Bayesian Analysis (2019) **TBA**, Number TBA, pp. 1–29

Intuitive Joint Priors for Variance Parameters

Geir-Arne Fuglstad[*], Ingeborg Gullikstad Hem[†], Alexander Knight[‡], Håvard Rue[§], and Andrea Riebler[¶]

Abstract. Variance parameters in additive models are typically assigned independent priors that do not account for model structure. We present a new framework for prior selection based on a hierarchical decomposition of the total variance along a tree structure to the individual model components. For each split in the tree, an analyst may be ignorant or have a sound intuition on how to attribute variance to the branches. In the former case a Dirichlet prior is appropriate to use, while in the latter case a penalised complexity (PC) prior provides robust shrinkage. A bottom-up combination of the conditional priors results in a proper joint prior. We suggest default values for the hyperparameters and offer intuitive statements for eliciting the hyperparameters based on expert knowledge. The prior framework is applicable for `R` packages for Bayesian inference such as `INLA` and `RStan`.

Three simulation studies show that, in terms of the application-specific measures of interest, PC priors improve inference over Dirichlet priors when used to penalise different levels of complexity in splits. However, when expressing ignorance in a split, Dirichlet priors perform equally well and are preferred for their simplicity. We find that assigning current state-of-the-art default priors for each variance parameter individually is less transparent and does not perform better than using the proposed joint priors. We demonstrate practical use of the new framework by analysing spatial heterogeneity in neonatal mortality in Kenya in 2010–2014 based on complex survey data.

Keywords: additive models, hierarchical variance decomposition, latent Gaussian models, penalised complexity, joint prior distributions, variance parameters.

1 Introduction

Bayesian hierachical models (BHMs) are ubiquitous in science due to their flexibility and interpretability (Gelman and Hill, 2007; Gelman et al., 2013; Banerjee et al., 2014). In this paper, we consider BHMs where the latent level consists of an additive combination of model components that are classified as fixed effects and random effects. This subclass covers a range of common model classes such as generalised linear mixed

[*]Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz' vei 1, 7034 Trondheim, Norway. Corresponding author: geir-arne.fuglstad@ntnu.no

[†]Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz' vei 1, 7034 Trondheim, Norway, ingeborg.hem@ntnu.no

[‡]Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz' vei 1, 7034 Trondheim, Norway, alexander.knight@ntnu.no

[§]CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia, haavard.rue@kaust.edu.sa

[¶]Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz' vei 1, 7034 Trondheim, Norway, andrea.riebler@ntnu.no

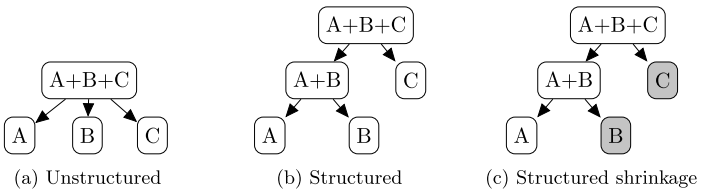
© 2019 International Society for Bayesian Analysis https://doi.org/10.1214/19-BA1185

models (GLMMs) and generalised additive mixed models (GAMMs) (Fahrmeir and Lang, 2001). In additive models, the total latent variance of the sum of the random effects decomposes into the sum of the variance contributed by each random effect, and each random effect has a variance parameter that controls its *a priori* contribution. We present a general framework for constructing joint priors for these variance parameters for BHMs, and suggest robust shrinkage priors for the reduced class of latent Gaussian models (LGMs) where the model components are Gaussian conditional on the model parameters (Rue et al., 2009, 2017; Bakka et al., 2018; Krainski et al., 2018).

There is no concensus on priors for variance parameters in BHMs (Lambert et al., 2005; Gelman, 2006; Gelman et al., 2017). The default prior in the `R` package `INLA` (Lindgren and Rue, 2015) is an inverse-gamma distribution InvGamma$(1, 5 \cdot 10^{-5})$ (Blangiardo and Cameletti, 2015), and the `R` package `RStan` (Carpenter et al., 2017; Stan Development Team, 2018a) has implicit priors that are uniform on the range of legal values for the parameters (Stan Development Team, 2018b). WinBUGS, OpenBUGS and JAGS used InvGamma$(0.001, 0.001)$ distributions in their examples (Spiegelhalter et al., 1996; Plummer, 2017), and the Stata manual employs InvGamma$(0.01, 0.01)$ priors (StataCorp, 2017). Conjugacy provides InvGamma(ϵ, ϵ) distributions with computational advantages, but their use may result in severe problems (Gelman, 2006) and they are generally inappropriate for variances of random effects (Lunn et al., 2009). Gelman (2006) proposed heavier tails through Half-Cauchy(25) distributions on the standard deviations, and others have investigated bounded uniform densities on the variances or the logarithms of the variances (Lambert et al., 2005) and bounded uniform priors on the standard deviations (Martinez-Beneito, 2013). Recently, Simpson et al. (2017) proposed a principle-based, robust prior termed penalised complexity (PC) prior that offers shrinkage towards zero variance. In the case of LGMs, the PC prior is an exponential distribution on the standard deviation.

However, general-purpose priors may not be suitable for a given application (Gelman et al., 2017) and independent priors for each random effect cannot exploit the structure of the model (Simpson et al., 2017, Section 7). For example, in disease mapping, prior elicitation is more meaningful for the total variance of the random effects than their separate variances (Wakefield, 2006), and, for animal models in genetic settings, the proportion of variability in a phenotypic trait being accounted for by genes is important (Holand et al., 2013). Further, the intraclass correlation (ICC) (McGraw and Wong, 1996) in a random intercept model is linked to a generalised version of the coefficient of determination (Gelman and Hill, 2007), also known as R^2, which expresses the proportion of the total variance explained by the model components. However, putting a prior on R^2 requires a joint prior on the two variance parameters in the random intercept model. Additionally, in the context of regression, Som et al. (2014) discuss block g-priors where regression coefficients are partitioned and shrinkage is applied to the R^2 of each partition.

Consider a simple multilevel model with responses $y_{i,j,k}|\eta_{i,j,k} \sim \text{Poisson}(\exp(\eta_{i,j,k}))$, where $\eta_{i,j,k} = a_i + b_{i,j} + c_{i,j,k}$ for experiment k on individual j in group i. We will term the group effect, individual effect and measurement effect for A, B, and C, respectively, and write the latent model as A+B+C for short hand. The total latent variance t of A+B+C decomposes as $t = \sigma^2_{\text{A}} + \sigma^2_{\text{B}} + \sigma^2_{\text{C}}$, where σ^2_{A}, σ^2_{B} and σ^2_{C} are the variances of A, B and C,



(a) Unstructured (b) Structured (c) Structured shrinkage

Figure 1: Hierarchical model decomposition. Gray boxes indicate preferred branches.

respectively. This standard parametrization facilitates independent priors on the variances and can be used to achieve the desired *a priori* marginal properties for the random effects. However, it is difficult to encode *a priori* knowledge on joint properties such as the size of t or preference for A over B or A+B over C in a transparent and intuitive way.

An obvious alternative is to parametrize the variance parameters as t and the proportion of t assigned to each random effect $(\omega_A, \omega_B, \omega_C)$, where $0 \leq \omega_A, \omega_B, \omega_C \leq 1$ and $\omega_A + \omega_B + \omega_C = 1$. This is illustrated in Figure 1a by splitting A+B+C into the models A, B and C. This parametrization is suitable for expressing ignorance about how the variance should be attributed to the random effects. A simple way to assign the joint prior is to set $(\omega_A, \omega_B, \omega_C) \sim \text{Dir}(a, a, a)$, $a > 0$, where Dir denotes the Dirichlet distribution (Balakrishnan and Nevzorov, 2003). This prior has no preference for one of the random effects over the other and is invariant to the ordering of the random effects, and we can select $a > 0$ to make the prior suitably vague. Together with the conditional prior $\pi(t|\omega_A, \omega_B, \omega_C)$, this implicitly defines a proper joint prior for $(\sigma^2_A, \sigma^2_B, \sigma^2_C)$ that is invariant to permutations in the order of the random effects, but can incorporate prior knowledge on t. This has a similar flavor as the Dirichlet-Laplace prior by Bhattacharya et al. (2015), which is a global-local shrinkage prior (Polson and Scott, 2010) that induces sparsity in regression. However, in this paper we will focus on random effects and not fixed effects.

The simple split strategy is not always suitable and Riebler et al. (2016) demonstrated that for the BYM (Besag, York and Mollié) model, which is a sum of a Besag random effect and an unstructured random effect, a PC prior that penalises the added complexity of the structured effect relative to the unstructured effect improves inference. For A+B+C, fewer levels of hierarchy may be preferred so that B is preferred to A and C is preferred over A+B. This knowledge about relative complexity of the random effects can be incorporated by splitting A+B+C hierarchically as shown in Figure 1b. Here we first split A+B+C into A+B and C through $\omega_1 = (\sigma^2_A + \sigma^2_B)/t$, and then split A+B into A and B through $\omega_2 = \sigma^2_A/(\sigma^2_A + \sigma^2_B)$, where $0 \leq \omega_1, \omega_2 \leq 1$. The joint prior for $(\sigma^2_A, \sigma^2_B, \sigma^2_C)$ is then constructed by first selecting $\pi(\omega_2)$, then $\pi(\omega_1|\omega_2)$, and finally $\pi(t|\omega_1, \omega_2)$. Priors inducing shrinkage towards $\omega_2 = 0$ and $\omega_1 = 0$ can be chosen in the lower and upper split, respectively. The shrinkage can be illustrated graphically as shown in Figure 1c. For LGMs, PC priors offer a robust choice, but the framework is general and other priors can be selected by the analyst. For example, if shrinkage is only required at the top level, a Dirichlet prior for $(\omega_2, 1 - \omega_2)$ could be combined with a shrinkage prior for $\omega_1|\omega_2$.

The ideas generalize to more random effects through the selection of a hierarchical decomposition of the model in the form of a tree, and the selection of a conditional distribution for the attribution of the total variance to the branches for each split. The joint prior is calculated in a bottom-up approach using these conditional distributions. We suggest default values for the hyperparameters of the Dirichlet distribution based on the marginal prior distributions for the proportions of variance assigned to each branch of the split. This ensures that the default setting for the prior is well-behaved as the number of branches in a split increases. Default values for the PC priors can be selected based on moderate shrinkage of the proportion of variance. Additionally, we discuss how to include expert knowledge through interpretable statements on the total variance and the distribution of variance in the tree. The joint prior can contain a mix of expert knowledge and default values that provide a weakly informative prior (Gelman et al., 2008; Simpson et al., 2017). This means the prior framework with joint priors is appropriate for default priors for software packages such as `INLA` and `RStan`.

The properties of the proposed priors are compared to the properties of default priors from software and vague priors from literature. This is a fair comparison since even though the new priors account for model structure, they do not incorporate strong expert knowledge and are suggested to be used in a default way in Bayesian software. The comparison is performed through three simulation studies: a simple random intercept model with Gaussian responses, a latin square experiment with Gaussian responses, and a spatial model with Binomial responses. To ease the presentation of the comparisons and not overload the reader with results, we choose a set of targets for each simulation study and compare the posteriors resulting from the different prior choices with respect to the targets. Additional results are provided in the Supplementary Materials (Fuglstad et al., 2019a). Furthermore, we provide example code in the Supplementary Materials for producing results for different priors for the latin square model in Section 5.2. The code is described in Section S4.3 in the Supplementary Materials.

We start by introducing the general framework in Section 2, then we introduce LGMs and suitable priors for developing a new class of priors for LGMs in Section 3. The new class of priors for LGMs is introduced in Section 4 and is applied to simulation studies with Gaussian responses in Section 5. In Section 6 we present one simulation study with Binomial response and explain how the approach can be used in practice. The paper ends with a discussion in Section 7.

2 Tree-based hierarchical variance decomposition

In this section we cover basic notation, and formally introduce additive models, hierarchical variance decomposition, and the new framework for joint priors for variances.

2.1 Additive models

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a vector of $n > 0$ observations. We model the expected values $\mathrm{E}(y_i) = g^{-1}(\eta_i)$, $i = 1, \ldots, n$, through a vector of linear predictors $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ and a link function $g : \mathbb{R} \to \mathbb{R}$. We consider models where the likelihood has parameters $\boldsymbol{\theta}_{\mathrm{L}}$

and factors as $\pi(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\theta}_{\mathrm{L}}) = \prod_{i=1}^{n} \pi(y_i|\eta_i, \boldsymbol{\theta}_{\mathrm{L}})$. This covers models such as GLMMs and GAMMs. We term $\boldsymbol{\eta}$ and its description as the latent part of the model.

We assume that the linear predictor is described as

$$\eta_i = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{N} u_{j,k_j[i]}, \quad i = 1, \ldots, n, \tag{2.1}$$

where β_0 is the intercept, $\boldsymbol{x}_i$ is the vector of covariates associated with observation i, $\boldsymbol{\beta}$ is a vector of coefficients, and $\boldsymbol{u}_j = (u_1, \ldots, u_{m_j})$ is a random vector and $k_j[i]$ is the associated element of $\boldsymbol{u}_j$ for observation i for $j = 1, \ldots, N$. The two first terms will be called fixed effects and the last N terms will be called random effects. To focus on the joint prior for variance parameters, we will assume that each random effect $\boldsymbol{u}_j$ has a single model parameter, which is a variance σ_j^2. In general, the random effects may have other parameters such as correlation parameters and we discuss how to handle this in Section 7.

We denote the vector of model parameters by $\boldsymbol{\theta}_{\mathrm{M}} = (\sigma_1^2, \ldots, \sigma_N^2)$. The BHM is completed by specifying the latent model through $\pi(\boldsymbol{u}_j|\sigma_j^2)$ for $j = 1, \ldots, N$, and the prior $\pi(\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_{\mathrm{L}}, \boldsymbol{\theta}_{\mathrm{M}})$. We follow common practice so that the prior satisfies $\pi(\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_{\mathrm{L}}, \boldsymbol{\theta}_{\mathrm{M}}) = \pi(\beta_0)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\theta}_{\mathrm{L}})\pi(\boldsymbol{\theta}_{\mathrm{M}})$. The major improvement over common practice is that we will develop a framework for selecting intuitive joint priors for the variance parameters that does not require that $\pi(\boldsymbol{\theta}_{\mathrm{M}}) = \prod_{j=1}^{N} \pi(\sigma_j^2)$.

2.2 Hierarchical variance decomposition

The additivity in (2.1) causes the total latent variance $\mathrm{Var}[\eta_i|\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_{\mathrm{M}}]$ of linear predictor i to decompose as the variance contributed by each random effect $\mathrm{Var}[u_{k_j[i]}|\beta_0, \boldsymbol{\beta}, \sigma_j^2]$, $j = 1, \ldots, N$, for $i = 1, \ldots, n$. If random effect j is homogeneous, the variance parameter of random effect j will be a marginal variance in the sense that $\mathrm{Var}[u_{k_j[i]}|\beta_0, \boldsymbol{\beta}, \sigma_j^2] = \sigma_j^2$ for $i = 1, \ldots, n$. If all random effects are homogeneous, the total latent variance of the linear predictors is homogeneous, $t = \mathrm{Var}[\eta_1|\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_{\mathrm{M}}] = \cdots = \mathrm{Var}[\eta_n|\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_{\mathrm{M}}] = \sigma_1^2 + \ldots + \sigma_N^2$. If random effect j is heterogenous so that $\mathrm{Var}[u_{k_j[i]}|\beta_0, \boldsymbol{\beta}, \sigma_j^2]$ varies for different values of i, the variance parameter σ_j^2 is selected to be comparable to a marginal variance; see the discussion in Section 3.1. We term the parameter $t = \sigma_1^2 + \ldots + \sigma_N^2$ the total latent variance.

We describe the attribution of t to the individual random effects through a tree $\mathcal{T}$. The construction of $\mathcal{T}$ starts with a root node $T_0 = \{1, \ldots, N\}$ that contains all the random effects, and in the first step we introduce $K_1 > 1$ child nodes $T_1, \ldots, T_{K_1}$ that partition T_0 into $T_0 = T_1 \cup \cdots \cup T_{K_1}$. We continue this recursively for each child node until all leaf nodes are singletons. This results in a tree $\mathcal{T}$ with S splits where there are K_s child nodes for split $s = 1, \ldots, S$. We have $S \leq N - 1$, where $S = 1$ is achieved by directly splitting the root node to singletons as in Figure 1a and the maximum value is achieved by only using dual splits such as in Figure 1b.

For each split s, the parent node P_s is split into K_s child nodes $C_1, \ldots, C_{K_s}$ and we will define a vector of parameters $\boldsymbol{\omega}_s = (\omega_{s,1}, \ldots, \omega_{s,K_s})$, $s = 1, \ldots, S$. The child nodes

describe a partitioning of the random effects in the parent node, and we let $\boldsymbol{\omega}_s$ describe the proportion of the total variance in the parent node, $\sum_{j\in P_s}\sigma_j^2$, that is assigned to each child node through

$$\boldsymbol{\omega}_s = \frac{1}{\sum_{j\in P_s}\sigma_j^2}\left(\sum_{j\in C_1}\sigma_j^2, \ldots, \sum_{j\in C_{K_s}}\sigma_j^2\right), \quad s = 1, \ldots, S.$$

We denote the $K-1$ simplex by $\Delta^K = \{(x_1, \ldots, x_K) | \sum_{k=1}^K x_k = 1, x_k \geq 0 \ \forall k\}$ so that the restrictions are $\boldsymbol{\omega}_s \in \Delta^{K_s}$ for $s = 1, \ldots, S$. This means that the parameters ω_{s,K_s} are superfluous for $s = 1, \ldots, S$, but we keep them for ease of notation and interpretability.

For any split $s = 1, \ldots, S$, we term a child node and its decendants as a branch of the split. The description of the model structure through a tree structure defines a reparametrization of $(\sigma_1^2, \ldots, \sigma_N^2)$ to $(t, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_S)$, where S is the number of splits in the tree. The examples discussed in the introduction can be rephrased in this terminology, and demonstrate that there is no unique selection of the tree.

Example 1 (Tree structure)**.** Consider three random effects A, B and C with marginal variances $(\sigma_\mathrm{A}^2, \sigma_\mathrm{B}^2, \sigma_\mathrm{C}^2)$. Let the root node be $T_0 = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$.

Figure 1a, describes the case that the root node is partitioned into three children $T_1 = \{\mathrm{A}\}$, $T_2 = \{\mathrm{B}\}$ and $T_3 = \{\mathrm{C}\}$. This leads to a reparametrization $(t, \boldsymbol{\omega})$, where $t = \sigma_\mathrm{A}^2 + \sigma_\mathrm{B}^2 + \sigma_\mathrm{C}^2$ and $\boldsymbol{\omega} = (\sigma_\mathrm{A}^2, \sigma_\mathrm{B}^2, \sigma_\mathrm{C}^2)/t$.

Figure 1b shows the case that T_0 is first partitioned into $T_1 = \{\mathrm{A}, \mathrm{B}\}$ and $T_2 = \{\mathrm{C}\}$, and then T_1 is partitioned into $T_3 = \{\mathrm{A}\}$ and $T_4 = \{\mathrm{B}\}$. This results in a reparametrization $(t, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$, where $t = \sigma_\mathrm{A}^2 + \sigma_\mathrm{B}^2 + \sigma_\mathrm{C}^2$, $\boldsymbol{\omega}_1 = (\sigma_\mathrm{A}^2 + \sigma_\mathrm{B}^2, \sigma_\mathrm{C}^2)/t$ and $\boldsymbol{\omega}_2 = (\sigma_\mathrm{A}^2, \sigma_\mathrm{B}^2)/(\sigma_\mathrm{A}^2 + \sigma_\mathrm{B}^2)$. △

2.3 Hierachical decomposition priors

The tree-based hierarchical variance decomposition facilitates the construction of joint priors that include prior belief about the relative sizes of groups of random effects. The tree structure must be selected so that the desired comparisons can be made. Trees such as shown in Figure 1a are useful for expressing ignorance about the attribution of variance to the random effects, whereas trees such as shown in Figure 1b are useful for imposing shrinkage to one of the branches in each dual split. Generally, a tree may consist of a mixture of splits where the analyst wants to be informative and splits where the analyst wants to express ignorance.

We propose to construct a joint prior for the marginal variance parameters in a bottom-up approach where the prior for a given split only depends on descendant nodes of the parent node.

Assumption 1 (Bottom-up approach)**.** *For a tree structure with S splits, $\pi(\{\boldsymbol{\omega}_s\}_{s=1}^S) = \prod_{s=1}^S \pi(\boldsymbol{\omega}_s | \{\boldsymbol{\omega}_j\}_{j\in D(s)})$, where $D(s)$ is the set of descendant splits for split $s = 1, \ldots, S$.*

This means that the joint prior for the decomposition uses a directed acyclic graph so that parameters that belong to subsplits in different branches of a split are marginally independent. We combine the prior for the decomposition of the variance with a conditional prior on the total variance of the random effects to form what we will call *hierarchical decomposition* (HD) priors.

Definition 1 (Hierarchical decomposition (HD) priors)**.** Consider a BHM with an additive latent structure with N random effects with marginal variance parameters $\sigma_1^2, \ldots, \sigma_N^2$. Assume that the model structure is described by a tree that recursively partitions the set of random effects into singletons. Then a hierarchical decomposition (HD) prior is given by

$$\pi(\sigma_1^2, \ldots, \sigma_N^2) = \pi(t|\{\boldsymbol{\omega}_s\}_{s=1}^S) \prod_{s=1}^S \pi(\boldsymbol{\omega}_s|\{\boldsymbol{\omega}_j\}_{j\in D(s)}),$$

where $t = \sigma_1^2 + \ldots + \sigma_N^2$, S is the number of splits, and $D(s)$ denotes the set of descendant splits for the parent node in split s and $\boldsymbol{\omega}_s$ describes the proportions of the total variance of a parent node assigned to its branches for $s = 1, \ldots, S$.

3 Latent Gaussian models and priors for the splits

This section introduces LGMs and the priors we will use for the splits to build the intuitive class of joint priors for the variance parameters for LGMs.

3.1 Latent Gaussian models

LGMs constitute a subclass of BHMs with additive latent structure where the model components are Gaussian conditional on the model parameters. We write the additive model in (2.1) in vector form, $\boldsymbol{\eta} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \sum_{j=1}^N \mathbf{A}_j \boldsymbol{u}_j$, where $\mathbf{1} = (1, \ldots, 1)$ is a column vector of length n, $\mathbf{X}$ is the $n \times p$ design matrix that contains the covariates for each observation as rows, and $\mathbf{A}_j$ are sparse $n \times m_j$ matrices that select the appropriate elements of the random effects for $j = 1, \ldots, N$. The latent Gaussian structure is achieved by $\beta_0 \sim \mathcal{N}(0, \sigma_{\mathrm{I}}^2)$, $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \sigma_{\mathrm{F}}^2 \mathbf{I}_p)$, and $\boldsymbol{u}_j|\sigma_j^2 \sim \mathcal{N}_{m_j}(\mathbf{0}, \sigma_j^2 \Sigma_j)$ for $j = 1, \ldots, N$. It is common to give σ_{I}^2 and σ_{F}^2 suitably vague values, and we will assume that σ_{I}^2 and σ_{F}^2 are fixed and focus on the variance parameters $\sigma_1^2, \ldots, \sigma_N^2$.

For non-intrinsic Gaussian random effects, such as independent and identically distributed (i.i.d.) random effects, stationary autoregressive processes and Matérn Gaussian random fields, the covariance matrix Σ of the random effect $\boldsymbol{u}$ is chosen to be a correlation matrix and the variance parameter σ^2 is the marginal variance. However, this does not work for intrinsic Gaussian Markov random fields (GMRFs) (Rue and Held, 2005) such as the Besag model (Besag et al., 1991), the first-order random walk and the second-order random walk (Rue and Held, 2005, Chapter 3). In this case there is no well-defined concept of a marginal variance since they are defined through singular precision matrices that cannot be inverted to find a covariance matrix. We follow Sørbye and Rue (2014) and choose the variance parameter σ^2 to be a representative value for the marginal variance.

3.2 Introducing shrinkage towards branches

Penalising complexity

The fundamental basis for introducing robust shrinkage in our proposed class of priors are the PC priors introduced in Simpson et al. (2017), which uses a set of principles to derive model-component-specific prior distributions. The main idea is to regard a single model component as a flexible extension of a so-called base model. In the simplest case of an unstructured random effect, the base model would be to remove the effect entirely from the linear predictor by letting the variance parameter go to zero. The idea is to follow Occam's razor and favour a simpler, more sparse or more intuitive model as long as the data does not indicate otherwise. The PC priors have been used successfully in a variety of contexts such as BYM models (Riebler et al., 2016), correlation parameters (Guo et al., 2017), autoregressive processes (Sørbye and Rue, 2018) and Matérn Gaussian random fields (Fuglstad et al., 2019b).

Simpson et al. (2017) proposed to compute the complexity of the alternative model relative to the base model using the Kullback-Leibler divergence (KLD) defined as

$$\mathrm{KLD}(\pi(\boldsymbol{u}|\xi) \mid\mid \pi(\boldsymbol{u}|\xi=0)) = \int \pi(\boldsymbol{u}|\xi) \log\left(\frac{\pi(\boldsymbol{u}|\xi)}{\pi(\boldsymbol{u}|\xi=0)}\right) \mathrm{d}\boldsymbol{u}, \tag{3.1}$$

where ξ is the flexibility parameter, and $\xi = 0$ at the base model. The KLD is consequently transformed to an interpretable distance measure between two densities f_1 and f_2: $d(f_1 \mid\mid f_2) = \sqrt{2\mathrm{KLD}(f_1 \mid\mid f_2)}$. In contrast to defining a prior for ξ directly, a prior is defined for d. See Simpson et al. (2017) for detailed motivation.

We follow Simpson et al. (2017) and select an exponential distribution, where information provided by the user is used to determine the rate λ. Usually this information is provided by a probability statement about the tail probability of the prior,

$$P(X(\xi) > U) = \alpha.$$

Here, $X(\xi)$ is an interpretable transformation of the parameter of the flexible extension, U can be thought of as a sensible upper bound, and α is a small probability. A user can express their knowledge by constraining tail probabilities of $X(\xi)$ as above. Selecting U near a large plausible value for $X(\xi)$ and α small encodes weak information about ξ (Simpson et al., 2017). This means that it is *a priori* unlikely that the value of $X(\xi)$ exceeds U. Finally, the prior can be transformed to the corresponding prior for the flexibility parameter ξ. An attractive feature of this principle-based construction is that the resulting priors are proper and have a natural link to Jeffreys' priors.

Shrinking a marginal variance parameter

In the case of a single Gaussian random effect with marginal variance σ^2, the PC prior with base model $\sigma^2 = 0$ is an exponential prior on σ. The rate parameter λ can be set, for example, by an *a priori* statement $\mathrm{P}(\sigma > U) = 0.05$ so that the 95th percentile of the prior for σ is $U > 0$. Then the prior is an exponential prior with rate parameter $\lambda = -\log(\alpha)/U$ which we denote as $\sigma \sim \mathrm{PC}_{\mathrm{SD}}(U, \alpha)$; see Simpson et al. (2017) for details and derivation.

Shrinking a weight parameter

Consider the situation that the linear predictor only contains two random effects A and B with variances σ^2_{A} and σ^2_{B}, respectively. The proportion of $t = \sigma^2_{\text{A}} + \sigma^2_{\text{B}}$ assigned to each random effect is described by $\boldsymbol{\omega} = (1-\omega, \omega) = (\sigma^2_{\text{A}}, \sigma^2_{\text{B}})/(\sigma^2_{\text{A}} + \sigma^2_{\text{B}})$. If one *a priori* prefers the attribution $\boldsymbol{\omega} = \boldsymbol{\omega}^0 = (1-\omega_0, \omega_0)$, shrinkage can be induced in the joint prior for the variance parameters using a PC prior where $\boldsymbol{\omega} = \boldsymbol{\omega}^0$ is the base model. Here we apply the KLD from (3.1) to express distance from the base model $\boldsymbol{\omega}^0$ to the alternative model $\boldsymbol{\omega}$, and penalise deviations from the base model according to the difference in model complexity.

Theorem 1 (PC prior for dual split)**.** *Let $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ be random effects of an LGM that enter the linear predictor through $\mathbf{A}_1\boldsymbol{u}_1 \sim \mathcal{N}_n(\mathbf{0}, \sigma_1^2\tilde{\Sigma}_1)$ and $\mathbf{A}_2\boldsymbol{u}_2 \sim \mathcal{N}_n(\mathbf{0}, \sigma_2^2\tilde{\Sigma}_2)$. Assume that $\tilde{\Sigma}_1+\tilde{\Sigma}_2$ is non-singular.*[1] *Let $\omega = \sigma_2^2/(\sigma_1^2+\sigma_2^2)$ and $\Sigma(w) = (1-\omega)\tilde{\Sigma}_1+\omega\tilde{\Sigma}_2$. Then the distance from the base model $\Sigma(\omega_0)$ to the alternative model $\Sigma(\omega)$ is given by $d(\omega) = \sqrt{\text{tr}(\Sigma(\omega_0)^{-1}\Sigma(\omega)) - n - \log|\Sigma(\omega_0)^{-1}\Sigma(\omega)|}$ for $0 \leq \omega_0 \leq 1$.*

The PC prior for ω with base model $\omega_0 = 0$ is

$$\pi(\omega) = \begin{cases} \frac{\lambda\left|d'(\omega)\right|}{1-\exp(-\lambda d(1))}\exp\left(-\lambda d\left(\omega\right)\right), & 0 < w < 1,\ \tilde{\Sigma}_1 \text{ non-singular}, \\ \frac{\lambda}{2\sqrt{\omega}(1-\exp(-\lambda))}\exp(-\lambda\sqrt{\omega}), & 0 < \omega < 1,\ \tilde{\Sigma}_1 \text{ singular}, \end{cases}$$

where $\lambda > 0$ is the hyperparameter. We suggest to set λ so that the median is $\omega_{\text{m}} = 0.25$.

For base model $0 < \omega_0 < 1$, the PC prior whose median is equal to ω_0 is

$$\pi(\omega) = \begin{cases} \frac{\lambda\left|d'(\omega)\right|}{2[1-\exp(-\lambda d(0))]}\exp\left(-\lambda d\left(\omega\right)\right), & 0 < \omega < \omega_0, \\ \frac{\lambda\left|d'(\omega)\right|}{2[1-\exp(-\lambda d(1))]}\exp\left(-\lambda d\left(\omega\right)\right), & \omega_0 < \omega < 1, \end{cases}$$

where $\lambda > 0$ is a hyperparameter. We suggest to set λ so that

$$\text{P}(\text{logit}(1/4) + \text{logit}(\omega_0) < \text{logit}(\omega) < \text{logit}(\omega_0) + \text{logit}(3/4)) = 1/2.$$

Base model equal to $\omega_0 = 1$ follows directly by reversing the roles of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$.

Proof. See Section S1.1 in the Supplementary Materials. □

The default values in each case are specified as to place most of the prior mass in a small interval on the ω scale around ω_0, but to also ensure large deviations from ω_0 are *a priori* plausible; in this sense they are weakly informative (Gelman, 2006; Gelman et al., 2008). Sections 5.1 and 5.2 show that the results from the inference are stable to changes in these hyperparameters; which in turn shows that these λ's provide weak information. If the analyst has expert knowledge this should be used instead of the default values. Large ω might be 0.75 for test-retest reliability in a psychology study (Cicchetti, 1994) but 0.4 for the genetic heritability of a trait (Shen et al., 2016).

[1] If this were not the case, some elements of the sum of $\mathbf{A}_1\boldsymbol{u}_1$ and $\mathbf{A}_2\boldsymbol{u}_2$ would be exactly equal and we would choose a subset of maximal size so that $\tilde{\Sigma}_1 + \tilde{\Sigma}_2$ was non-singular for comparing the effects of $\mathbf{A}_1\boldsymbol{u}_1$ and $\mathbf{A}_2\boldsymbol{u}_2$.

3.3 Expressing a priori ignorance about a split

Exchangeability

In some cases the analyst does not want to express an *a priori* preference for any of the branches in a split in the tree. This can be achieved indirectly through a series of dual splits. For example, by replacing the split in Figure 1a by the series of dual splits as shown in Figure 1b where the left-hand side has a base model of 2/3 in the first split and the left-hand side has a base model of 1/2 for the second split. In total this is specifying a base model of 1/3 of the total variance to each random effect, but the resulting prior is not invariant to permutations of A, B and C in Figure 1b. See Section S2 of the Supplementary Materials for details. When the goal is to express ignorance about the decomposition of the variance, one can use a base model of equal attribution of the total variance to each random effect and choose an exchangeable prior for $(\sigma^2_{\mathrm{A}}, \sigma^2_{\mathrm{B}}, \sigma^2_{\mathrm{B}})$. This can be done, for example, through a Dirichlet prior.

Dirichlet prior

The Dirichlet prior of order $K \geq 2$ with parameters $a_1, \ldots, a_K > 0$ is given by

$$\pi(\boldsymbol{\omega}) = \frac{1}{B(a_1, \ldots, a_K)} \prod_{k=1}^{K} \omega_k^{a_k - 1}, \quad \boldsymbol{\omega} = (\omega_1, \ldots, \omega_K) \in \Delta^K,$$

where B is the multivariate beta function, and Δ^K is the $K-1$ simplex. Since there is no preference for any random effect, we consider the symmetric Dirichlet distribution where $a_1 = \ldots = a_K = a > 0$, where a is the hyperparameter that must be selected by the analyst. For $a = 1$ the prior is uniform, for $a < 1$ the prior has peaks at the vertrices of Δ^K, and for $a > 1$ the mode is $\boldsymbol{\omega} = (1, \ldots, 1)/K$. The prior is invariant to permutations of the elements of $\boldsymbol{\omega}$ for any value of $a > 0$ and it is computationally cheap for arbitrary dimensions K.

The hyperparameter a can be selected by considering the marginal properties of $\pi(\boldsymbol{\omega})$. The marginal prior $\pi(\omega_1) \propto \omega_1^{a-1}(1-\omega_1)^{(K-1)a-1}$, $0 < \omega_1 < 1$, is a Beta distribution whose quantiles are dependent both on the values of a and K. We select a by requiring $\mathrm{P}(\mathrm{logit}(1/4) < \mathrm{logit}(\omega_1) - \mathrm{logit}(\omega_0) < \mathrm{logit}(3/4)) = 1/2$. By symmetry the same marginal properties are satisfied for ω_i, $i = 2, \ldots, K$.

4 Hierarchical decomposition priors for LGMs

In this section we introduce the new class of intuitive joint priors for the variance parameters in LGMs.

4.1 Accounting for model structure

In the general formulation of HD priors in Definition 1, the prior is composed of conditional priors that for each split depends on all descendant splits. This is impractical

because computing PC priors would require new KLDs to be computed every time the prior is evaluated. We take a pragmatic approach where we decide on a set of base models, which expresses our best prior guess, and condition on these.

Assumption 2 (Simplified conditioning)**.** *For a given tree with S splits and base models $\{\boldsymbol{\omega}_1^0, \ldots, \boldsymbol{\omega}_S^0\}$, we replace $\pi(\boldsymbol{\omega}_s|\{\boldsymbol{\omega}_j\}_{j\in \mathrm{D}(s)})$ with $\pi(\boldsymbol{\omega}_s|\{\boldsymbol{\omega}_j = \boldsymbol{\omega}_j^0\}_{j\in \mathrm{D}(s)})$, $s = 1, \ldots, S$.*

Under this assumption a new class of HD priors for LGMs are constructed by combining intuition about shrinkage and ignorance through independent priors for the splits.

Prior class 1 (HD priors for LGMs)**.** *Assume the LGM contains N random effects with variances $\sigma_1^2, \ldots, \sigma_N^2$ and that the hierarchical decomposition of the variance is described through a tree with S splits. Under base models $\{\boldsymbol{\omega}_1^0, \ldots, \boldsymbol{\omega}_S^0\}$, the prior is*

$$\pi(\sigma_1^2, \ldots, \sigma_N^2) = \pi(t|\{\boldsymbol{\omega}_s\}_{s=1}^S) \prod_{s=1}^{S} \pi(\boldsymbol{\omega}_s|\{\boldsymbol{\omega}_j = \boldsymbol{\omega}_j^0\}_{j\in \mathrm{D}(s)}),$$

where the total latent variance is $t = \sigma_1^2 + \ldots + \sigma_N^2$, and $\boldsymbol{\omega}_i \in \Delta^{l_s}$, where l_s is the number of branches in split s, $s = 1, \ldots, S$.

For each of the S splits, the analyst can express ignorance through a Dirichlet prior or sequence of PC priors as described in Section 3.3, or express preference to the selected base models as described in Section 3.2. The selection of $\pi(t|\{\boldsymbol{\omega}_s\}_{s=1}^S)$ must be done in the context of the likelihood as described in Section 4.2.

This prior is computationally inexpensive since the overall prior probability density factorises into independent conditional distributions that consist of PC priors, which can be precomputed, and Dirichlet priors, which are cheap to compute.

We demonstrate the use of HD priors through one example where the analyst wants to express ignorance and one example where the analyst wants to penalise complexity.

Example 2 (Non-nested random effects)**.** Consider responses $y_1, \ldots, y_n$, described by the Gaussian linear model $y_i|\eta_i \sim \mathcal{N}(\eta_i, \sigma_{\mathrm{R}}^2)$ with

$$\eta_i = \mu + h_1(\mathrm{Age}_i) + h_2(\mathrm{Weight}_i) + h_3(\mathrm{Income}_i), \quad i = 1, 2, \ldots, n,$$

where μ is the intercept, h_1, h_2 and h_3 are smooth effects of the covariates expressed as second-order random walks (Rue and Held, 2005), and σ_{R}^2 is the residual variance. Assume that one has no *a priori* preference for the three smooth effects, and decide to encode the decomposition of the total latent variance as shown Figure 1a, where A, B and C represents the three smooth of covariates effects. Let $\boldsymbol{\omega}_1$ denote the proportions of variance assigned to model components and let t denote the total latent variance. We construct an HD prior by assigning a Dirichlet prior to $\boldsymbol{\omega}_1$, and handle $t|\boldsymbol{\omega}_1$ as discussed in Section 4.2. △

Example 3 (Shrinkage in multilevel models)**.** The latent part of the multilevel model in Section 1 can be written in vector form as $\boldsymbol{\eta} = \mathbf{A}_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}} + \mathbf{A}_{\mathrm{B}}\boldsymbol{u}_{\mathrm{B}} + \mathbf{A}_{\mathrm{C}}\boldsymbol{u}_{\mathrm{C}}$, where $\mathbf{A}_{\mathrm{A}}$, $\mathbf{A}_{\mathrm{B}}$ and $\mathbf{A}_{\mathrm{C}}$ are sparse matrices selecting the appropriate group, individual and

measurement effects, respectively. Assume we use an LGM, then $\boldsymbol{u}_1 \sim \mathcal{N}_G(\mathbf{0}, \sigma^2_{\mathrm{A}}\mathbf{I}_G)$, $\boldsymbol{u}_2 \sim \mathcal{N}_{GP}(\mathbf{0}, \sigma^2_{\mathrm{B}}\mathbf{I}_{GP})$ and $\boldsymbol{u}_3 \sim \mathcal{N}_{GPK}(\mathbf{0}, \sigma^2_{\mathrm{C}}\mathbf{I}_{GPK})$, where G is the number of groups, P is the number of individuals per group, and K is the number of measurements per individual.

If we prefer shrinkage towards fewer levels in the multilevel model as shown in Figure 1c, we decompose the total latent variance $t = \sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}} + \sigma^2_{\mathrm{C}}$ through two splits. For the split at the root node, we decompose t according to the proportions $\boldsymbol{\omega}_1 = (\sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}}, \sigma^2_{\mathrm{C}})/t$. Then in the second split we decompose $\sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}}$ according to the proportions $\boldsymbol{\omega}_2 = (\sigma^2_{\mathrm{A}}, \sigma^2_{\mathrm{B}})/(\sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}})$.

We use an HD prior where we apply base models $\boldsymbol{\omega}^0_1 = (0, 1)$, which prefers C over A+B, and $\boldsymbol{\omega}^0_2 = (0, 1)$, which prefers B over A. Due to the desire for shrinkage we apply PC priors and use Theorem 1 with base model $\boldsymbol{\omega}^0_2$ to compute $\pi(\boldsymbol{\omega}_2)$. We define $\tilde{\boldsymbol{u}}_1 = \mathbf{A}_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}} + \mathbf{A}_{\mathrm{B}}\boldsymbol{u}_{\mathrm{B}}$ and $\tilde{\boldsymbol{u}}_2 = \mathbf{A}_{\mathrm{C}}\boldsymbol{u}_{\mathrm{C}}$. Then if we condition on $\boldsymbol{\omega}_2$, the top split in Figure 1c compares $\tilde{\boldsymbol{u}}_1|\boldsymbol{\omega}_2 \sim \mathcal{N}_n(\mathbf{0}, (\sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{B}})(\omega_{2,1}\mathbf{A}_{\mathrm{A}}\mathbf{A}^{\mathrm{T}}_{\mathrm{A}} + \omega_{2,2}\mathbf{A}_{\mathrm{B}}\mathbf{A}^{\mathrm{T}}_{\mathrm{B}}))$ and $\tilde{\boldsymbol{u}}_2 \sim \mathcal{N}_n(\mathbf{0}, \sigma^2_3\mathbf{A}_3\mathbf{A}^{\mathrm{T}}_3)$, and the conditional prior $\pi(\boldsymbol{\omega}_1|\boldsymbol{\omega}_2 = \boldsymbol{\omega}^0_2)$ can be computed using Theorem 1 with base model $\boldsymbol{\omega}^0_1$ conditional on $\boldsymbol{\omega}_2 = \boldsymbol{\omega}^0_2$. The joint prior is then $\pi(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) = \pi(\boldsymbol{\omega}_1|\boldsymbol{\omega}_2 = \boldsymbol{\omega}^0_2)\pi(\boldsymbol{\omega}_2)$, and an appropriate prior is chosen for $\pi(t|\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ as described in Section 4.2. △

4.2 Accounting for the likelihood

Meaningful priors for the total latent variance t depend on the likelihood and prior beliefs about the responses in the specific application (Gelman et al., 2017). We provide tools for expressing scale-invariance for the variances of the random effects and the measurement error when the responses are Gaussian, or shrinkage for the total latent variance of the random effects.

Under a Gaussian likelihood, the selection of the unit of measurement by the analyst affects the sizes of the variances. However, when the residual variance σ^2_{R} is expected to be well-identified, we can define the prior on t relative to σ^2_{R} and shrink t by preferring to describe the total variance $V = t + \sigma^2_{\mathrm{R}}$ in the model by σ^2_{R}. This can be complemented by a scale-independent Jeffreys' prior on V to achieve a scale-invariant joint prior for the variance parameters.

Prior class 2 (HD priors with Gaussian likelihoods)**.** *Assume an HD prior from Prior class 1 is desired for an LGM with Gaussian responses with residual variance σ^2_{R}. First select the prior on the decomposition of the total latent variance t. Then augment the tree by an extra node on the top with variance $V = t + \sigma^2_{\mathrm{R}}$. The new top node has one branch with residual variance and the other branch is the subtree describing the latent model. Let $\boldsymbol{\omega}_{\mathrm{R}} = (1 - \sigma^2_{\mathrm{R}}/V, \sigma^2_{\mathrm{R}}/V)$ and assume shrinkage through a PC prior $\pi(\boldsymbol{\omega}_{\mathrm{R}}|\{\boldsymbol{\omega}_s = \boldsymbol{\omega}^0_s\}^S_{s=1})$ with base model $\boldsymbol{\omega}^0_{\mathrm{R}} = (0, 1)$.*

If V is assigned a scale-invariant prior, the full joint prior is

$$\pi(V, \boldsymbol{\omega}_{\mathrm{R}}, \{\boldsymbol{\omega}_s\}^S_{s=1}) \propto \pi(\boldsymbol{\omega}_{\mathrm{R}}|\{\boldsymbol{\omega}_s = \boldsymbol{\omega}^0_s\}^S_{s=1})\pi(\{\boldsymbol{\omega}_s\}^S_{s=1})/V, \quad V > 0, \boldsymbol{\omega}_{\mathrm{R}} \in \Delta^2,$$

and $\boldsymbol{\omega}_s \in \Delta^{l_s}$, where l_s is the number of branches in split s, for $s = 1, \ldots, S$.

Proof. The scale-invariant prior is $\pi(V|\boldsymbol{\omega}_{\mathrm{R}}, \{\boldsymbol{\omega}_s\}_{s=1}^S) \propto 1/V$, and $\pi(\boldsymbol{\omega}_{\mathrm{R}}, \{\boldsymbol{\omega}_s\}_{s=1}^S) = \pi(\boldsymbol{\omega}_{\mathrm{R}}|\{\boldsymbol{\omega}_s\}_{s=1}^S)\pi(\{\boldsymbol{\omega}_s\}_{s=1}^S)$ □

If the likelihood is binomial with a logit link function, a scale for the random effects is induced through their effects on the odds-ratio. Similarly, for a Poisson likelihood with a log link function, there is a scale for the random effects through their effects on the relative risk. In these cases, scale-invariance is not meaningful and we can induce shrinkage on the total variance of the random effects by using the PC prior for variance from Simpson et al. (2017).

Prior class 3 (HD priors with shrinkage on latent variance)**.** *Assume an HD prior from Prior class 1 is desired for an LGM where shrinkage on the total latent variance is appropriate. First select the prior on the decomposition of the total latent variance t. Then t can be shrunk towards 0 by a PC prior $\pi(t|\{\boldsymbol{\omega}_s\}_{s=1}^S)$ with base model $t_0 = 0$. This results in*

$$\pi(t, \{\boldsymbol{\omega}_s\}_{s=1}^S) = \frac{\lambda}{2\sqrt{t}} \exp(-\lambda\sqrt{t})\pi(\{\boldsymbol{\omega}_s\}_{s=1}^S),$$

$t > 0$, and $\boldsymbol{\omega}_i \in \Delta^{l_s}$, where l_s is the number of branches in split s, for $s = 1, \ldots, S$, and $\lambda > 0$ is a hyperparameter.

Proof. The conditional PC prior for t with base model $t_0 = 0$ is given by $\pi(t|\{\boldsymbol{\omega}_s\}_{s=1}^S) = \lambda \exp(-\lambda\sqrt{t})/(2\sqrt{t})$, $t > 0$ (Simpson et al., 2017). □

We illustrate how the hyperparameter can be selected by considering the prior on the total latent variance in the case of a Binomial likelihood.

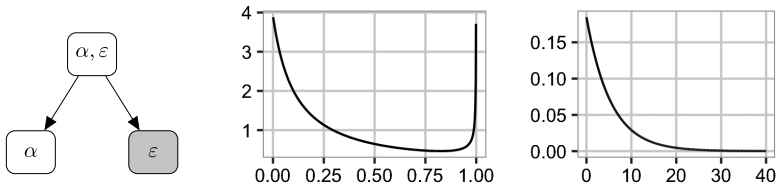
Example 4 (Shrinking latent variance)**.** Let $\mathrm{logit}(p) = \mu + x$, where $x \sim \mathcal{N}(0, t)$, for a $t > 0$, and μ is considered fixed. The latent variance t is difficult to interpret directly due to the non-linear link function, but we can interpret it through the effect on the odds-ratio, $p/(1-p) = \exp(\mu)\exp(x)$. The hyperparameter λ in Prior class 3 can, for example, be set so that the relative change in the odds-ratio, $\exp(x)$, is between 1/2 and 2 with probability 90%, $\mathrm{P}(1/2 < \exp(x) < 2) = 0.90$. △

5 Case studies: Gaussian responses

In this section we investigate the performance of HD priors compared to a set of commonly used standard priors for two simulation studies with Gaussian responses.

5.1 Random intercept model

The *random intercept model* is given by $y_{i,j} = \alpha_i + \varepsilon_{i,j}$ for $j = 1, \ldots, n_i$, $i = 1, \ldots, n_{\mathrm{g}}$, where n_i is the size of group i, and n_{g} is the number of groups. The random intercepts are i.i.d. Gaussian with variance σ_α^2 and the residual effects are i.i.d. Gaussian with variance σ_{R}^2. The total latent variance is $t = \sigma_\alpha^2$ and the total variance is $V = \sigma_{\mathrm{R}}^2 + \sigma_\alpha^2$. We introduce the proportion of the total variance explained by the latent model $\omega = \sigma_\alpha^2/V$, and decompose V as $\sigma_\alpha^2 = \omega V$ and $\sigma_{\mathrm{R}}^2 = (1-\omega)V$. We desire shrinkage towards the



(a) Model structure (b) $\pi(\omega)$ (c) $\pi(d(\omega))$

Figure 2: Model structure and prior for ω in the random intercept model with 10 individuals in each group and prior median $\omega_{\text{m}} = 0.25$. The prior is independent of the number of groups. a) Tree structure, b) prior for ω, and c) prior for distance $d(\omega)$.

base model $\omega^0 = 0$ and use an HD prior based on the tree structure in Figure 2a, where the prior on ω is calculated using Theorem 1 and we use the scale-invariant prior from Prior class 2. The specification of the hyperparameter of the HD prior is done through the median ω_{m} of $\pi(\omega)$. The resulting prior for ω is shown in Figure 2b for $\omega_{\text{m}} = 0.25$ and the corresponding prior for the distance $d(\omega)$ discussed in Section 3.2 is shown in 2c. Further details can be found in Section S3.1 of the Supplementary Materials.

The intraclass correlation (ICC) for the random intercept model is given by $\sigma_\alpha^2/(\sigma_{\text{R}}^2 + \sigma_\alpha^2)$, which equals the weight parameter ω. Thus the shrinkage of the ICC is completely controlled in the construction of the prior and expert knowledge about the ICC can be incorporated directly. Further, ω can be linked to a generalised version of the coefficient of determination, R^2, suggested by Gelman and Hill (2007); see Section S3.2 in the Supplementary Materials for details.

We use the `R`-package `RStan` (Stan Development Team, 2018a) to perform the inference for the simulation study. We use HD priors from Prior class 2 with shrinkage from PC priors on ω with hyperparameters $\omega_{\text{m}} = 0.25$ (P-HD-25), $\omega_{\text{m}} = 0.5$ (P-HD-50) and $\omega_{\text{m}} = 0.75$ (P-HD-75), and an HD prior from Prior class 2 where the PC prior is replaced by a Dirichlet prior on $(\omega, 1-\omega)$ (P-HD-D) with default hyperparameter. Additional priors are Jeffreys' prior on the residual variance combined with different priors on the random intercepts variance or standard deviation: the default INLA prior InvGamma$(1, 5\times10^{-5})$ (P-INLA), Half-Cauchy(25) (P-HC), and $\text{PC}_{\text{SD}}(3, 0.05)$ (P-PC). This gives seven joint priors. Each scenario in the simulation study consists of 500 datasets which are simulated from the random intercept model for $n_{\text{g}} \in \{5, 10, 50\}$, and 10, 50, or varying number of individuals in each group. We select true values $\omega \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and select true total variance $V = 1$ in every scenario.

We evaluate the performance of the different priors with respect to posterior inference for total variance V and ICC ω. We use the bias of $\log(V)$ and $\text{logit}(\omega)$, calculated using the estimated median minus the true value, and the 80% empirical coverage, found by counting the number of times the true value is contained in the 80% equal-tailed credible interval. We use the same settings for the call to the `stan` function for all priors and scenarios in the simulation study. `RStan` reports a *divergent transition* for each iteration of the Markov chain Monte Carlo (MCMC) sampler that runs into numerical instabilities (Carpenter et al., 2017). In Figure S3.1 in the Supplementary Materials we

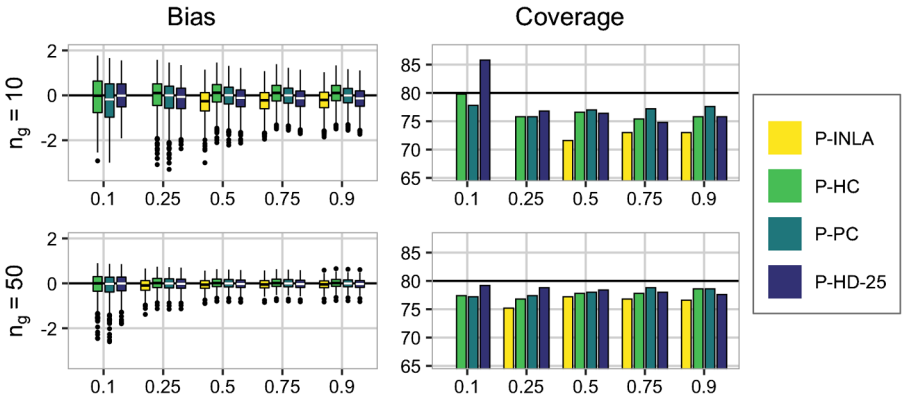


Figure 3: Results for logit(ω) for the random intercept simulation study. True value of ω shown on the x-axis, the number of groups is shown on left-hand side, and the group size is 10. Results for P-INLA are only shown when it leads to stable inference.

report the proportion of datasets that resulted in at most 0.1% divergent transitions for each prior and scenario. This is used as a measure of stability of the inference scheme for each prior, and the dataset and prior combinations causing unstable inference are removed from the study.

The results in Figure 3 are for $n_g \in \{10, 50\}$ and group size 10, and show that P-HD-25 performs at least as good in terms of bias and coverage of logit(ω) as P-INLA, P-HC and P-PC. The magnitude of the bias decreases and the coverage approaches 80% for all four priors when the number of groups increases, which is expected as the amount of information about the parameters in the datasets increases. Figures S3.3–S3.7 in the Supplementary Materials show that the HD priors perform at least as good in terms of bias and coverage for logit(ω) as P-INLA, P-HC and P-PC also for the other combinations of the number of groups and group sizes, and that the same conclusions as for logit(ω) also holds for $\log(V)$.

Furthermore, Figures S3.3–S3.7 show that the behaviour of the four HD priors is stable with respect to the choice of ω_{m} when group size is 10, and that P-HD-D performs worse than P-HD-25, P-HD-50 and P-HD-75 for all values of the true weight except 0.5. For 10 groups with two observations per group, the risk of overfitting is high because low information about the parameters may lead to overestimating the weight parameter and estimating spurious signals in the group effect. In this setting, P-HD-25 leads to overfitting for true weight equal to 0.1, but underfitting for true weight equal to 0.25, 0.5, 0.75 and 0.9. P-HD-50, P-HD-75 and P-HD-D result in overfitting for true weight equal to 0.1 and 0.25, but underfitting for true weight equal to 0.5, 0.75 and 0.9. See Section S3.4 in the Supplementary Materials for additional details.

Figure S3.1 shows that P-INLA is the only prior that is heavily affected by divergent transitions during the inference for scenarios with 10 or 50 groups. Part of the problem with P-INLA is that it results in a bi-modal posterior for σ_α^2; see Figure S3.2. The

new HD priors are preferred for the random intercept model due to their intuitive definition, where the structure of the shrinkage is directly available in Figure 2a, and interpretability of the parametrization which aids prior elicitation.

5.2 Latin square experiment

Consider an experiment where a latin square design (Hinkelmann and Kempthorne, 1994) is used to control for two nuisance sources of noise. For example, a field split into rows and columns where different levels of strength of a new fertilizer is applied to each plot. We assume there are nine possible levels of the treatment so that a 9×9 grid of plots is necessary for a full latin square design. We focus on random effects and exclude fixed effects from the model, and assume that the responses can be modelled by

$$y_{i,j} = \alpha_i + \beta_j + \gamma_{k[i,j]} + \varepsilon_{i,j}, \quad i, j = 1, \ldots, 9, \tag{5.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_9) \sim \mathcal{N}_9(\mathbf{0}, \sigma_\text{r}^2 \mathbf{I}_9)$ is an i.i.d. effect of row, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_9) \sim \mathcal{N}_9(\mathbf{0}, \sigma_\text{c}^2 \mathbf{I}_9)$ is an i.i.d. effect of column, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_9)$ is the effect of the treatment, $k[i, j]$ denotes the treatment assigned to row i and column j, and $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}, \ldots, \varepsilon_{9,9}) \sim \mathcal{N}_{81}(\mathbf{0}, \sigma_\text{R}^2 \mathbf{I}_{81})$ is the residual noise.

We believe that the effect of the treatment is ordered, and that the treatment effect consists of a smooth signal of interest $\boldsymbol{\gamma}^{(1)} = (\gamma_1^{(1)}, \ldots, \gamma_9^{(1)})$ and random noise $\boldsymbol{\gamma}^{(2)} = (\gamma_1^{(2)}, \ldots, \gamma_9^{(2)})$ we have to control for. The signal is given a second-order random walk model described by $\mathcal{N}_9(\mathbf{0}, \sigma_\text{RW2}^2 \mathbf{Q}_\text{RW2}^{-1})$, where σ_RW2^2 is the variance and $\mathbf{Q}_\text{RW2}^{-1}$ is a slight abuse of notation to describe the intrinsic second-order random walk defined by the precision matrix $\mathbf{Q}_\text{RW2}$, and the noise is $\boldsymbol{\gamma}^{(2)} \sim \mathcal{N}_9(\mathbf{0}, \sigma_\text{t}^2 \mathbf{I}_9)$. We use the constraints $\sum_{i=1}^9 \gamma_i^{(1)} = 0$ and $\sum_{i=1}^9 i\gamma_i^{(1)} = 0$ to remove the implicit intercept and linear effect, respectively.

We set the true standard deviations equal, $\sigma_\text{r} = \sigma_\text{c} = \sigma_\text{t} = \sigma_\text{R} = 0.1$, and let the true effect of treatment be given by $x_i = C\left((i-5)^2 - 20/3\right)$, $i = 1, \ldots, 9$. We entertain three scenarios: $C = 0$ for no effect of treatment (S1), $C = 0.05$ for medium effect of treatment (S2) and $C = 0.2$ for strong effect of treatment (S3). More details on the true treatment effect is included in Section S4.1 in the Supplementary materials, see especially Figure S4.2. We simulate 500 datasets for each scenario and analyse them with four choices of priors.

The three default priors used are Jeffreys' prior for σ_R^2 combined with InvGamma$(1, 5 \times 10^{-5})$ for σ_r^2, σ_c^2, σ_t^2 and σ_RW2^2 (P-INLA), or Half-Cauchy(25) (P-HC) or $\text{PC}_\text{SD}(3, 0.05)$ (P-PC) for σ_r, σ_c, σ_t and σ_RW2. We select an HD prior from Prior class 2 using the model structure in Figure 4a, where the triple split has a Dirichlet prior and the two other splits have PC priors (P-HD-D3). We also decompose the triple split into the two dual splits as shown in Figure 4b, and use a PC prior on all four splits according to the shrinkage structure in the figure (P-HD-25). In all cases we use default values for the hyperparameters. See Section S2 in the Supplementary Materials for more details on changing a triple split to two dual splits. Figures S4.3, S4.4, S4.10 and S4.11 in the Supplementary Materials show that the implementation of the triple split has little influence on the targets of the analysis.

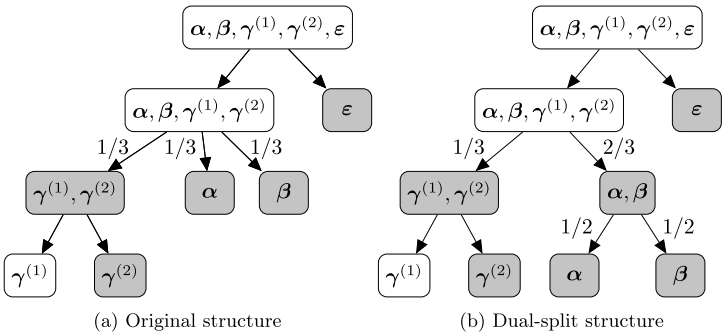


Figure 4: Model structure for the latin square simulation study. Gray nodes indicate base models. $(1/3, 1/3, 1/3)$, $(1/3, 2/3)$, and $(1/2, 1/2)$ indicates that the base model for the split is a combination of the branches. a) Original, and b) alternative structure.

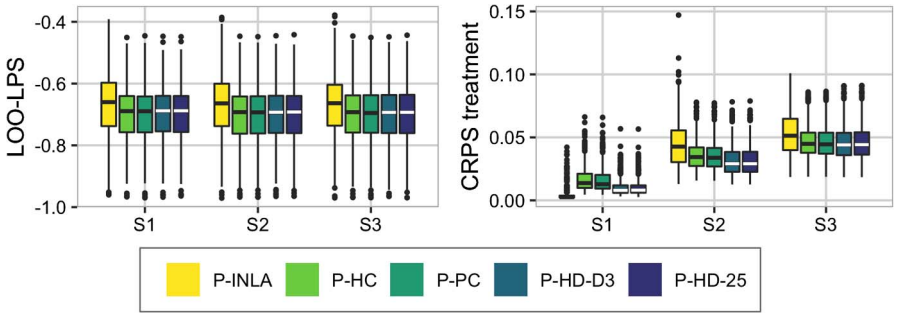


Figure 5: Results from the latin square experiment simulation study.

The targets of the analysis are the posterior distribution of the structured treatment effect $\boldsymbol{\gamma}^{(1)}$ and the model fit. The former will be assessed by the continuous rank probability score (CRPS) (Gneiting and Raftery, 2007) and the latter by the leave-one-out log predictive score (LOO-LPS) $-\frac{1}{81}\sum_{i=1}^{81}\log\pi(y_i|\boldsymbol{y}_{-i})$. The CRPS is a proper scoring rule and given by $\frac{1}{9}\sum_{i=1}^{9}\int_{-\infty}^{\infty}\left(F_i(x)-\mathbb{I}(x\geq x_i)\right)^2\mathrm{d}x$, where F_i is the cumulative distribution function for the posterior of $\gamma_i^{(1)}$, x_i is the true effect of treatment i, and $\mathbb{I}$ is the Heaviside function, and is estimated using the procedure of Jordan et al. (2017). We report the proportion of datasets leading to no more than 0.1% divergent transitions for each prior and scenario, and use this as a measure on stability of the inference. These numbers can be seen in Figure S4.5 in the Supplementary Materials, and show that all priors lead to similar stability. The datasets leading to more than 0.1% divergent transitions for one or more priors are removed from the study.

The main results from the simulation study are displayed in Figure 5. Low LOO-LPS indicates good model fit and low CRPS indicates good predictive power for the treatment effect. P-INLA gives a poorer model fit than the other priors, and with respect

to predictive power, the HD priors P-HD-D3 and P-HD-25 perform best for S2 and S3. The high predictive power of P-INLA for S1 is due to the fact that P-INLA has a peak at low variance and produces a posterior for the treatment effect with mean closer to zero and lower variance. Overall, the HD prior performs well across all scenarios. The results are stable to changes in the construction of the HD prior and the choice of hyperparameters; see Section S4.2 in the Supplementary Materials for details. The HD priors are preferable to the other priors because of their intuitive parametrization and the interpretability of the *a priori* assumptions placed on the joint prior of the variance parameters. Further, P-HD-D3 is preferred to P-HD-25 since they perform similar and P-HD-D3 is more intuitive.

6 Case studies: Binomial responses

In this section we study neonatal mortality counts arising from complex surveys through a simulation study, and show how to practically apply the HD priors.

6.1 Background

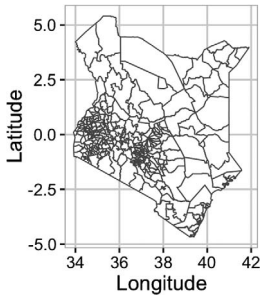
Neonatal mortality is an important indicator of health and well-being in a country and is included in Goal 3.2 of the Sustainable Development Goals (SDGs) (General Assembly of the United Nations, 2015), and mapping child mortality is an important area of current research (Golding et al., 2017; Wakefield et al., 2019; Li et al., 2019). We define neonatal mortality as the rate of deaths within the first month of life per live birth. An important source of data for neonatal mortality is the nationally-representative household surveys performed by Demographic and Health Surveys (DHS). The survey performed by DHS in 2014 in Kenya targets its 47 counties, which is the relevant administrative level for health policies (Kenya National Bureau of Statistics et al., 2015). The target of the simulation study in Section 6.2 and the analysis in Section 6.3 is the spatial heterogeneity in neonatal mortality in Kenya in the time period 2010 to the time of the survey.

From the survey we can extract the number of live births, $b_{i,j,k}$, and the number of neonatal deaths, $y_{i,j,k}$, in household k in cluster j in county i. We also have an indicator $x_{i,j}$ specifying whether the cluster is rural (0) or urban (1) and each household has an inclusion probability $\pi_{i,j,k}$ of being included in the survey sample. See the Section S5.1 in the Supplementary Materials for more background.

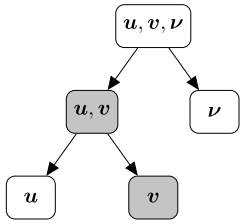
6.2 Simulation study

In this section we use the $n = 290$ constituencies shown in Figure 6a.[2] We assume that $m_i = 6$ clusters are visited in constituency i, $i = 1, \ldots, n$, and consider births $b_{i,j}$ and neonatal deaths $y_{i,j}$ in cluster j in constituency i. We assume that there are $b_{i,j} = 25$ live births in each cluster and the outcomes are simulated according to the

[2]Preliminary investigations revealed that 47 counties provided too little information to learn about model structure in the data. We instead use the 290 constituencies of Kenya for the simulations study.



(a) The 290 constituencies of Kenya.



(b) Model structure. Gray nodes indicate base models.

Figure 6: Map and model structure for the Kenya neonatal mortality simulation study.

model $y_{i,j}|p_{i,j} \sim \text{Binomial}(b_{i,j}, p_{i,j})$ for

$$\text{logit}(p_{i,j}) = \eta_{i,j} = \mu + u_i + v_i + \nu_{i,j}, \quad j = 1, \ldots, m_i, \ i = 1, \ldots, n,$$

where μ is a joint intercept, $\boldsymbol{u} = (u_1, \ldots, u_n)$ has a Besag distribution with variance σ^2_{B} and a sum-to-zero constraint, $\boldsymbol{v} = (v_1, \ldots, v_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma^2_{\text{IID}}\mathbf{I}_n)$, and $\boldsymbol{\nu} = (\nu_{1,1}, \ldots, \nu_{n,m_n}) \sim \mathcal{N}_M(\mathbf{0}, \sigma^2_{\text{C}}\mathbf{I}_M)$ with $M = m_1 + \ldots + m_n = 6 \cdot 290 = 1740$.

We use the structure for the prior shown in Figure 6b to make an HD prior from Prior class 3 with PC priors on all splits according to the base models indicated in the figure (P-HD-25) and an HD prior from Prior class 3 where a Dirichlet prior distributes variance to the three model components (P-HD-D). In all cases, the splits have default hyperparameter values and we select the hyperparameter in the PC prior on total variance, $t = \sigma^2_{\text{B}} + \sigma^2_{\text{IID}} + \sigma^2_{\text{C}}$, so that $\text{P}(t > 3) = 0.05$. Further, we use $\text{InvGamma}(1, 5 \times 10^{-5})$ for σ^2_{B}, σ^2_{IID} and σ^2_{C} (P-INLA), Half-Cauchy(25) for σ_{B}, σ_{IID} and σ_{C} (P-HC), and the joint prior proposed in Riebler et al. (2016) (P-PC), where σ^2_{B} and σ^2_{IID} has a PC prior of the type introduced in this paper with $\text{P}(\sigma^2_{\text{B}}/(\sigma^2_{\text{B}} + \sigma^2_{\text{IID}}) < 0.5) = 2/3$ and σ^2_{C} is given an independent PC prior $\sigma_{\text{C}} \sim \text{PC}_{\text{SD}}(3, 0.05)$.

Based on the final report from the survey (Kenya National Bureau of Statistics et al., 2015) the estimated national level of neonatal mortality is 0.022 for 2010–2014, and we set $\mu = \text{logit}(0.022)$. Further, we choose $\sigma^2_{\text{C}} = 0.1$ and create five scenarios by combining this with $\sigma^2_{\text{IID}} = \sigma^2_{\text{B}} = 0$ (S1), $\sigma^2_{\text{IID}} = 0.4$ and $\sigma^2_{\text{B}} = 0$ (S2), $\sigma^2_{\text{IID}} = \sigma^2_{\text{B}} = 0.2$ (S3), $\sigma^2_{\text{IID}} = 0.04$ and $\sigma^2_{\text{B}} = 0.36$ (S4), and $\sigma^2_{\text{IID}} = 0$ and $\sigma^2_{\text{B}} = 0.4$ (S5). We simulate 500 datasets for each scenario. The main targets of the simulation study are the structured part of the spatial heterogeneity through the posterior of $\boldsymbol{u}$, the degree of structure in the spatial heterogeneity through $\omega^{(2)} = \sigma^2_{\text{B}}(\sigma^2_{\text{B}} + \sigma^2_{\text{IID}})^{-1}$, and how well the underlying neonatal mortality is estimated through the posterior of the intercept μ. The performance is assessed through the CRPS (see Section 5.2) of $\boldsymbol{u}$, the bias of the posterior median of $\omega^{(2)}$, and the bias of the posterior median and the coverage of the 80% equal-tailed credible interval for μ. We use the proportion of datasets leading to at most 0.1% divergent transitions as a measure of stability in the inference, these

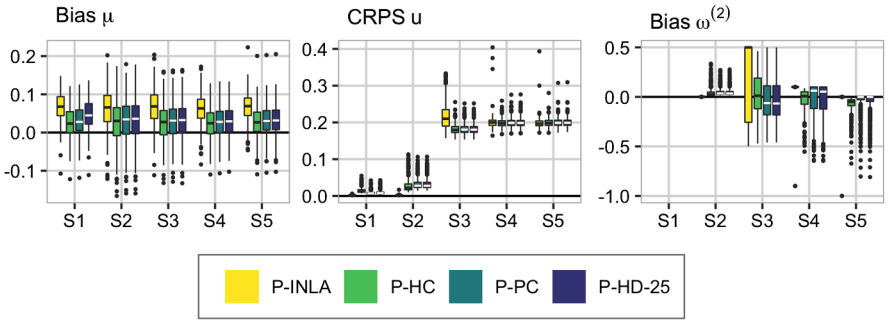


Figure 7: Main results from the Kenya neonatal mortality simulation study. Left to right: bias of the intercept μ, CRPS of $\boldsymbol{u}$ and bias of $\omega^{(2)}$. Scenario shown on the x-axes.

numbers can be seen in Figure S5.1 in the Supplementary Materials, and show that P-INLA leads to more unstable inference than the others.

Figure 7 shows the main results from the simulation study. We drop datasets that cause more than 0.1% divergent transitions for at least one of the priors from each scenario. All priors have a tendency to overestimate the intercept, with P-INLA doing worse than the others, P-INLA gives close to exact estimates when the true value of $\omega^{(2)}$ is 0 (in S2) and 1 (in S5), but performs worse than the other priors for S3 and S4. Figure S5.2 in the Supplementary Materials shows that P-HD-25 performs better than P-HD-D except in S3 where the Dirichlet prior is closest to the truth, and that $\omega^{(1)}$ tends to be underestimated under all the priors. P-HD-25 is preferred because overall it performs at least as good as the other priors P-HC and P-PC, and P-HD-25 is an intuitive and well-behaved prior that takes the hierarchical structure of the model into account.

6.3 Neonatal mortality in Kenya

This section follows the notation introduced in Section 6.1. The survey consists of 13183 households with one or more live births, distributed over 1593 clusters that are distributed over $n = 47$ counties. In total there are 376 deaths among 17664 children. Figure 8c shows the counties and the weighted neonatal mortality by the inverse inclusion probabilities, and it is unclear if there is a structured spatial pattern. The neonatal mortality is assumed to follow a survival model with constant hazard through the first month of life, and we use a latent Gaussian model with a binomial likelihood, $y_{i,j,k}|b_{i,j,k}, p_{i,j,k} \sim \text{Binomial}(b_{i,j,k}, p_{i,j,k})$, a logit link function, and a linear latent Gaussian model

$$\eta_{i,j,k} = \text{logit}(p_{i,j,k}) = \mu + x_{i,j}\beta + u_i + v_i + \nu_{i,j} + \varepsilon_{i,j,k}, \tag{6.1}$$

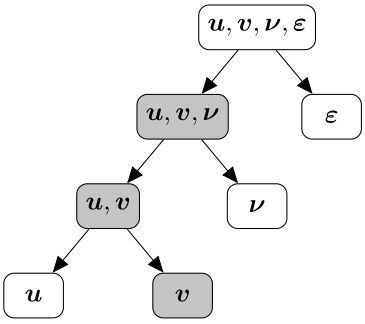
where μ is an overall intercept, β is the effect of urban, $\boldsymbol{u}$ is a Besag model with variance σ^2_{11}, $\boldsymbol{v}$ is a Gaussian i.i.d. effect of county with variance σ^2_{12}, $\boldsymbol{\nu}$ is a Gaussian i.i.d. effect of cluster with variance σ^2_2, and $\boldsymbol{\varepsilon}$ is a Gaussian i.i.d. effect of household with variance σ^2_3. In this model, $\boldsymbol{u}$ and $\boldsymbol{v}$ provide structured and unstructured, respectively, between-county variation, $\boldsymbol{\nu}$ provides between-cluster variation, and $\boldsymbol{\varepsilon}$ provides within-cluster

variation. The Besag effect has a sum-to-zero constraint to make the overall intercept identifiable. The random effects of cluster and household are necessary to account for the dependence induced between sampled households due to the clustering in the sampling design. We assume that there is no difference between the effect of urbanicity between different counties.

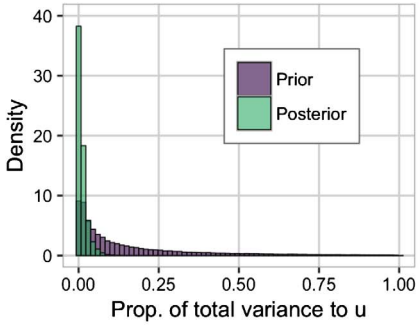
The model has four variance parameters that must be assigned a joint prior. The first step is to choose the tree structure. For simplicity's sake, the alternatives to the full model (6.1) we would entertain are first $\eta_{i,j,k} = \mu + x_{i,j}\beta + v_i$, then we would add u_i, so $\nu_{i,j}$, and at last $\varepsilon_{i,j,k}$. We prefer coarser unstructured effects over finer unstructured effects since we would like to explain the data at a coarser level if possible, and we prefer the unstructured spatial effect over the structured spatial effect since we want to reduce the risk of estimating spurious spatial signals. This gives the nested tree structure in Figure 8a where the household effect, cluster effect and Besag effect are sequentially split off from the total latent variance. We construct an HD prior based on the tree structure with PC priors with default hyperparameter values for the splits, and induce shrinkage on the total latent variance as in Prior class 3 with a PC prior where $\mathrm{P}(\text{Total variance} > 11.296) = 0.05$. This corresponds to *a priori* equal-tailed 90% credible interval of $(0.1, 10)$ for the effect of the random effects on the odds-ratio, $\exp(u_i + v_i + \nu_{i,j} + \varepsilon_{i,j,k})$. This allows for high variation in the data and is used because the data is observed at the household level. The splits in Figure 8a are given PC priors with default hyperparameters and bases models as indicated in the figure.

The model is parameterized by total standard deviation σ_{T}, and proportion of household variance to total variance of the random effects $\omega^{(1)}$, proportion of cluster variance to the sum of cluster and county variance $\omega^{(2)}$, and the proportion of structured spatial variance to county variance $\omega^{(3)}$. The priors and posteriors of the proportions $\omega^{(1)}$, $\omega^{(2)}$ and $\omega^{(3)}$ are shown in Figure 8e. The total standard deviation has a posterior median of 1.47, and the prior and posterior can be seen in Figure S5.3 in the Supplementary Materials. The results show that the data only weakly informs about the proportion of structured to unstructured spatial effects, which indicates that the data provide no strong evidence in favor of or against a structured spatial effect. Also the posterior of $\omega^{(2)}$ is similar to the prior, but there is a strong signal in the posterior of $\omega^{(1)}$ that there is non-negligible household-level dependence. A plausible explanation for the weak signals in $\omega^{(2)}$ and $\omega^{(3)}$ is that there is substantial noise coming from high variance in the household-level random effect and weak information from the Binomial likelihood due to few successes and few numbers of trials.

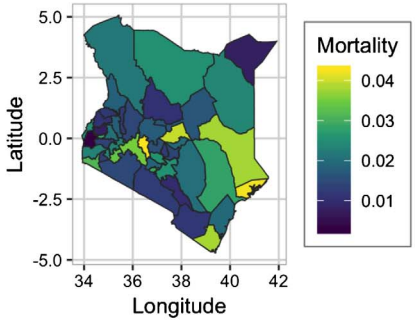
As shown in Figure 8b the proportion of the total latent variance attributed to the structured spatial effect is low and the posterior median is 0.56%. The estimated spatial effect in Figure 8d only explains a small part of the variation seen in the observed data in Figure 8c. One should be careful to draw conclusions about spatial variation based on Figure 8d because the data is only weakly informative about the split between the structured and the unstructured spatial random effects $\omega^{(3)}$, and there is only weak evidence for the spatial effect being different from 0 as shown in Figure S5.5 in the Supplementary Materials. The fact that the comparisons of priors and posteriors for $\omega^{(2)}$ and $\omega^{(3)}$ directly informs about the weak signal in the data is an advantage of



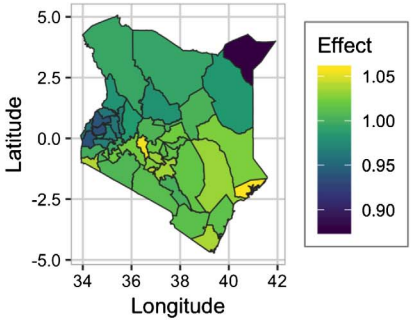
(a) Model structure.



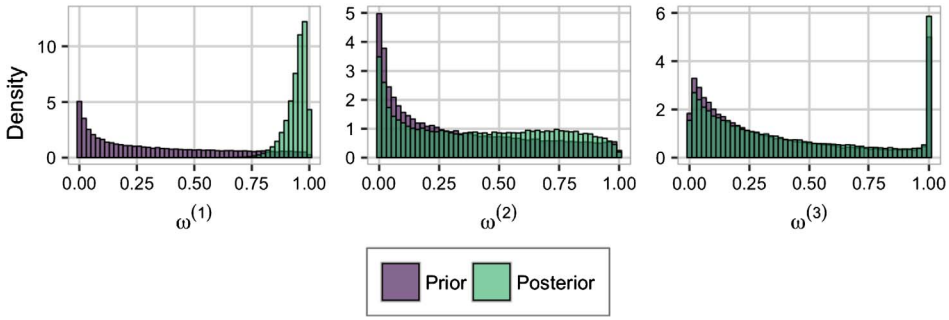
(b) Variance of $\boldsymbol{u}$ relative to total variance.



(c) Weighted average of neonatal mortality.



(d) Posterior median of e^{u}.



(e) The priors and posteriors for the proportion of household variance to total variance of the random effects $\omega^{(1)}$, the proportion of cluster variance to cluster- and household-level variance $\omega^{(2)}$, and the proportion of structured spatial variance to total between-county variance $\omega^{(3)}$.

Figure 8: Description of model structure, map of observed mortality, and results for neonatal mortality in Kenya.

the parametrization through proportions of variance, and a strong argument for setting priors on $\omega^{(2)}$ and $\omega^{(3)}$ rather than independent priors on the variance of each effect since the resulting posteriors for $\omega^{(2)}$ and $\omega^{(3)}$ are strongly dependent on the resulting implicit priors for $\omega^{(2)}$ and $\omega^{(3)}$.

One could argue for other splits in the tree in Figure 8a such as preferring finer level effects to coarser level effects because one does not want to estimate spurious cluster-level or county-level effects, but the key point of this application is that it is easy to set up the prior based on *a priori* assumptions and the assumptions are available to other scientists at a glance. With the traditional approach of independent priors, the resulting prior on the total variance of the random effects and the distribution of this total variance to the different random effects is obfuscated. Furthermore, if expert knowledge indicates that stronger relative shrinkage of the variances than the default setting is needed, the medians of the conditional priors for $\omega^{(1)}$, $\omega^{(2)}$ and $\omega^{(3)}$ can be reduced.

7 Discussion

Independent priors for the variance parameters in a BHM result in an implicit prior on the total variance of the random effects, t, and the attribution of t to the random effects. Additive models are typically built in a modular fashion, but these implicit priors are not consistent with respect to adding or removing random effects. In the case of Gaussian responses, both the prior for t and the prior for t relative to the size of the residual variance change. The proposed HD priors overcomes these shortcomings, and respect the defined model structure and are consistent for t and the attribution of t to the different random effects for different selections of random effects.

The HD priors admit a visual representation through trees that allow transparent communication of the assumptions made in constructing the priors and facilitate discussion around the assumptions. The tree clearly specifices where shrinkage has been applied, and in some cases lead to more intuitive parametrization that is more suitable for elicitation of priors. For the random intercept model, the tree-based hierarchical variance decomposition leads to a parameterisation in terms of t and the ICC. A prior on these parameters is more interpretable than separate priors on the group variance and individual variance, which obfuscates the joint effect of the priors. The increased interpretability of joint priors compared to independent priors addresses concerns raised about transparency for point processes where prior sensitivity is a major concern (Sørbye et al., 2018).

The mix of robust PC priors for shrinkage and simple Dirichlet priors for expressing ignorance, allows principled priors that respect the relative complexity of the random effects when shrinkage is necessary, and intuitive exchangeability when no random effects are preferred or no model structure is apparent. The simulation studies show that this approach performs better than a completely unstructured approach with a Dirichlet prior attributing t to the different random effects, but that Dirichlet priors perform well for subgroups of the random effects where there is no nested structure or difference in complexity.

HD priors with default settings for the hyperparameters performs well, but there are corner cases like no treatment effect in the latin square experiment and no structured spatial effect for the binomial data, which are best handled by the default `INLA` prior. However, this prior has a peak in the prior distribution for low variances and generally

performs surprisingly bad. The HD priors perform comparable to component-wise PC priors and separate half-Cauchy priors for the marginal variances. The main benefit of the HD priors over other default priors is their combination of intuitive graphical representation with robust inference that behaves well across a range of different scenarios.

The calculation of PC priors is more complex in the context of correlation parameters, but multivariate PC priors have been developed for more complex random effects such as autoregressive processes (Sørbye and Rue, 2017) and spatial Matérn models (Fuglstad et al., 2019b). These can be integrated into the HD prior framework by first defining priors on the correlation parameters, and then constructing the joint prior for the variance parameters with the correlation parameters fixed to reasonable values. This follows the pragmatic mindset of Assumption 2 of producing priors that are computationally feasible, intuitive and practically useful.

A key focus for future work is to exploit sparsity in the precision matrices of the random effects. This is important when shrinkage is desired through PC priors because many models such as random walks, Besag models, and Gaussian random fields (Lindgren et al., 2011) have dense covariance matrices, but can be expressed through sparse precision matrices. Assume that the total variance is split between random effects with sparse precision matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$, where $\mathbf{Q}_1$ corresponds to the base model. Let $0 < \omega < 1$, then the KLD used in Theorem 1 consists of the trace of $\mathbf{Q}_1[(1-\omega)\mathbf{Q}_1^{-1} + \omega\mathbf{Q}_2^{-1}]$, which can be computed quickly through the techniques in Rue and Held (2010, Section 12.1.7.10), and the determinant $\det[\mathbf{Q}_1[(1-\omega)\mathbf{Q}_1^{-1} + \omega\mathbf{Q}_2^{-1}]] = \det[(1-\omega)\mathbf{Q}_2 + \omega\mathbf{Q}_1](\det[\mathbf{Q}_2])^{-1}$, which can be computed quickly through Cholesky factorizations.

We aim to further broaden the advantages of the HD priors in the future by constructing a joint prior for the variance parameters and the fixed effects. However, this will require re-thinking of the concept of total latent variance as it is the values of the coefficients of the fixed effects and not their variance that determines the amount of variance they explain. Instead of starting with the concept of marginal variances, it is natural to begin with the classical concept of explained variance and use ideas from block-wise g-priors (Som et al., 2014) to distribute variance inside a group of covariates. In a multilevel model this would connect the attribution of explained variance to different levels to generalised coefficients of determinations. Additionally, towards non-parametric regression by including a combination of a linear effect of a covariate and a smooth effect of a covariate, and explicitly putting a prior on the degree of non-linearity (Simpson et al., 2017, Section 7). However, there are still open questions and this addition is outside the scope of this paper.

The choice of tree structure for HD priors should be guided by the application at hand, for example, by considering the relative complexity of the random effects. When expert knowledge is available, the default values for the hyperparameters should be replaced by values elicited based on expert knowledge. We believe that the advantages of the HD priors over independent priors mean that they should be used as the default option in software for Bayesian analysis. However, it is necessary to make the selection and computation of HD prior for a specific problem easier for analysts. We plan to address this by providing a separate `R` package, which is compatible with `INLA`, that

provides a graphical user interface for selecting the tree structure and selecting priors for the splits, and has the option to pre-compute priors for use in `RStan`. This will allow analysts to experiment with different *a priori* assumptions and produce graphical figures that summarize their assumptions and can be communicated to fellow scientists. This will encourage transparancy and clarity in *a priori* assumptions in the scientific community.

Supplementary Material

Supplement to "Intuitive joint priors for variance parameters"
(DOI: 10.1214/19-BA1185SUPP; .zip). The Supplementary Materials consist of a supplementary document providing additional results and discussion, and example code for the latin square model. The code is described in the Section S4.3 of the supplementary document. http://www.some-url-address.org/dowload/0000.zip

References

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., and Lindgren, F. (2018). "Spatial modeling with R-INLA: A review." *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6): e1443. MR3873676. doi: https://doi.org/10.1002/wics.1443. 2

Balakrishnan, N. and Nevzorov, V. B. (2003). *A primer on statistical distributions*. Hoboken, NJ: John Wiley & Sons. MR1988562. doi: https://doi.org/10.1002/0471722227. 3

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC. MR3362184. 1

Besag, J., York, J., and Mollié, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43(1): 1–20. MR1105822. doi: https://doi.org/10.1007/BF00116466. 7

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. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 3

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. West Sussex, United Kingdom: John Wiley & Sons. MR3364017. 2

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). 2, 14

Cicchetti, D. V. (1994). "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." *Psychological assessment*, 6(4): 284. 9

Fahrmeir, L. and Lang, S. (2001). "Bayesian inference for generalized additive mixed models based on Markov random field priors." *Journal of the Royal Statistical Society: Series C*, 50(2): 201–220. MR1833273. doi: https://doi.org/10.1111/1467-9876.00229. 2

Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H. , and Riebler, A. (2019a). "Supplement to "Intuitive joint priors for variance parameters"." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1185SUPP. 4

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019b). "Constructing priors that penalize the complexity of Gaussian random fields." *Journal of the American Statistical Association*, 114(525): 445–452. MR3941267. doi: https://doi.org/10.1080/01621459.2017.1415907. 8, 24

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 2, 9

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC. MR3235677. 1

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, volume 1. New York, New York: Cambridge University Press. 1, 2, 14

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008). "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics*, 2(4): 1360–1383. MR2655663. doi: https://doi.org/10.1214/08-AOAS191. 4, 9

Gelman, A., Simpson, D., and Betancourt, M. (2017). "The prior can often only be understood in the context of the likelihood." *Entropy*, 19(10): 555. 2, 12

General Assembly of the United Nations (2015). "Resolution adopted by the General Assembly on 25 September 2015." A/RES/70/1. 18

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477): 359–378. MR2345548. doi: https://doi.org/10.1198/016214506000001437. 17

Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D. C., et al. (2017). "Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals." *The Lancet*, 390(10108): 2171–2182. 18

Guo, J., Riebler, A., and Rue, H. (2017). "Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors." *Statistics in Medicine*, 36(19): 3039–3058. MR3670407. doi: https://doi.org/10.1002/sim.7313. 8

Hinkelmann, K. and Kempthorne, O. (1994). *Design and Analysis of Experiments,*

Volume 1: Introduction to Experimental Design. John Wiley & Sons. MR2129060. doi: https://doi.org/10.1002/0471709948. 16

Holand, A. M., Steinsland, I., Martino, S., and Jensen, H. (2013). "Animal models and integrated nested Laplace approximations." *G3: Genes, Genomes, Genetics*, g3-113. 2

Jordan, A., Krüger, F., and Lerch, S. (2017). "Evaluating probabilistic forecasts with the R package scoringRules." *arXiv preprint* arXiv:1709.04743. 17

Kenya National Bureau of Statistics, Ministry of Health/Kenya, National AIDS Control Council/Kenya, Kenya Medical Research Institute, and National Council for Population and Development/Kenya (2015). *Kenya Demographic and Health Survey 2014*. Rockville, MD, USA. URL http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf. 18, 19

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilio, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations using R and INLA*. Boca Raton, FL: CRC press. Github version www.r-inla.org/spde-book. 2

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005). "How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS." *Statistics in Medicine*, 24(15): 2401–2428. MR2151713. doi: https://doi.org/10.1002/sim.2112. 2

Li, Z., Hsiao, Y., Godwin, J., Martin, B. D., Wakefield, J., Clark, S. J., et al. (2019). "Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa." *PloS one*, 14(1): e0210645. 18

Lindgren, F. and Rue, H. (2015). "Bayesian spatial modelling with R-INLA." *Journal of Statistical Software*, 63(19): 1–25. MR2490553. doi: https://doi.org/10.1016/S0169-7161(05)25033-2. 2

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4): 423–498. MR2853727. doi: https://doi.org/10.1111/j.1467-9868.2011.00777.x. 24

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). "The BUGS project: Evolution, critique and future directions." *Statistics in Medicine*, 28(25): 3049–3067. MR2750401. doi: https://doi.org/10.1002/sim.3680. 2

Martinez-Beneito, M. A. (2013). "A general modelling framework for multivariate disease mapping." *Biometrika*, 100(3): 539–553. MR3094436. doi: https://doi.org/10.1093/biomet/ast023. 2

McGraw, K. O. and Wong, S. P. (1996). "Forming inferences about some intraclass correlation coefficients." *Psychological methods*, 1(1): 30. 2

Plummer, M. (2017). "JAGS version 4.3. 0 user manual [Computer software manual]." *Retrieved from* sourceforge.net/projects/mcmc-jags/files/Manuals/4.x. 2

Polson, N. G. and Scott, J. G. (2010). "Shrink globally, act locally: Sparse Bayesian regularization and prediction." *Bayesian statistics*, 9: 501–538. MR3204017. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. 3

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). "An intuitive Bayesian spatial model for disease mapping that accounts for scaling." *Statistical Methods in Medical Research*, 25(4): 1145–1165. MR3541089. doi: https://doi.org/10.1177/0962280216660421. 3, 8, 19

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Boca Raton, Florida: CRC press. MR2130347. doi: https://doi.org/10.1201/9780203492024. 7, 11

Rue, H. and Held, L. (2010). "Discrete Spatial Variation." In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (eds.), *Handbook of Spatial Statistics*, Handbooks of Modern Statistical Methods, chapter 12, 171–200. Boca Raton, FL: CRC Press. MR2730942. doi: https://doi.org/10.1201/9781420072884-c12. 24

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society: Series B*, 71(2): 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 2

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). "Bayesian computing with INLA: A review." *Annual Review of Statistics and Its Application*, 4(1): 395–421. MR3634300. doi: https://doi.org/10.1214/16-STS576. 2

Shen, K.-K., Doré, V., Rose, S., Fripp, J., McMahon, K. L., de Zubicaray, G. I., Martin, N. G., Thompson, P. M., Wright, M. J., and Salvado, O. (2016). "Heritability and genetic correlation between the cerebral cortex and associated white matter connections." *Human brain mapping*, 37(6): 2331–2347. 9

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). "Penalising model component complexity: a principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 2, 4, 8, 13, 24

Som, A., Hans, C. M., and MacEachern, S. N. (2014). "Block hyper-g priors in Bayesian regression." *arXiv preprint* arXiv:1406.6419. MR3321977. 2, 24

Sørbye, S. H., Illian, J. B., Simpson, D. P., Burslem, D., and Rue, H. (2018). "Careful prior specification avoids incautious inference for log-Gaussian Cox point processes." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68: 543–564. MR3937462. doi: https://doi.org/10.1111/rssc.12321. 23

Sørbye, S. H. and Rue, H. (2014). "Scaling intrinsic Gaussian Markov random field priors in spatial modelling." *Spatial Statistics*, 8: 39–51. MR3326820. doi: https://doi.org/10.1016/j.spasta.2013.06.004. 7

Sørbye, S. H. and Rue, H. (2017). "Penalised complexity priors for stationary autoregressive processes." *Journal of Time Series Analysis*, 38(6): 923–935. MR3714116. doi: https://doi.org/10.1111/jtsa.12242. 24

Sørbye, S. H. and Rue, H. (2018). "Fractional Gaussian noise: Prior specification and model comparison." *Environmetrics*, 29(5–6): e2457. 8

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). "BUGS 0.5* Examples Volume 2 (version ii)." *MRC Biostatistics Unit*. 2

Stan Development Team (2018a). "RStan: the R interface to Stan." R package version 2.18.1. URL http://mc-stan.org/. 2, 14

Stan Development Team (2018b). "Stan Modeling Language Users Guide and Reference Manual, version 2.18.0." *Technical report*. URL http://mc-stan.org. 2

StataCorp (2017). *Stata Bayesian analysis, Reference manual*. StataCorp LLC, College Station, TX, 15 edition. 2

Wakefield, J. (2006). "Disease mapping and spatial regression with count data." *Biostatistics*, 8(2): 158–183. 2

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2019). "Estimating under-five mortality in space and time in a developing world context." *Statistical Methods in Medical Research*, 28(9): 2614–2634. MR4000184. doi: https://doi.org/10.1177/0962280218767988. 18