Generalised Decomposition Priors on R2

DAGStat 2025

Javier Aguilar, Paul-Christian Bürkner

25.03.2025

TU Dortmund
javier.aguilar@tu-dortmund.de

Prior specification

The prior is a fundamental part of Bayesian modeling.

Hard question: How to specify the prior?

Prior specification

The prior is a fundamental part of Bayesian modeling.

Hard question: How to specify the prior?

Fundamental question: How to make it easier for the user to specify priors?

Prior specification

The prior is a fundamental part of Bayesian modeling.

Hard question: How to specify the prior?

Fundamental question: How to make it easier for the user to specify priors?

- Build upon **intuition** about the phenomenon.
- Propose **semi automated** and **parsimonious** priors.
- Propose priors on **predictive quantities** that are better understood by the user and move uncertainty.

Priors in Linear Regression

How to select the prior for b?

1. **Scaled Gaussians**

$$b \mid \sigma, \Sigma_b \sim N(0, \sigma^2 \Sigma_b), \quad \sigma \sim p(\sigma), \quad \Sigma_b \sim p(\Sigma_b)$$

Priors in Linear Regression

How to select the prior for b?

1. **Scaled Gaussians**

$$b \mid \sigma, \Sigma_b \sim N(0, \sigma^2 \Sigma_b), \quad \sigma \sim p(\sigma), \quad \Sigma_b \sim p(\Sigma_b)$$

2. **Shrinkage priors**

$$b_i \mid \Psi_i \sim N(0, \Psi_i), \quad \Psi_i \sim G(\cdot)$$
$$p(b_i) = \int N(b_i | 0, \Psi_i) dG(\Psi_i)$$

- Usual decomposition $\Psi_i = \phi_i \tau^2$
- Includes Spike and Slab, Horseshoe, Dirichlet Laplace, Beta Prime etc.

Prior specification

Question: What is the effect on R^2?

$$b_k \sim \mathcal{N}(0, 1), \quad \sigma \sim \text{Exp}(1)$$

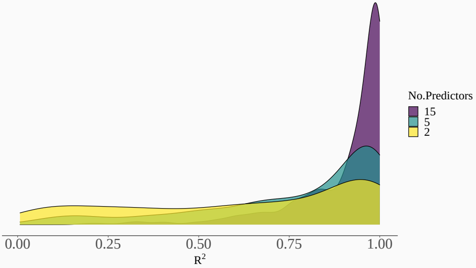


Figure 1: Implied prior distribution on R^2

The GDR2 prior

Basic Idea:

- Set a prior on R^2 to encode domain knowledge.
- Decompose total variance τ^2 among b_k
- Span and jointly regularize all coefficients!

$$R^2 = \frac{\text{var}(x'b)}{\text{var}(x'b) + \sigma^2}$$

The GDR2 prior: Construction

- Assumptions:
  - Centered coefficients
    $\mathbb{E}(b_i) = \mathbb{E}(u_{igj}) = 0,$
  - Scaled variances
    $\mathrm{var}(b_i) = \sigma^2 \lambda_i^2$

The GDR2 prior: Construction

- Assumptions:

  - Centered coefficients
    $\mathbb{E}(b_i) = \mathbb{E}(u_{igj}) = 0,$
  - Scaled variances
    $\text{var}(b_i) = \sigma^2 \lambda_i^2$

- Variance decomposition of the linear predictor $x'b$

$$\text{var}(x'b) = \sum_{i=1}^{K} \sigma^2 \lambda_i^2$$

The GDR2 prior: Construction

- Assumptions:
  - Centered coefficients $\mathbb{E}(b_i) = \mathbb{E}(u_{igj}) = 0,$
  - Scaled variances $\text{var}(b_i) = \sigma^2 \lambda_i^2$
- Variance decomposition of the linear predictor $x'b$

$$\text{var}(x'b) = \sum_{i=1}^{K} \sigma^2 \lambda_i^2$$

- Total explained variance τ^2

$$\tau^2 := \sum_{i=1}^{K} \lambda_i^2$$

The GDR2 prior: Construction

- Assumptions:
  - Centered coefficients $\mathbb{E}(b_i) = \mathbb{E}(u_{igj}) = 0$,
  - Scaled variances $\text{var}(b_i) = \sigma^2 \lambda_i^2$
- Variance decomposition of the linear predictor $x'b$

$$\text{var}(x'b) = \sum_{i=1}^{K} \sigma^2 \lambda_i^2$$

- Total explained variance τ^2

$$\tau^2 := \sum_{i=1}^{K} \lambda_i^2$$

- Rewrite R^2

$$R^2 = \frac{\sigma^2 \tau^2}{\sigma^2 \tau^2 + \sigma^2} = \frac{\tau^2}{\tau^2 + 1}.$$

The GDR2 prior: R2 prior

1) Set

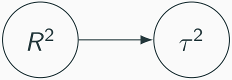
$$R^2 \sim \text{Beta}(\mu_{R^2}, \varphi_{R^2}) \iff \tau^2 \sim \text{BP}(\mu_{R^2}, \varphi_{R^2})$$

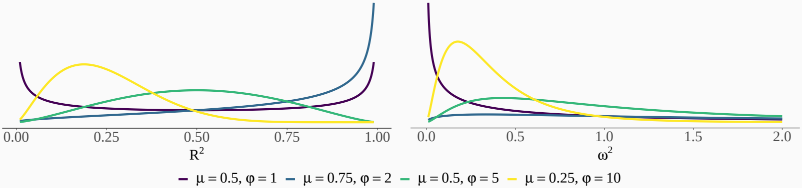
- $\mu_{R^2} \in (0, 1)$ **prior mean**
- $\varphi_{R^2} > 0$ **prior precision**

1) Set

$$R^2 \sim \text{Beta}(\mu_{R^2}, \varphi_{R^2}) \iff \tau^2 \sim \text{BP}(\mu_{R^2}, \varphi_{R^2})$$

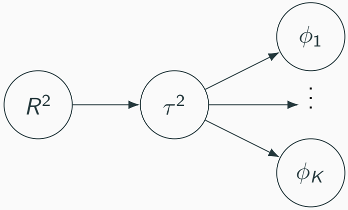
- $\mu_{R^2} \in (0, 1)$ **prior mean**
- $\varphi_{R^2} > 0$ **prior precision**





The GDR2 prior: Variance Partitioning

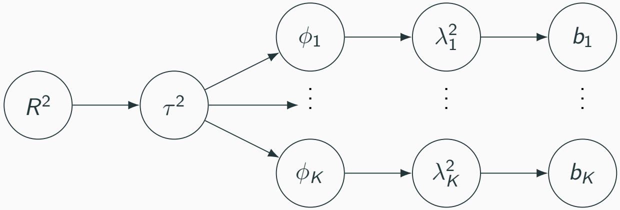
Let ϕ follow a distribution on the simplex $\mathcal{S}^{K-1}$



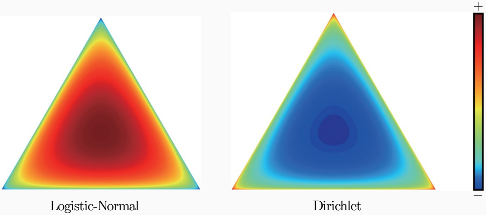
- How should we specify the distribution of ϕ?
- Which behavior do we want to exhibit in ϕ?

The GDR2 prior: Coefficients

Set $\lambda_i = \phi_i \tau^2$ and $b \mid \sigma, \lambda \sim N(0, \sigma \lambda_i^2)$



Distributions in the Simplex



Dirichlet $\phi \sim \text{Dir}(\alpha)$

- Tractable analytical properties
- α is easy to understand, but the mean determines covariance.

Logistic Normal

$\eta \sim \mathcal{N}(\mu, \Sigma), \quad \phi = \text{softmax}(\eta)$

- Higher flexibility
- Challenging to select μ, Σ

Hyperparameter specification

Prior for R^2

- User informed
- Since $\text{var}(b) = \mathbb{E}(\tau^2)\text{cov}(\phi)$, set values to imply a heavy tail for b. Set $(1 - \mu_{R^2})\varphi_{R^2} \leq 1/2$

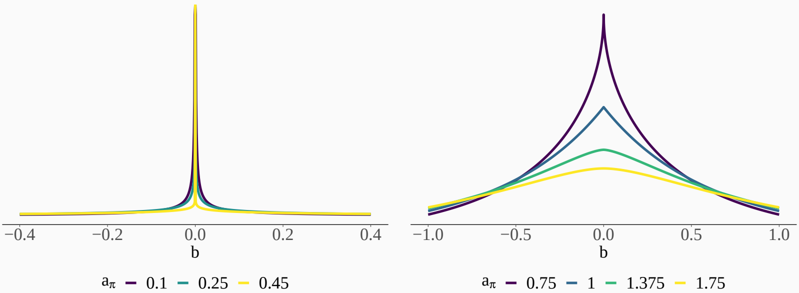
Hyperparameter specification

Prior for R^2

- User informed
- Since $\text{var}(b) = \mathbb{E}(\tau^2)\text{cov}(\phi)$, set values to imply a heavy tail for b. Set $(1 - \mu_{R^2})\varphi_{R^2} \leq 1/2$

Priors for ϕ: Dirichlet distribution

- Set $\alpha = (a_\pi, ..., a_\pi), a_\pi > 0$
- If $a_\pi \leq 1/2$ then we get unbounded marginals for b around the origin

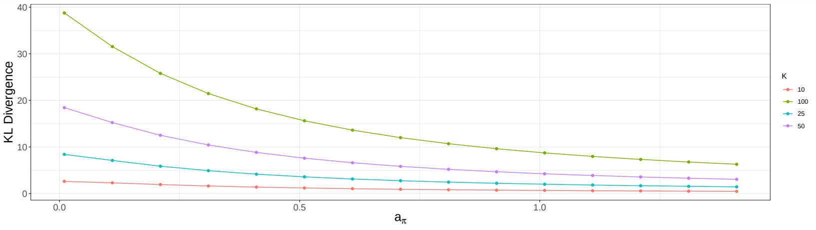


Hyperparameter specification: LN and KL Matching

Priors for ϕ: Logistic Normal distribution

Idea: Minimize KL between Dir(α) and $LN(\mu, \Sigma)$.

- Closed form expression.
- Automated and cheap
- Exact matching is neither wished nor achievable as KL doesnt vanish.



Hyperparameter specification: LN

Priors for ϕ: Logistic Normal distribution

How to specify μ?

- $\mu = 0$ means all proportions are equally weighted and $\mathbb{E}[\phi_k] = 1/K$.
- If $\mu_k = c_k, \Sigma = \sigma_\phi^2 I$ then

$$\mathbb{E}[\phi_k] = \mathbb{E}\left(\frac{e^{\eta_k}}{\sum_j e^{\eta_j}}\right) \approx \frac{e^{c_k}}{\sum_j e^{c_j}}$$

- One can also form groups within μ
- Other cases are more involved

Hyperparameter specification: LN and KL Matching

Priors for ϕ: Logistic Normal distribution

Idea: How to specify Σ? Study the implied prior on the size of the logits η (log ratios)

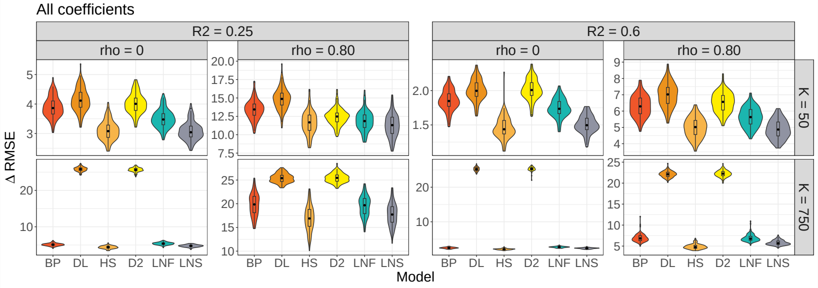
- If $\mu = 0, \Sigma = \sigma_\phi^2 I$ and $\eta \sim N(0, \sigma_\phi^2 I)$ then $\|\eta\|$ concentrates around $\sqrt{K}\sigma_\phi$ since

$$\mathbb{P}\left(\left|\frac{\|\eta\|}{\sqrt{K}\sigma_\phi} - 1\right| \geq t\right) \leq 2\exp\left(-\frac{K\sigma_\phi^2 t^2}{2C}\right),\ C > 0$$

- There σ_ϕ specifies a budget.
- $\sigma_\phi \to 0$, the logits concentrate near zero, resulting in $\phi_k \to 1/K$
- σ_ϕ increases, the logits spread
- Example: If $\sigma_\phi = \sqrt{\gamma/K}, \gamma > 0$ then $\|\eta\|^2 \approx \gamma$ regardless of K
- Other cases are more involved

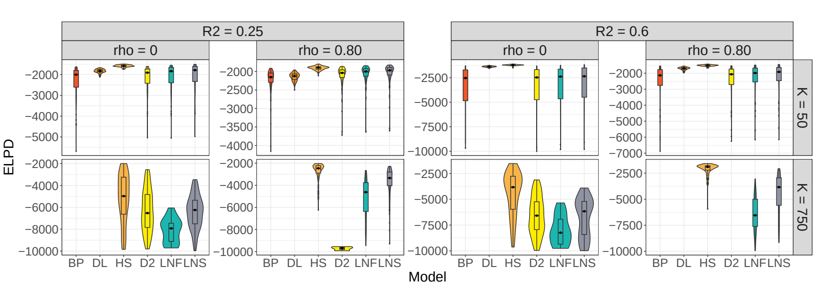
Simulation: Comparison with other priors

Parameter recovery



Simulation: Comparison with other priors

Out of sample predictive performance measured via ELPD



Future directions

1. How can the data inform us about the mean and covariance of ϕ?
2. Theoretical properties of the prior
3. Even with KL matching we are getting promising results, hence proper hyperparameter selection should improve performance.

Takeaway Message

Key Takeaway:

- Opens the way to think about different relationships among variance components.
- Joint priors based on quantities of interest is a promising avenue for research.

Thank You for Your Attention!

Feel free to reach out for questions or collaborations.

Contact Information:

- javier.aguilar@tu-dortmund.de
- `https://jear2412.github.io`
- jear2412

References

[1] Javier Enrique Aguilar and Paul-Christian Bürkner. Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior. *Electronic Journal of Statistics*, 17(1):1711 – 1767, 2023. doi: 10.1214/23-EJS2136. URL https://doi.org/10.1214/23-EJS2136.

[2] Javier Enrique Aguilar and Paul-Christian Bürkner. Generalized decomposition priors on r2, 2025. URL https://arxiv.org/abs/2401.10180.