performing model for a given dataset. Let $\{m_1, \ldots, m_l\}$ represent the set of models, and denote the best model with respect to $\mathcal{F}$ by m^*. The Delta metric, $\Delta\mathcal{F}_i$, for the ith model is defined as $\Delta\mathcal{F}_i = \mathcal{F}(m_i) - \mathcal{F}(m^*)$, where $i = 1, \ldots, l$. This approach highlights differences between models while accounting for variations caused by randomness in the simulated datasets, thus focusing the evaluation on meaningful distinctions.

Metrics for diagnosing and evaluating the Markov Chain Monte Carlo (MCMC) sampler, including Effective Sample Size (ESS) and Rhat (Geyer, 1992; Brooks et al., 2011; Vehtari et al., 2021), are provided in the Supplementary Material (Aguilar and Bürkner, 2025). Overall, the results indicate effective sampling from the posterior, as ESS values are sufficiently high and Rhat values fall within acceptable ranges. Additionally, all models demonstrate comparable computational efficiency, with similar runtimes observed across the board.

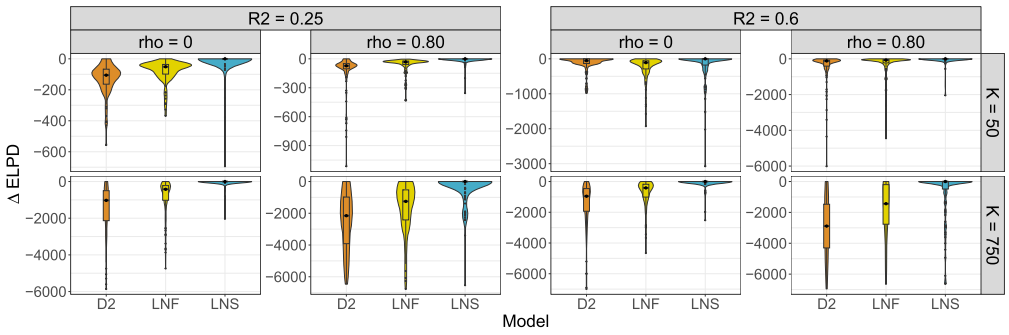

Figure 6: **Simulated coefficients setup.** ΔELPD evaluated on test datasets of size $N = 500$.

Results: Simulated Coefficients Setup

We present results for the scenario where the coefficients are generated with an AR(1) correlation structure, characterized by $\rho_b = 0.8$ and marginal variances $\sigma_b^2 = 9$, alongside an average sparsity of $\nu = 0.75$. In real-world applications, it is reasonable to encounter clusters of coefficients that are correlated with each other, even though these correlations do not necessarily translate into significant effects on the response variable. Figure 6 shows the distribution of ΔELPD across simulations under various conditions, visualized as violin plots with embedded boxplots (Hintze and Nelson, 1998; Stryjewski, 2010). The results demonstrate that Logistic Normal (LN) decompositions consistently outperform their Dirichlet counterparts. Notably, the LN priors exhibit the best average performance across all scenarios. The LNS version is the clear winner in predictive performance.

Figure 7 depicts the distribution of ΔRMSE across various simulation conditions. Similar to predictive performance, the LN priors demonstrate the most robust parameter recovery among the competing R^2-based models, regardless of the values of ρ, R^2, or K. The overall RMSE results reflect a balance between the model's posterior error for truly zero and nonzero coefficients. To explore this further, Figure 7 separately illustrates ΔRMSE for truly zero and nonzero coefficients.

These results suggest that LN decompositions achieve more effective shrinkage compared to Dirichlet counterparts, reducing false positives and improving the detection of truly zero coefficients. Dirichlet decompositions impose less shrinkage, which can result in poorer performance when coefficients are truly zero. The trends for nonzero coefficients are similar. Of particular note is the case with $K = 750$, a challenging scenario where shrinkage priors are expected to overshrink. In this high-dimensional setting, LN decompositions show improved behavior when moving from $\rho = 0$ to $\rho = 0.80$. This improvement is also present when transitioning from a low signal ($R^2 = 0.25$) to a moderate signal ($R^2 = 0.60$). Overall, LN decompositions offer a more robust and adaptable approach to shrinkage in sparse, high-dimensional contexts.

In summary, using a zero mean vector and a diagonal covariance matrix for LN

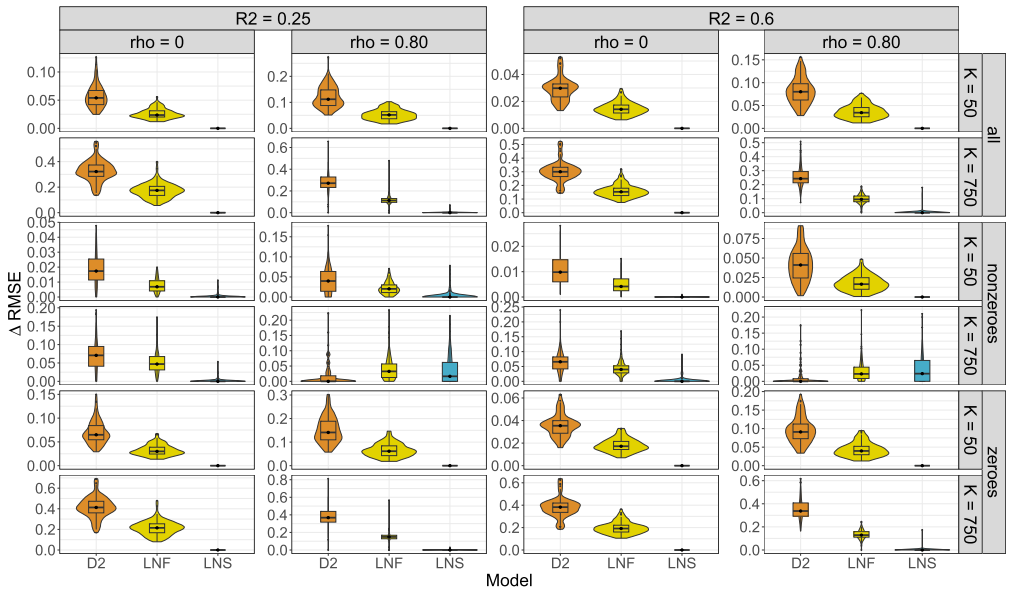

Figure 7: **Simulated coefficients setup.** Violin plots with embedded box plots for ΔRMSE under different simulation conditions. ΔRMSE has been partitioned with respect to truly zero and nonzero coefficients.

priors appears to be an effective strategy, particularly when prior knowledge about relationships between coefficients is limited—a common scenario in practice. These results highlight the potential of LN-based approaches as robust default alternatives to Dirichlet-based shrinkage priors. Even with default parameter settings, LN priors demonstrate competitive or superior performance across diverse scenarios.

Results: Fixed Coefficients Setup

We show the results when the fixed value is equal to 3. Figure 8 reveals that the LN decompositions consistently excel in out-of-sample predictive performance when compared to the Dirichlet counterpart. The LNS stands out as the most efficient predictive model. The most pronounced disparities emerge in scenarios characterized by high correlation and dimensionality ($\rho = 0.8, K = 750$), where the performance of the Dirichlet prior is notably subpar. The results for ΔRMSE for all of the coefficients and the zero coefficients behave similarly to the simulated coefficients setup. The LN decompositions perform clearly and uniformly better the Dirichlet for all the coefficients and for the zero coefficient only. This indicates again that LN decompositions are better at both overall RMSE and noise detection. We only show the case for nonzero coefficients in Figure 9. A complete overview over the other cases is available in the Supplementary Material (Aguilar and Bürkner, 2025).

The Dirichlet decomposition excels only in the in low-signal, high-correlation set-

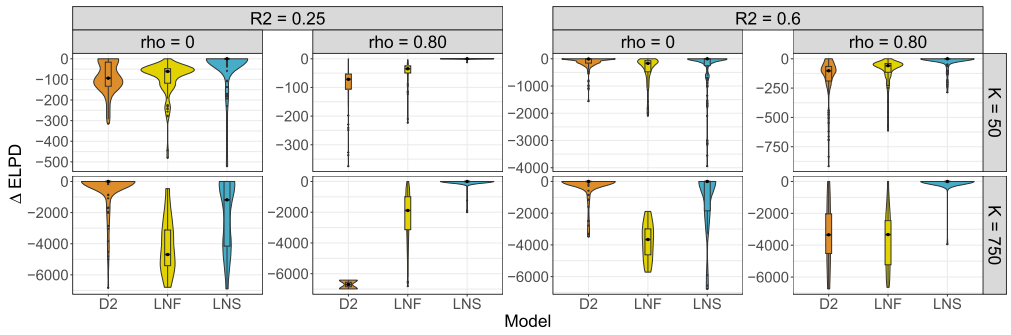

Figure 8: **Fixed coefficients setup.** ΔELPD evaluated on the test datasets for the fixed coefficients when the signal have a value of 3.

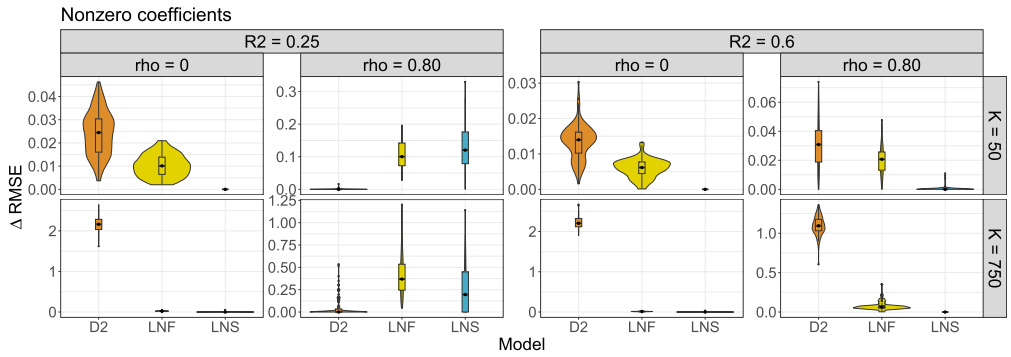

Figure 9: **Fixed coefficients setup.** ΔRMSE for truly nonzero coefficients.

tings. LN decompositions achieves better performance in all other scenarios. These results suggest that the effectiveness of LN decompositions could be enhanced even more by incorporating prior structure that mitigates correlations in the design matrix. If we focus on the $\rho = 0.8, K = 750$ case in both Figures 8 and 9, we can see that there is a tradeoff between out-of-sample predictive performance and detecting signals in that case. The Dirichlet lacks sufficient shrinkage (worsening out-of-sample predictions), while the LN may overshrink some signals but improves them overall.

3.2 Real-World Case Studies

We assess the predictive performance of our proposed prior using three high-dimensional real-world datasets, each with distinct correlation structures among covariates. The Cereal dataset contains starch content measurements for 15 observations with 145 infrared spectra predictors, sourced from the R package `chemometrics` (Filzmoser, 2023). The Cookie dataset includes fat content measurements from 72 samples with 700 Near - infrared spectra predictors, originally introduced by Osborne et al. (1984) and available

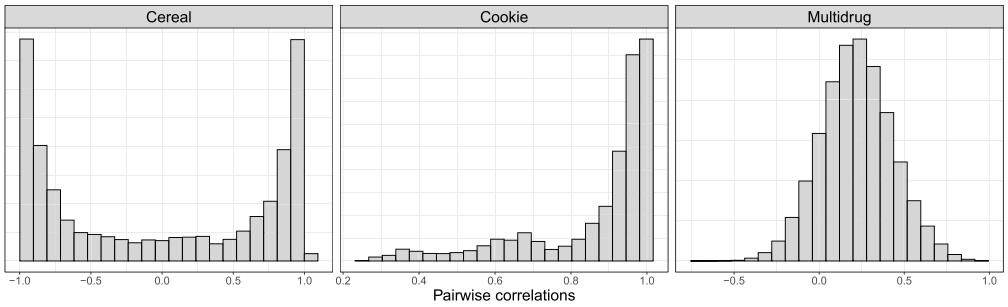

Figure 10: Histograms of pairwise correlations of the covariates of the different datasets being considered.

in the orphaned `ppls` R package (Boulesteix, 2014). The Multidrug dataset, derived from a pharmacogenomic study on drug concentration and gene expression (Szakács et al., 2004), comprises 60 observations and 853 covariates after removing missing values. The Cereal and Multidrug datasets have been analyzed in prior work on shrinkage priors (Polson and Scott, 2011; Griffin and Brown, 2013; Zhang et al., 2020), while the Cookie dataset has been studied in Osborne et al. (1984); Ghosh and Ghattas (2015); Zhang et al. (2020). These datasets exhibit varied predictor dependencies: mixed correlations in Cereal, strong positive correlations in Cookie, and low-to-moderate correlations in Multidrug (Zhang et al., 2020), as visualized in Figure 10.

We use ELPD to quantify out-of-sample predictive performance. Since we don't have an independent test set available, we use cross-validation to estimate out-of-sample ELPD as follows: We split each dataset into training and test sets, with 75% of observations used for training and the remaining 25% for testing. We repeat this 100 times and use the resulting ELPD values, which are subsequently summed to obtain an overall ELPD estimate. We compare the performance of the Dirichlet, LNF, and LNS decompositions with specific parameter settings, namely $(\mu_{R^2}, \varphi_{R^2}, a_\pi) = (0.5, 1, 0.5)$.

We summarise the results in Table 2. LNS consistently demonstrates superior ELPD performance compared to both Dirichlet and LNF across the diverse datasets. The latter

Dataset	Cereal ($N = 15, K = 145$)	Cookie ($N = 72, K = 700$)	Multidrug ($N = 60, K = 853$)
Model	ΔELPD (SE)	ΔELPD (SE)	ΔELPD (SE)
LNS	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
LNF	−101.4 (85.1)	−673.1 (471.5)	−2950.4 (743.8)
D2	−96.8 (42.2)	−518.9 (274.4)	−5464.2 (794.7)

Table 2: Differences in ELPD and standard deviations for the different datasets considered, computed through pairwise comparisons with the model having the highest ELPD (in the first row). The initial value is zero, and subsequent rows display negative values that indicate the difference with the best model (Vehtari et al., 2023). See Vehtari et al. (2016) for details on standard error calculations.

two priors clearly differ from each other only for the Multidrug data where LNF was better than Dirichlet. This showcases that priors that exert more shrinkage on complex, high-dimensional datasets can be beneficial in terms of predictive performance.

4 Discussion

We introduced the Generalized Decomposition R2 prior framework for high-dimensional Bayesian regression, building on and extending the R2D2 prior (Zhang et al., 2020; Aguilar and Bürkner, 2023). Our key innovation lies in allowing to vary the prior used for the explained variance decomposition, enabling a more nuanced exploration of dependency structures. Previous methods are confined to Dirichlet decompositions, and thus can only express negative covariances tied to the means of the explained variance proportions. We overcome this limitation by employing the logistic normal distribution, facilitating the expression of covariance structures by transitioning from the simplex to the unconstrained real space.

Our simulations and real-world case studies demonstrate substantial gains in predictive performance when using logistic normal priors for the proportions of explained variance. The flexible dependency structures encoded in the logistic normal allow for a broader range of prior assumptions compared to the Dirichlet distribution. Specifically, we show that use of logistic normal based decompositions leads to superior out-of-sample predictive performance compared to the use of Dirichlet counterparts. In terms of parameter recovery, the Logistic normal prior models showed on par and superior performance compared to Dirichlet in most scenarios.

Our primary goal was to build well-predicting models, using the full set of available covariates and strong shrinkage on the corresponding coefficients. That said, as is the case for continuous shrinkage priors more generally, they do not create exact sparsity directly. Rather, redundant coefficients are shrunk to values close to zero, but not to zero exactly. As a result, predictive capabilities are not necessarily directly visible in few nonzero coefficients but rather implicitly distributed also among the almost zero coefficients, at least in high-dimensional settings (Piironen and Vehtari, 2017). In other words, we argue that our models (and other continuous shrinkage priors models) should not be used directly for variable selection (for more details and discussion see Piironen et al. (2020); Zhang and Bondell (2018); Tadesse and Vannucci (2021)). Rather, we recommend a two-step procedure where, after fitting a well predicting reference model with all covariates in the first step, we apply a separate, dedicated variable selection procedure in the second step, for example, projection predictive variable selection (Piironen et al., 2020; Catalina et al., 2020; McLatchie et al., 2023).

In future research, one important direction will be to study the influence and optimal choice of the logistic normal hyperparameters in more detail. While the hyperparameter choices proposed here already showed strongly improved performance compared to Dirichlet, it is still unclear how much room for improvement remains within the class of logistic normal decomposition priors. One promising direction we see in this context is the inclusion of covariate grouping or covariate dependency information in the prior

(Boss et al., 2023). But exactly how this information should be encoded, and if its inclusion is actually beneficial for the prior's performance, remains to be studied.

Acknowledgments

We would like to thank Aki Vehtari and his research group at Aalto University for their valuable comments and insightful discussions on an earlier version of this work. In particular, we are grateful to David Kohns for his detailed and constructive feedback.

Funding

Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2075-390740016 and DFG Project 500663361. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). We acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by DFG Project 271512359.

Supplementary Material

Supplementary Material for: "Generalized Decomposition Priors on R2" (Aguilar and Bürkner, 2025) (DOI: 10.1214/25-BA1524SUPP; .pdf). In this supplementary material, we present further derivations and results.

References

Agresti, A. and Hitchcock, D. B. (2005). "Bayesian inference for categorical data analysis." *Statistical Methods and Applications*, 14(3): 297–330. MR2211337. doi: https://doi.org/10.1007/s10260-005-0121-y. 2, 10

Aguilar, J. E. and Bürkner, P.-C. (2023). "Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior." *Electronic Journal of Statistics*, 17(1): 1711–1767. MR4609453. doi: https://doi.org/10.1214/23-EJS2136. 2, 3, 5, 6, 7, 9, 13, 16, 19, 25

Aguilar, J. E. and Bürkner, P.-C. (2025). "Supplement to "Generalized Decomposition Priors on R2"." *Bayesian Analysis*. doi: https://doi.org/10.1214/25-BA1523SUPP. 3, 12, 14, 18, 19, 20, 22, 26

Aitchison, J. and Shen, S. M. (1980). "Logistic-Normal Distributions: Some Properties and Uses." *Biometrika*, 67(2): 261–272. URL http://www.jstor.org/stable/2335470 MR0581723. doi: https://doi.org/10.2307/2335470. 11, 12, 13, 16

Aitchison, J. J. (1986). *The statistical analysis of compositional data / J. Aitchison.*. Monographs on statistics and applied probability (Series). Chapman and Hall. MR0865647. doi: https://doi.org/10.1007/978-94-009-4109-0. 2, 8, 10, 11, 12, 13, 16, 17

Armagan, A., Clyde, M., and Dunson, D. B. (2011). "Generalized beta mixtures of Gaussians." In *Advances in neural information processing systems*, 523–531. 2

Armagan, A., Dunson, D. B., and Lee, J. (2013). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23(1): 119. MR3076161. 2, 4

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233. MR4034859. doi: https://doi.org/10.1080/02331888.2019.1662017. 5

Barndorff-Nielsen, O. and Jørgensen, B. (1991). "Some parametric models on the simplex." *Journal of Multivariate Analysis*, 39(1): 106–116. MR1128675. doi: https://doi.org/10.1016/0047-259X(91)90008-P. 10

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). "The Horseshoe+ Estimator of Ultra-Sparse Signals." *Bayesian Analysis*, 12(4): 1105–1131. MR3724980. doi: https://doi.org/10.1214/16-BA1028. 2, 4, 5

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). "Lasso Meets Horseshoe: A Survey." *Statistical Science*, 34(3): 405–427. MR4017521. doi: https://doi.org/10.1214/19-STS700. 2, 5

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(512): 1479–1490. PMID: 27019543. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 2, 4, 9, 18

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer. URL https://books.google.de/books?id=kTNoQgAACAAJ MR2247587. doi: https://doi.org/10.1007/978-0-387-45528-0. 12, 14

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association*, 112(518): 859–877. MR3671776. doi: https://doi.org/10.1080/01621459.2017.1285773. 17

Blei, D. M. and Lafferty, J. D. (2007). "A correlated topic model of Science." *The Annals of Applied Statistics*, 1(1): 17–35. MR2393839. doi: https://doi.org/10.1214/07-AOAS114. 2, 10, 12, 17

Boogaart, K. and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*, 209–253. MR3099409. doi: https://doi.org/10.1007/978-3-642-36809-7. 10, 12, 13

Boss, J., Datta, J., Wang, X., Park, S. K., Kang, J., and Mukherjee, B. (2023). "Group Inverse-Gamma Gamma Shrinkage for Sparse Linear Models with Block-Correlated Regressors." *Bayesian Analysis*, 1–30. MR4770323. doi: https://doi.org/10.1214/23-BA1371. 26

Boulesteix, N. K. A.-L. (2014). *ppls: Penalized Partial Least Squares*. R package version 1.6-1. URL http://cran.nexr.com/web/packages/ppls/index.html 24

Brooks, Gelman, S., and Jones, A. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 1 edition. 18, 20

Bürkner, P.-C. (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, 80(1): 1–28. URL https://www.jstatsoft.org/index.php/jss/article/view/v080i01 9

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): 1–32. URL https://www.jstatsoft.org/index.php/jss/article/view/v076i01 3, 18

Carvalho, Polson, and Scott (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. URL http://www.jstor.org/stable/25734098 MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 1, 2, 4, 5, 6, 18

Castillo, I. and van der Vaart, A. (2012). "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069–2101. MR3059077. doi: https://doi.org/10.1214/12-AOS1029. 5

Catalina, A., Bürkner, P.-C., and Vehtari, A. (2020). "Projection Predictive Inference for Generalized Linear and Additive Multilevel Models." 25

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). "Scalable Inference for Logistic-Normal Topic Models." In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/285f89b802bcb2651801455c86d78f2a-Paper.pdf 2, 10, 17

Chow, D. D. K. (2022). "Schlömilch integrals and probability distributions on the simplex." URL https://arxiv.org/abs/2201.11013 10

Connor, R. J. and Mosimann, J. E. (1969). "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution." *Journal of the American Statistical Association*, 64(325): 194–206. URL http://www.jstor.org/stable/2283728 MR0240895. 10, 11

Creus-Martí, I., Moya, A., and Santonja, F.-J. (2021). "A Dirichlet Autoregressive Model for the Analysis of Microbiota Time-Series Data." *Complex.*, 2021: 9951817:1–9951817:16. URL https://api.semanticscholar.org/CorpusID:237715479 13

Efron, B. (2011). "Tweedie's Formula and Selection Bias." *Journal of the American Statistical Association*, 106(496): 1602–1614. PMID: 22505788. MR2896860. doi: https://doi.org/10.1198/jasa.2011.tm11181. 5

Filzmoser, P. (2023). *chemometrics: Multivariate Statistical Analysis in Chemometrics*. R package version 1.4.4. URL https://CRAN.R-project.org/package=chemometrics 23

Follett, L. and Yu, C. (2019). "Achieving parsimony in Bayesian vector autoregressions

with the horseshoe prior." *Econometrics and Statistics*, 11: 130–144. URL https://www.sciencedirect.com/science/article/pii/S2452306219300036 MR3980254. doi: https://doi.org/10.1016/j.ecosta.2018.12.004. 4

Frederic, P. and Lad, F. (2008). "Two Moments of the Logitnormal Distribution." *Communications in Statistics – Simulation and Computation*, 37(7): 1263–1269. MR2528273. doi: https://doi.org/10.1080/03610910801983178. 13

Gelman, A. (2006a). "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis*, 1. 6

Gelman, A. (2006b). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 9

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. URL https://books.google.de/books?id=ZXL6AQAAQBAJ MR3235677. 1

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press. MR2262986. doi: https://doi.org/10.1137/040607964. 1

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and Other Stories*. Analytical Methods for Social Research. Cambridge University Press. 5

Geyer, C. J. (1992). "Practical Markov Chain Monte Carlo." *Statistical Science*, 7(4): 473–483. doi: https://doi.org/10.1214/ss/1177011137 20

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). "Convergence rates of posterior distributions." *The Annals of Statistics*, 28(2): 500–531. MR1790007. doi: https://doi.org/10.1214/aos/1016218228. 2

Ghosh, J. and Ghattas, A. (2015). "Bayesian Variable Selection Under Collinearity." *The American Statistician*, 69. MR3391636. doi: https://doi.org/10.1080/00031305.2015.1031827. 24

Ghosh, S., Yao, J., and Doshi-Velez, F. (2019). "Model Selection in Bayesian Neural Networks via Horseshoe Priors." *Journal of Machine Learning Research*, 20(182): 1–46. URL http://jmlr.org/papers/v20/19-236.html MR4048993. 4

Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. URL https://books.google.de/books?id=qRuVoAEACAAJ MR3307991. 1, 2, 14

Good, I. J. (1962). "Theory of Probability Harold Jeffreys (Third edition, 447 + ix pp., Oxford Univ. Press, 84s.)." *Geophysical Journal International*, 6: 555–558. MR0187257. 9

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press. http://www.deeplearningbook.org. MR3617773. 12

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.21.1. URL https://mc-stan.org/rstanarm 9

Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., and Quinn, T. (2023). "Aitchison's Compositional Data Analysis 40 Years on: A Reappraisal." *Statistical Science*, 38(3): 386–410. MR4630375. doi: https://doi.org/10.1214/22-STS880. 2, 10, 12, 13

Griffin, J. E. and Brown, P. J. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188. MR2596440. doi: https://doi.org/10.1214/10-BA507. 2, 4, 18

Griffin, J. E. and Brown, P. J. (2013). "Some Priors for Sparse Regression Modelling." *Bayesian Analysis*, 8(3): 691–702. MR3102230. doi: https://doi.org/10.1214/13-BA827. 18, 24

Gupta, R. D. and Richards, D. S. P. (2001). "The History of the Dirichlet and Liouville Distributions." *International Statistical Review / Revue Internationale de Statistique*, 69(3): 433–446. URL http://www.jstor.org/stable/1403455 10

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. ISSN. CRC Press. URL https://books.google.de/books?id=f-A_CQAAQBAJ MR3616141. 2

Hintze, J. L. and Nelson, R. D. (1998). "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician*, 52(2): 181–184. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480559 21

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67. URL http://www.jstor.org/stable/1267351 MR0611894. doi: https://doi.org/10.1080/01966324.1981.10737061. 1

Hoffman, M. D. and Gelman, A. (2011). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." 18

Holmes, J. B. and Schofield, M. R. (2022). "Moments of the logit-normal distribution." *Communications in Statistics – Theory and Methods*, 51(3): 610–623. MR4368427. doi: https://doi.org/10.1080/03610926.2020.1752723. 13

Jeffrey, A., Zwillinger, D., Gradshteyn, I., and Ryzhik, I. (2007). "8–9 – Special Functions." In *Table of Integrals, Series, and Products (Seventh Edition)*, 859–1048. Boston: Academic Press, seventh edition. URL https://www.sciencedirect.com/science/article/pii/B9780080471112500169 MR2360010. 16

Johndrow, J., Orenstein, P., and Bhattacharya, A. (2020). "Scalable Approximate MCMC Algorithms for the Horseshoe Prior." *Journal of Machine Learning Research*, 21(73): 1–61. URL http://jmlr.org/papers/v21/19-536.html MR4095352. 2, 4

Kohns, D. and Szendrei, T. (2024). "Horseshoe prior Bayesian quantile regression." *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1): 193–220. MR4719324. doi: https://doi.org/10.1093/jrsssc/qlad091. 4

Kruijer, W., Rousseau, J., and Vaart, A. (2010). "Adaptive Bayesian Density Estimation with Location-Scale Mixtures." *Electronic Journal of Statistics*, 4. MR2735885. doi: https://doi.org/10.1214/10-EJS584. 8

Kruschke, J. K. (2015). "Chapter 6 – Inferring a Binomial Probability via Exact Mathematical Analysis." In Kruschke, J. K. (ed.), *Doing Bayesian Data Analysis (Second Edition)*, 123–141. Boston: Academic Press, second edition. URL https://www.sciencedirect.com/science/article/pii/B9780124058880000064 8

Kullback, S. and Leibler, R. A. (1951). "On Information and Sufficiency." *The Annals of Mathematical Statistics*, 22(1): 79–86. MR0039968. doi: https://doi.org/10.1214/aoms/1177729694. 16

Lin, J. (2016). "On The Dirichlet Distribution by Jiayu Lin." 7, 8, 10

McLatchie, Y., Rögnvaldsson, S., Weber, F., and Vehtari, A. (2023). "Robust and efficient projection predictive inference." MR4859120. doi: https://doi.org/10.1214/24-sts949. 25

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A. (2021). "Prior knowledge elicitation: The past, present, and future." 3

Mimno, D., Wallach, H., and McCallum, A. (2008). "Gibbs sampling for logistic normal topic models with graph-based priors." 17

Ongaro, A. and Migliorati, S. (2013). "A generalization of the Dirichlet distribution." *Journal of Multivariate Analysis*, 114: 412–426. URL https://www.sciencedirect.com/science/article/pii/S0047259X12001753 MR2993896. doi: https://doi.org/10.1016/j.jmva.2012.07.007. 10, 11

Osborne, B., Fearn, T., Miller, A. R., and Douglas, S. (1984). "Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs." *Journal of the Science of Food and Agriculture*, 35: 99–105. URL https://api.semanticscholar.org/CorpusID:94286061 23, 24

Pavone, F., Piironen, J., Bürkner, P.-C., and Vehtari, A. (2020). "Using reference models in variable selection." 5

Pawlowsky-Glahn, V. and Buccianti, A. (2011). "Compositional data analysis: theory and applications." URL https://api.semanticscholar.org/CorpusID:117956959 MR2920574. doi: https://doi.org/10.1002/9781119976462. 11, 12, 13

Piironen, J., Paasiniemi, M., and Vehtari, A. (2020). "Projective inference in high-dimensional problems: Prediction and feature selection." *Electronic Journal of Statistics*, 14(1): 2155–2197. MR4097052. doi: https://doi.org/10.1214/20-EJS1711. 5, 25

Piironen, J. and Vehtari, A. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018–5051. MR3738204. doi: https://doi.org/10.1214/17-EJS1337SI. 2, 4, 5, 25

Polson, N. G. and Scott, J. G. (2011). "Local shrinkage rules, Levy processes, and regularized regression." MR2899864. doi: https://doi.org/10.1111/j.1467-9868.2011.01015.x. 24

Polson, N. G. and Scott, J. G. (2012). "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Analysis*, 7(4): 887–902. MR3000018. doi: https://doi.org/10.1214/12-BA730. 6

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 5

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer New York. URL https://books.google.ch/books?id=6oQ4s8Pq9pYC MR2723361. 20

Ročková, V. (2018). "Bayesian estimation of sparse signals with a continuous spike-and-slab prior." *Annals of Statistics*, 46: 401–437. URL https://api.semanticscholar.org/CorpusID:85503428 MR3766957. doi: https://doi.org/10.1214/17-AOS1554. 2

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian shrinkage for high-dimensional regression." *Science China Mathematics*, 66: 409–442. URL https://api.semanticscholar.org/CorpusID:88516067 MR4535982. doi: https://doi.org/10.1007/s11425-020-1912-6. 5

Stan Development Team (2024). "Stan Modeling Language Users Guide and Reference Manual, Version 2.36." URL http://mc-stan.org/ 3, 18

Stein, C. M. (1981). "Estimation of the Mean of a Multivariate Normal Distribution." *The Annals of Statistics*, 9(6): 1135–1151. Publisher: Institute of Mathematical Statistics. URL https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-6/Estimation-of-the-Mean-of-a-Multivariate-Normal-Distribution/10.1214/aos/1176345632.full MR0630098. 4

Stryjewski, L. (2010). "40 years of boxplots." URL https://api.semanticscholar.org/CorpusID:36975036 21

Stuart, A. and Ord, K. (2009). *Kendall's Advanced Theory of Statistics: Volume 1: Distribution Theory*. Number vol. 1 ;vol. 1994 in Kendall's Advanced Theory of Statistics. Wiley. URL https://books.google.ch/books?id=tW18thQWJQIC MR1285356. 14

Szakács, G., Annereau, J.-P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K. J., Reinhold, W., Guo, Y., Kruh, G. D., Reimers, M., Weinstein, J. N., and Gottesman, M. M. (2004). "Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells." *Cancer Cell*, 6(2): 129–137. URL https://www.sciencedirect.com/science/article/pii/S1535610804002065 24

Tadesse, M. and Vannucci, M. (2021). *Handbook of Bayesian Variable Selection*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press. URL https://books.google.de/books?id=Cn1TEAAAQBAJ 2, 4, 25

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. MR1379242. 1, 2

Tosh, C., Greengard, P., Goodrich, B., Gelman, A., Vehtari, A., and Hsu, D. (2022). "The piranha problem: Large effects swimming in a small pond." 17

Van der Pas, S. (2021). "Theoretical guarantees for the horseshoe and other global-local shrinkage priors." In *Handbook of Bayesian Variable Selection*, 133–160. Chapman and Hall/CRC. 1, 4, 5

Van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10(1): 976–1000. MR3486423. doi: https://doi.org/10.1214/16-EJS1130. 1, 2, 4, 6

Van der Pas, S., Szabó, B., and van der Vaart, A. (2017). "Uncertainty Quantification for the Horseshoe (with Discussion)." *Bayesian Analysis*, 12(4): 1221–1274. MR3724985. doi: https://doi.org/10.1214/17-BA1065. 2, 5

Van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2014). "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal of Statistics*, 8(2): 2585–2618. MR3285877. doi: https://doi.org/10.1214/14-EJS962. 2

Van Erp, S., Oberski, D., and Mulder, J. (2019). "Shrinkage priors for Bayesian penalized regression." *Journal of Mathematical Psychology*, 89: 31–50. MR3903921. doi: https://doi.org/10.1016/j.jmp.2018.12.004. 4, 5

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2023). "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models." R package version 2.6.0. URL https://mc-stan.org/loo/ 24

Vehtari, A., Gelman, A., and Gabry, J. (2016). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432. MR3647105. doi: https://doi.org/10.1007/s11222-016-9696-4. 20, 24

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)." *Bayesian Analysis*, 16(2): 667–718. MR4298989. doi: https://doi.org/10.1214/20-BA1221. 20

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6(none): 142–228. MR3011074. doi: https://doi.org/10.1214/12-SS102. 20

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2022). "Pareto Smoothed Importance Sampling." MR4749108. 20

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cam-

bridge University Press. URL https://books.google.de/books?id=J-VjswEACAAJ MR3837109. doi: https://doi.org/10.1017/9781108231596. 14

Wang, Y. and Polson, N. G. (2024). "Pochhammer Priors for Sparse Count Models." URL https://arxiv.org/abs/2402.09583 10

West, M. (1987). "On scale mixtures of normal distributions." *Biometrika*, 74(3): 646–648. MR0909372. doi: https://doi.org/10.1093/biomet/74.3.646. 4, 9

Wong, T.-T. (1998). "Generalized Dirichlet distribution in Bayesian analysis." *Applied Mathematics and Computation*, 97(2): 165–181. URL https://www.sciencedirect.com/science/article/pii/S0096300397101400 MR1643091. doi: https://doi.org/10.1016/S0096-3003(97)10140-0. 11

Xun, G., Li, Y., Zhao, W. X., Gao, J., and Zhang, A. (2017). "A Correlated Topic Model Using Word Embeddings." In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 4207–4213. doi: https://doi.org/10.24963/ijcai.2017/588 17

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). "On the Computational Complexity of High-Dimensional Bayesian Variable Selection." *The Annals of Statistics*, 44(6): 2497–2532. URL http://www.jstor.org/stable/44245760 MR3576552. doi: https://doi.org/10.1214/15-AOS1417. 2

Zhang, Y. and Bondell, H. D. (2018). "Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors." *Bayesian Analysis*, 13(3): 823–844. doi: https://doi.org/10.1214/17-BA1076 2, 5, 7, 25

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior." *Journal of the American Statistical Association*, 0(0): 1–13. doi: https://doi.org/10.1080/01621459.2020.1825449 2, 3, 4, 7, 9, 13, 18, 24, 25