

Doctoral thesis

Doctoral theses at NTNU, 2021:215

Ingeborg Gullikstad Hem

Robustifying Bayesian Hierarchical Models Using Intuitive Prior Elicitation

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences



Norwegian University of
Science and Technology

Ingeborg Gullikstad Hem

Robustifying Bayesian Hierarchical Models Using Intuitive Prior Elicitation

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU

Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

© Ingeborg Gullikstad Hem

ISBN 978-82-326-6915-8 (printed ver.)
ISBN 978-82-326-6411-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2021:215

Printed by NTNU Grafisk senter

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of Philosophiae Doctor (PhD) at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. The research was funded by The Research Council of Norway, and the work was carried out at the Department of Mathematical Sciences during the years 2017–2021.

First, I want to say thank you so much to my supervisors, Geir-Arne Fuglstad and Andrea Riebler, for many great discussions and amazing guidance throughout my PhD. You both know so much, and I have not met a single problem you could not help me solve. I have really enjoyed working with you these four years.

I would also like to say thank you to the technical staff at the department. Numerical experiments and simulations have been a large part of my PhD, which I would not have been able to carry out without their support.

Thank you to all my colleagues, both past and present, for making the work with my thesis a really fun experience. I am grateful for my family and friends for always being supportive and for making my everyday life wonderful. I really appreciate you all. Last, but not least, I am especially grateful for Tale, you have been a huge support these years. Tusen takk!

Ingeborg Gullikstad Hem
Trondheim, March 2021

Contents

1	Introduction	1
1.1	Bayesian statistics	3
1.2	Prior knowledge	5
1.3	The challenge of choosing prior distributions	7
1.4	Component-wise variance priors	8
1.4.1	Application-based priors	9
1.4.2	Principle-based priors	11
1.5	Model-wise variance priors	14
1.5.1	Distributing the total variance along a prior tree	14
1.5.2	Priors for the total variance	16
1.5.3	Priors for variance proportions	17
1.6	Example: Random intercept model	18
1.7	Bayesian inference	20
1.8	Discussion	22
2	Scientific papers	25

Chapter 1

Introduction

Bayesian modelling has over the recent years had a raise in popularity among both statisticians and applied scientists through easy-to-use software (e.g., van de Schoot et al., 2021). They allow for fitting complex models, but lack thorough explanations on how to choose prior distributions for these models. There is an increase in recommendations about being sceptical to default prior distributions, i.e., general-purpose priors chosen without any specific application in mind (e.g., Gelman et al., 2017, 2020; Smid and Winter, 2020; Stan Development Team, 2021a). However, it can be difficult to navigate through the huge amount of proposed priors. Everyone comes with their own suggestions, and it can be difficult to transform the prior knowledge, when it exists, into a prior distribution function.

This thesis presents the new prior framework *hierarchical decomposition (HD) priors*. The HD prior framework makes it possible to specify joint priors for variance parameters in Bayesian hierarchical models that are intuitive, easily communicated, and robust in the sense that it leads to stable inference and avoids estimating spurious random effects. It offers a simple way of including prior intuition and knowledge in the prior, and is a tool for easy communication of the prior. With this framework we can simplify the process of eliciting prior knowledge and choosing a good prior, reflecting our prior beliefs on the scale of the prior knowledge. The penalized complexity (PC) prior of Simpson et al. (2017) enables shrinkage to stabilize the inference and avoid overfitting. We extend the idea of the PC prior to the model level. Through a combination of

default settings and prior knowledge the HD prior framework yields a weakly-informative prior (Gelman et al., 2008; Simpson et al., 2017). The framework is focused on variance parameters. Including other parameters such as correlations is future work, and fixed effects are given vague Gaussian priors.

Often, one does not have good intuition about the absolute magnitudes of each variance parameter in the model. In such cases it may be easier to elicit information about the total variation in the data attributed to the random effects, and about the relative magnitudes of the individual variance parameters. This is the idea behind the HD prior framework. For example, in genomic modelling, experts in the field typically have prior knowledge about the phenotypic (total) variance and the heritability. Heritability is the proportion of phenotypic variance for some trait that can be attributed to the effect of the genes (Allenby et al., 1995; Holand et al., 2013; Mäki-Tanila and Hill, 2014). This prior knowledge can come from for example previous analyses on similar datasets, or intuition and knowledge about the species of interest. In the HD prior framework, we consider the total variation accounted for by the random effects, as we keep the fixed effects out of the framework.

The HD prior provides a flexible parameterization of the joint variance prior that is not restricted to the variance parameters, but can be customized to coincide with the available prior knowledge. This yields a transparent and easily communicated prior, making the elicitation process simpler (O’Hagan et al., 2006). For the genomic model where we have knowledge about the phenotypic variance and heritability, a prior parameterized with phenotypic variance and heritability is more intuitive than one with two independent variance parameters. In addition to being proper, the prior provides an option to easily investigate how much the prior influences the posterior. The latter indicates whether the data is able to provide information, or if the model is overfitting.

To evaluate the performance of the HD prior, it is applied in an extensive simulation study of wheat breeding, where one of the main goals is to identify the genetically best individuals in a population, and is shown to outperform both the standard maximum likelihood approach and independent priors on variance parameters. To make the HD prior easily available and applicable, a software is developed, which encourages to integrate prior choice as an active part of the Bayesian workflow.

The main contributions of this thesis are a novel framework for specifying joint variance priors in Bayesian hierarchical models, demonstrating the added value of incorporating expert knowledge in plant breeding programs, and an R

package that opens for easy inclusion of prior knowledge in an intuitive and transparent way. These contributions are summarized in Chapter 2 and are followed by the scientific papers. Details around aspects from and background for the enclosed papers are given in Chapter 1. First, a brief introduction to Bayesian statistics is given, before moving on to prior distributions. Some challenges, distribution families, and various approaches and methods are presented, and Chapter 1 ends with a short discussion.

1.1 Bayesian statistics

Thomas Bayes (1701–1761) was the origin of Bayesian statistics, and formulated what is now well-known as *Bayes' theorem* (Bayes, 1763). His work was communicated by Richard Price and published after his death. The famous theorem describes the posterior probability of some event A based on *a priori* knowledge relevant for the event, and states:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where B is an observed event and $P(B) \neq 0$ (see e.g. Casella and Berger, 2002). The *a priori* knowledge about the event is formulated through $P(A)$, and the theorem gives information on this event A given the observations done on event B .

The term “inference” covers statistical methods where one wants to learn something about a population from a sampled subset of the population. Through an analysis, we want to learn properties of some underlying probability distribution from observed data. We use a statistical model consisting of an observation model and a latent model describing the observed data for the population subset, and want to use this model to gain knowledge about the whole population. Bayesian inference is a statistical inference method where we strengthen the statistical model with beliefs, using new information and evidence to update the probability of some hypothesis.

The following formulation of Bayes' theorem is more relevant for this thesis:

$$\pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})}, \quad (1.1)$$

where $\mathbf{y} = (y_1, \dots, y_n)$ is observed data for some phenomenon, $\mathbf{x} = (x_1, \dots, x_n)$

is an underlying process describing these data, and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ are parameters describing the underlying process $\pi(\mathbf{x}|\boldsymbol{\theta})$ and observation model, also known as the likelihood, $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. $\pi(\mathbf{y})$ is the marginal likelihood. Prior knowledge about the parameters is included in the model through the *prior distribution* $\pi(\boldsymbol{\theta})$. This model is known as a *Bayesian hierarchical model* (BHM, Cressie and Wikle, 2011) and represents a flexible and widely used class of models (e.g., Gelman and Hill, 2007; Gelman et al., 2013; Banerjee et al., 2014). Models where the underlying, or latent, process \mathbf{x} consists of additive combinations of model effects are called additive models, and the useful *latent Gaussian models* are a subset of these (Rue et al., 2009). In a latent Gaussian model the individual model components in \mathbf{x} are all Gaussian conditional on the parameters $\boldsymbol{\theta}$, and the data is assumed to come from an exponential family of distributions such as the Gaussian or Poisson distributions (see e.g. Casella and Berger (2002, Chapter 3.4) for details). The latent Gaussian models have a latent process that can be formulated as:

$$\eta_i = \mu + \sum_{j=1}^p \beta_j z_{i,j} + \sum_{k=1}^m f_k(v_{i,k}), \quad i = 1, \dots, n, \quad (1.2)$$

where η_i is commonly known as the linear predictor due to its additive properties. The linear predictor is linked to the observations y_i through a link function $g(\cdot)$ such that the mean μ_i of observation y_i can be modelled through $\eta_i = g(\mu_i)$. In Equation (1.2), μ represents an intercept, β_j , $j = 1, \dots, p$, represents coefficients of covariates $\mathbf{z}_j = (z_{1,j}, z_{2,j}, \dots, z_{n,j})$ (fixed effects), and $\{f_k(\cdot)\}$, $k = 1, \dots, m$, are unknown functions of covariates $\mathbf{v}_k = (v_{1,k}, v_{2,k}, \dots, v_{n,k})$ (random effects). The coefficients β_j are assigned Gaussian distributions $\mathcal{N}(\mu_{\beta_j}, \sigma_{\beta_j}^2)$ where μ_{β_j} is commonly set to 0, and $\sigma_{\beta_j}^2$ is typically large. The same goes for the intercept μ .

In this thesis, the latent Gaussian models are considered, and the main focus is set on the properties of the prior distribution $\pi(\boldsymbol{\theta})$, more specifically the prior distributions on the variance parameters. The intercept μ and coefficients β_j are typically easy to identify (Goel and Degroot, 1981), and we do not consider properties of priors on them. We limit the scope to models where the random effects are multivariate Gaussian conditional on the variance parameter σ_k^2 , such that $\{f_k(\mathbf{v}_k)\} \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \boldsymbol{\Sigma}_k)$. $\boldsymbol{\Sigma}_k$ is the known covariance matrix of the effect with corresponding variance σ_k^2 . This covariance matrix can be dense or sparse. When the covariance matrix is the identity matrix, a special case of a sparse matrix, the effect is unstructured, such as a residual effect. We can set $\{f_k(\mathbf{v}_k)\} = \mathbf{u}_k = (u_{1,k}, u_{2,k}, \dots, u_{n,k})$ and get a simplified version of the linear

predictor in Equation (1.2):

$$\eta_i = \mu + \sum_{j=1}^p \beta_j z_{i,j} + \sum_{k=1}^m u_{k,i}. \quad (1.3)$$

Of note, if $\sigma_k^2 = 0$, the random effect \mathbf{u}_k is removed from the model in the sense that it does not contribute. We assign prior distributions to the variance parameters σ_k^2 of the random effects. These priors cannot be Gaussian since $\sigma_k^2 \geq 0$.

1.2 Prior knowledge

“Prior knowledge” is a loose term and can include everything from ideas, beliefs and intuitions to specific and highly detailed information about a phenomenon, before the phenomenon itself has happened or is investigated. Prior knowledge exists to some extent for more or less any situation, such as genomics (e.g., Dougherty and Dalton, 2013), medical image segmentation (e.g., Grau et al., 2004), weather prediction (e.g., Lorenc, 1986), and marketing (Allenby et al., 1995), and can be used to improve the analysis. This knowledge both can and should be utilized in a Bayesian framework through the prior distribution (O’Hagan et al., 2006). This is one of the great advantages with Bayesian inference; we can use the prior to express our prior beliefs. However, to transform an intuition into a probability distribution is not straight-forward; it is difficult to get the prior to express these prior beliefs we have. It is not obvious how to use single numerical values, or perhaps just an estimate or idea, to construct a prior distribution. Dallow et al. (2018) combines prior knowledge from several experts, “team beliefs”, and make a prior distribution, providing transparency about the beliefs and priors through a robust framework, and there is an increasing interest in using prior and expert knowledge in analyses through prior distributions, such as in social science (Zondervan-Zwijnenburg et al., 2017). Prior elicitation in a transparent way is today a relevant topic.

A prior distribution can be constructed in many ways. One can use prior knowledge from similar experiments performed in the past, or use the intuition and knowledge from one or several experts in the field of interest that expresses detailed and subjective information about the problem at hand (e.g., Spiegelhalter et al., 2004). One can use vague or weakly-informative priors that says something, but not much (e.g., van de Schoot et al., 2021). One can use

so-called non-informative prior distributions that contain objective information that does not use any prior knowledge at all (e.g., Gelman, 2006). One can use general principles developed independently of the application at hand (e.g., Kass and Wasserman, 1996). One can utilize conjugacy to choose a prior that will simplify and increase the efficiency of the computation (e.g., Gelman et al., 2013). One can use a uniform prior to restrict the parameter space assigning equal probability to all values in the chosen space (Lambert et al., 2005). All these approaches let us take advantage of pre-existing knowledge, also the non-informative priors. If we indeed know absolutely nothing about the problem at hand, we should use a prior that reflects that, and not use some default prior which may say something about the parameter you do not have grounds to assert. Gelman et al. (2020) points out that a uniform prior is often classified as non-informative, but depends on the parameterization of the model, and thus it contains some information (see also Lambert et al., 2005).

The need for a good prior distribution will decrease with increasing amounts of data, however, how much data required before vague priors will be sufficient depends on the complexity of the model and number of model parameters (Gelman et al., 2020). A simple model requires less data to tolerate vague priors than a complex model with many effects and parameters that must be estimated. An example of this is in quantitative genetics: the data needed for accurate estimation of nonadditive genetic effects is huge. The nonadditive effects are often confounded, and we seldom have enough observed data to estimate these effects without an informative prior (Sorensen and Gianola, 2007).

There is no one ultimate answer to the question about how to perform Bayesian inference - it varies with among others the problem at hand, model complexity, goal of analysis and prior knowledge. Gelman et al. (2020) has put together a comprehensive guide to Bayesian workflow. We do not go into details, but want to point out some of their aspects related to prior distributions. They stress that awareness and being critical of the decisions made in the model fitting process is important. The joint prior should be considered to ensure it does not become more informative than intended as many weakly-informative component-wise priors may lead to a much more informative joint prior. They argue that prior information can solve computational problems in the inference, and that priors must be considered for each model in the Bayesian workflow. If you change the model, you may need to adjust your prior.

1.3 The challenge of choosing prior distributions

The process in all kinds of inference starts with a research question one wants to answer. To answer this question, data are collected in the form of observations, one decides on a model, what components it should contain and on what likelihood distribution is suitable for the available data. It can also be done the other way around: data are collected after one has decided on what model is suitable for the problem at hand. In Bayesian inference, the next step is to choose the prior distributions, and then the question becomes: “Which prior distribution do I use for my model?”.

This is not an easy question we can answer with just a sentence. As pointed out by Gelman et al. (2020), this is a part of an iterative process. We do not consider the whole Bayesian workflow from the beginning to the end, but focus on the prior distribution part. There is no single recipe describing how to choose prior distributions, and there is no mutual agreement on which distributions are better or worse in general (e.g., Lambert et al., 2005; Gelman, 2006; Gelman et al., 2017). As indicated in Gelman et al. (2017), there are no general-purpose priors that suits every model and application. The question on which prior to use becomes even more complex for non-variance parameters, such as correlation parameters, which are even more important to assign good priors as they are further away from the data and it may thus be little information in the data about them (Goel and Degroot, 1981). This is however outside the scope of the thesis. In addition, the probability distribution itself should have favourable properties to ensure stable inference and avoid overfitting, such as how much mass there is in the tails or around the mode of the distribution. that the prior should be independent of the observed data, and be chosen before the data is seen.

Bayesian modelling has become increasingly more popular the recent years, especially through software such as Bayesian Analysis Toolkit (BAT, Caldwell et al., 2009), OpenBUGS (Lunn et al., 2009), Template Model Builder (TMB, Kristensen et al., 2016), JAGS (Plummer, 2017), Integrated Nested Laplace Approximations (INLA, Rue et al., 2009) through the R package INLA (see www.r-inla.org), Stan (Carpenter et al., 2017) through the R package `rstan` (Stan Development Team, 2020) and Stan extensions such as `rstanarm` (Goodrich et al., 2020) and `loo` (Goodrich et al., 2020), and more (see e.g., van de Schoot et al., 2021). Each software comes with default prior distributions for variance parameters. For example, the default in INLA is an inverse-gamma

distribution with shape 1 and scale 10^{-5} ($\text{InvGam}(1, 5 \cdot 10^{-5})$) (Blangiardo and Cameletti, 2015) (this corresponds to a gamma prior on inverse variance with shape 1 and rate 10^{-5}), and TMB has “non-informative priors in the Bayesian literature” as default (Kristensen et al., 2016). `rstan` has implicit priors that are uniform on the specified range of the parameter (Stan Development Team, 2018) which are improper if the parameter is allowed to take any value along the real line, however, the developers discourage the use of such very vague priors (Stan Development Team, 2021a). Bounded uniform priors have been investigated on both variances (Lambert et al., 2005) and standard deviations (Martinez-Beneito, 2013). Other proposed prior distributions include Half-Cauchy(25) on standard deviations (Gelman, 2006), and WinBUGS, OpenBUGS, JAGS and Stata used $\text{InvGam}(\varepsilon, \varepsilon)$ priors in their old examples and manuals (Spiegelhalter et al., 1996; Plummer, 2017; StataCorp, 2017). These distributions can provide computational advantages due to conjugacy, but they are generally inappropriate for variances of random effects (Lunn et al., 2009). Prior distributions are in other words a highly debated topic, but no overall conclusion on what prior distribution is the right one in general exist.

When choosing the default prior from some software or blindly using generic prior distributions found in the literature, the properties of the Bayesian framework are not fully utilized, and prior knowledge may go to waste. It is also difficult to know how this default or literature-based prior contributes to the inference without extensive testing of the model, and it may even contradict the prior knowledge (which with a default prior is not used). An inappropriate prior distribution can lead to slower inference than necessary because the posterior is difficult to explore, to overfitting, or it can lead to unstable or even failing inference.

1.4 Component-wise variance priors

The common approach is to choose individual prior distributions for each variance parameter σ_k^2 (see model description in Section 1.1) in the model in a component-wise fashion. This leads to a joint prior distribution

$$\pi(\sigma_1^2, \dots, \sigma_m^2) = \pi(\sigma_1^2) \times \dots \times \pi(\sigma_m^2)$$

where the prior on each variance is independent of the others. If prior knowledge about the absolute sizes of each variance is available, individual prior distribu-

tions on the variance parameters is a satisfactory solution. However, we are seldom in a situation where we have such specific information.

It is in general difficult to choose prior distributions for variance parameters (Fong et al., 2010), and independent component-wise variance priors have several properties that can lead to problems. First of all, it is problematic to exploit prior knowledge unless the knowledge exists on the scale of the variance. This is not the case in for example animal models, where the available prior information and knowledge often is on the heritability (Holand et al., 2013), or in disease mapping where it makes sense to elicit prior knowledge on the total random effect variances (Wakefield, 2006).

Further, challenges arise when we change the linear predictor, by adding or removing a model component, and with this get a new model. If we already had chosen a prior that reflects our prior beliefs on how much variance each model effect accounted for, we now need to restructure the prior so the new model match our beliefs. This problem increases when default priors for every variance parameter are used. By using the same prior distribution for all variances, one expresses that they account for an equal amount of the total variance in the data. However, when a model component is added or removed, the change in the prior reflects that the amount of variance in the observed data increases or decreases, which is seldom the intention, and is with default priors something that is not always considered.

As the total amount of variation in the data is the sum of the individual variance parameters for each model effect, it is an advantage to have some idea on what the total amount of variation the prior indicates. Unfortunately, we are not guaranteed that the sum of the distributions is a known distribution family we can interpret in a simple way. For example, the sum of two inverse-gamma distributed random variables is in general not inverse-gamma itself. There is potential for making the process of choosing priors for variance parameters simpler, more transparent and more intuitive, and at the same time constructing a prior that robustifies the modelling.

1.4.1 Application-based priors

A variance parameter σ^2 is always non-negative, and thus needs a prior distribution with zero mass for negative values on the variance scale. A range of probability distributions fulfills this requirement. It should have support for the variance being 0 to avoid overfitting. Note that distributions over the real line

can be used on log-variances, but this may make interpretation of the prior less intuitive.

The inverse-gamma distribution with shape $\alpha > 0$ and scale $\beta > 0$, given by

$$\text{InvGam}(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2} \right)^{\alpha+1} e^{-\beta/\sigma^2}, \quad \sigma^2 > 0,$$

is a popular prior choice for variances in Bayesian modelling with Gaussian data due to the conjugacy properties of this distribution. $\Gamma(\cdot)$ denotes the gamma function. An inverse-gamma distribution with shape α and scale β on the variance is equivalent to a gamma distribution with shape α and inverse-scale (rate) β on the inverse variance, commonly known as precision. Consider a model with likelihood $\pi(y|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$ where we use an inverse-gamma prior $\pi(\sigma^2) = \text{InvGam}(\alpha, \beta)$ and assume μ is known. For simplicity we set $\mu = 0$. We can compute the posterior distribution analytically:

$$\begin{aligned} \pi(\sigma^2|y) &= \frac{\pi(y|\sigma^2)\pi(\sigma^2)}{\pi(y)} \propto \frac{1}{\sigma} e^{-\frac{y^2}{2\sigma^2}} \left(\frac{1}{\sigma^2} \right)^{\alpha+1} e^{-\beta/\sigma^2} \\ &= \left(\frac{1}{\sigma^2} \right)^{\alpha+1+1/2} e^{-\frac{y^2/2+\beta}{\sigma^2}}, \quad \sigma^2 > 0. \end{aligned}$$

$\pi(y)$ is the marginal likelihood. This corresponds to $\pi(\sigma^2|y) = \text{InvGam}(\alpha + 1/2, y^2/2 + \beta)$, and the posterior becomes the prior with updated parameters and can be regarded as a prior for the next observation. For $\mathbf{y} = (y_1, \dots, y_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\sigma^2 \sim \text{InvGam}(\alpha, \beta)$ this generalizes to $(\sigma^2|\mathbf{y}) \sim \text{InvGam}(\alpha + n/2, \mathbf{y}^2/2 + \beta)$. Unstructured random effects in latent Gaussian models are assumed to follow a $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ distribution, and it is thus a natural choice to use an inverse-gamma distribution as the prior for the variance parameter. The same holds for multivariate Gaussian distributed data with structured covariance matrices. This relationship offers computational advantages, since the posterior is a known distribution family and one do not need to compute it. However, this prior has no support in 0, meaning the variance is forced to be positive, implying that the corresponding random effect is forced to be present in the model. This can lead to overfitting, and shows that the prior distributions used should be thoroughly inspected before performing inference.

Gelman (2006) suggest the heavy-tailed Half-Cauchy distribution on standard deviations, which is given by

$$\text{Half-Cauchy}(\sigma; \lambda) = \frac{2\lambda}{\pi(\lambda^2 + \sigma^2)}, \quad \sigma \geq 0,$$

for a scale $\lambda > 0$. This heavy tails allows the parameter, in this case the standard deviation, to be very large, and has mode in $\sigma = 0$.

A uniform prior distribution on an interval $[0, \sigma_{\text{MAX}}^2]$ is yet another option. It assigns equal preference to all values in the range, and values outside the range is given 0 probability.

All these priors require hyperparameters $(\alpha, \beta, \lambda, \sigma_{\text{MAX}}^2)$. To choose them can be a challenge, as it is difficult to have intuition on what value of, for example, the shape α gives a prior matching the existing prior knowledge. Priors based on principles can ease this process, as the principles themselves can aid the user in choosing hyperparameters.

1.4.2 Principle-based priors

Some priors are based on principles, instead of being chosen with a specific application or model in mind. Such priors are often a safer choice than literature-based ones, as they can be subjective in the sense of hyperparameters, while the idea of the prior is objective and it is developed independent of any model and situation. We mention the reference priors (Berger et al., 2009) with the special case Jeffreys' prior (Jeffreys, 1946) and the newly-developed penalized complexity priors (Simpson et al., 2017). These priors are constructed on a general model parameter and then transformed to the parameter of interest, and are applicable on any kind of model parameter.

Reference priors The reference prior is based on a principle stating that the prior should be dominated by the data as much as possible, and the posterior should be influenced as little as possible by the prior (Berger et al., 2009). The prior should have minimal information about the parameter. This is the same as maximizing the distance between the prior and posterior, so the posterior changes as much as possible when data is observed. As the prior naturally does not depend on the observed data, the expectation of this distance conditional on the model is maximized to construct the prior.

The reference prior is considered to be a non-informative, or objective, prior, as it aims to not affect the inference more than necessary. However, Gelman et al. (2020) argue that a non-informative prior does not generally exist, and the reference prior is one of these priors that are often taken as non-informative, but does in fact contain some information. The prior depends on the model

and on some assumed asymptotic properties of the soon-to-be observed data, meaning it is not completely without information. This model-dependence of the reference prior means it needs to be computed for each new application, such as the addition of a new covariate, which might be difficult.

Jeffreys' prior When we consider only one single parameter, the reference prior reduces to Jeffreys' prior (Jeffreys, 1946). Jeffreys' prior is a scale-independent, objective prior with density proportional to the square root of the determinant of the Fisher information matrix. It is useful for scale parameters, such as a variance, as it is invariant under reparameterization. Jeffreys' prior is often classified as a non-informative, or objective, prior where the application and prior knowledge is not taken into account. However, as it is a special case of a reference prior, the prior depends upon the model. Jeffreys' prior can be improper and should be used with care. It may lead to improper posteriors (e.g., Wakefield, 2006; Fong et al., 2010) and give misleading results if the impropriety is not discovered (Hobert and Casella, 1996).

For a variance parameter σ^2 , Jeffreys' prior is given by

$$\pi(\sigma^2) \propto 1/\sigma^2, \sigma^2 > 0.$$

and is improper without support in 0. This distribution is useful when we do not want to say anything about the scale of the variance parameter.

Penalized complexity priors The penalized complexity (PC) priors proposed by Simpson et al. (2017) are priors based on four general principles. They aim to avoid overfitting through penalizing a model that is more complex than there is support for in the data. This idea is in line with Occam's razor: the simplest explanation is probably the right one (Merriam-Webster, 2021), or in this case: the simplest model is preferred until the data tells otherwise. Since the PC prior is based on principles, rather than on application-specific information, it is weakly-informative, and can be tuned for the problem at hand. The idea is to define a *base model*, which for each model parameter ξ is a simpler version of the model. For example, if ξ is a variance parameter, a natural base model is a model where the variance is 0. The flexible extension of this model is then a model where the variance is different from 0. We summarize the four principles behind the PC prior, and refer to Simpson et al. (2017) for further details.

As we through the concept of Occam’s razor prefer simpler models over more complex models, deviating from the simple model should be penalized. The penalty is put on a measure of distance between the simple base model and the flexible extension. Simpson et al. (2017) suggest to use the Kullback-Leibler divergence (KLD, Kullback and Leibler, 1951) to compute this distance, which is defined as

$$\text{KLD}(\pi(\mathbf{x}|\xi)||\pi(\mathbf{x}|\xi = 0)) = \int \pi(\mathbf{x}|\xi) \log \left(\frac{\pi(\mathbf{x}|\xi)}{\pi(\mathbf{x}|\xi = 0)} \right) d\mathbf{x}, \quad (1.4)$$

for a model $\pi(\mathbf{x}|\xi)$ where $\xi = 0$ indicates the base model. The KLD is transformed to a more interpretable distance $d(\xi) = \sqrt{2\text{KLD}(\pi(\mathbf{x}|\xi)||\pi(\mathbf{x}|\xi = 0))}$, and measures the complexity of the model with varying ξ when compared to the model with $\xi = 0$. Then a prior is assigned to the distance, instead of directly to the parameter of interest ξ . Following Simpson et al. (2017), we choose a constant rate penalization for the distance d , which is achieved by using an exponential prior distribution $\pi(d) = \lambda \exp(-\lambda d)$.

We determine the hyperparameter $\lambda > 0$ (rate) with information from the user. The user must prior have intuition on the size of ξ , or on some property of the corresponding model component. This is related to the concept of weakly-informative priors. The prior information comes on the scale of some interpretable transformation $Q(\xi)$ of ξ which we use to control the density mass of the prior distribution. This is typically done through quantiles, $\text{Prob}(Q(\xi) > U) = \alpha$, where U is a plausible bound specified by the user, and α is the probability of the event such as an upper tail or the median. In this way, the hyperparameters U and α can be chosen so the prior is weakly-informative, or one can use prior and expert knowledge to make the prior more informative.

The prior is transformed from the distance space to the space of the flexibility parameter ξ , and is thus invariant to reparameterization. This is a large advantage, as we can specify the prior without taking the specified parameterization into account, but rather select the prior on a interpretable scale (Simpson et al., 2017). The PC priors have been shown to perform well in various contexts and for various parameters, such as BYM (Besag, York, and Mollié) models (Riebler et al., 2016), correlation parameters (Guo et al., 2017), autoregressive processes (Sørbye and Rue, 2017, 2018), and Matérn Gaussian random fields (Fuglstad et al., 2019).

For a standard deviation parameter σ , we have a linear transformation of the distance to the standard deviation, $d = \sigma$, and the PC prior is an exponential

distribution on the standard deviation:

$$\pi(\sigma) = \lambda e^{-\lambda\sigma}, \sigma \geq 0,$$

with rate parameter $\lambda = -\log(\alpha)/U$ where $\text{Prob}(U > \sigma) = \alpha$ (Simpson et al., 2017).

1.5 Model-wise variance priors

Often, expert knowledge consists of some approximate numeric value for a model parameter, and maybe also a corresponding uncertainty in this numeric value, but the knowledge is not necessarily on the variances σ_k^2 directly. To utilize the knowledge about the heritability in animal models, or the total variance of the random effects in disease mapping, we need a prior distributions on *proportions* of variances, and not on the variance components directly.

1.5.1 Distributing the total variance along a prior tree

Instead of considering each variance parameter independently, we can take the model as a whole and create a joint prior for all variance parameters. In this case, we first consider the *total latent variation* in the data, i.e., the total variation attributed to the random effects in the linear predictor $\boldsymbol{\eta}$ conditioned on the model parameters $\boldsymbol{\theta}$. Then we use *variance proportions* to distribute this total latent variance to the individual random effect components. This idea is explored by Simpson et al. (2017, Section 7), and they conclude that this approach opens for exploiting the structure of the model, in a way component-wise variance priors do not. This can be extended to a division of the linear predictor, where we distribute the latent variance for some components independent of other components. This is the idea behind the hierarchical decomposition (HD) prior framework.

To make this variance decomposition intuitive and easily communicated, we can imagine the linear predictor in the shape of a tree. Each random effect in the model is represented by a leaf node, which we combine according to our prior knowledge about the model and model structure. We use this tree structure to construct a joint prior for the variance parameters. Consider a model with a linear predictor consisting of three components:

$$\eta_{i,j,k} = a_i + b_{i,j} + c_{i,j,k}, \quad i = 1, \dots, p, \quad j = 1, \dots, m, \quad k = 1, \dots, n, \quad (1.5)$$

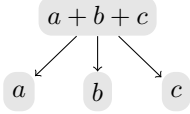


Figure 1.1: Tree structure describing the model in Equation (1.5) assigning equal amount of the total variance to each model effect.

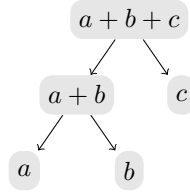


Figure 1.2: Tree structure describing the model in Equation (1.5) utilizing the nested structure of the model.

where $\mathbf{a} = (a_1, \dots, a_p) \sim \mathcal{N}_p(\mathbf{0}, \sigma_a^2 \mathbf{A})$, $\mathbf{b} = (b_1, \dots, b_m) \sim \mathcal{N}_m(\mathbf{0}, \sigma_b^2 \mathbf{B})$, and $\mathbf{c} = (c_1, \dots, c_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_c^2 \mathbf{C})$. σ_a^2 , σ_b^2 and σ_c^2 are marginal variance parameters and \mathbf{A} , \mathbf{B} and \mathbf{C} are covariance matrices. We denote the total variance, i.e., the sum of the variances of the random effects, $\sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2$. We want to visualize how the total variance is distributed among the three model effects. How this is done depends on the prior knowledge of the model and the hierarchical structure of the model. Figure 1.1 shows a tree structure where we do not use knowledge about the model structure, and give each component the same amount of variance in the prior. This is similar to using the same priors on each variance individually, however, we can easily control the total variance σ^2 . This is more complicated with component-wise priors, as the sum of the distributions may not belong to a known distribution family. The model in Equation (1.5) is nested, and we can utilize that in our tree structure, for example with a tree as in Figure 1.2.

Disease mapping models can contain several nested unstructured effects, and typically contain a spatial effect, for example at the coarsest level of nesting. Riebler et al. (2016) studied how the variance in a BYM (Besag, York and Mollié) model (Besag et al., 1991) can be divided between an unstructured noise effect and a structured spatial effect (Besag effect) in the same way we do with the HD prior. They argue that the two components should not be treated independently, but together, and a prior shrinking the structured effect can improve the inference. In this way the unstructured effect accounts for the variation in the BYM part of the model unless the data tells otherwise. See Section 1.5.3 for possible prior distributions for variance proportion parameters. As previously mentioned, in disease mapping it is natural to elicit prior knowledge about the total random effect variance, and how this is believed to be attributed to the

different random effects in the model, rather than asking about the individual variance parameters (Wakefield, 2006), making the model-wise approach highly relevant for such models.

The parameterization of the model is decided by the tree structure. For the tree structure in Figure 1.1, we get total variance $\sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2$, and two variance proportions $\omega_a = \sigma_a^2/\sigma^2$ and $\omega_b = \sigma_b^2/\sigma^2$ measuring the amount of total variance to effect **a** and **b**, respectively. The amount of total variance to effect **c** is $1 - \omega_a - \omega_b$. We get a different parameterization for the tree in Figure 1.2. See also Section 1.6. The scale of the prior knowledge varies between application, scientist, model and perhaps also the goal of the analysis. Prior elicitation will be easier if the parameterization of the model parameters coincides with the prior knowledge, and it is a huge advantage if the parameterization is flexible.

1.5.2 Priors for the total variance

The total variance is a variance parameter, and all distributions presented in Section 1.4.1 are applicable for such parameters.

Jeffreys' prior for the total variance is particularly handy when we have Gaussian data, as we do not need prior knowledge about the total variation in the data due to the scale-independence of the prior. In line with the principles of the penalized complexity (PC) prior, favouring the simplest model, it makes sense to shrink the total variance. If Jeffreys' prior is used, we can induce shrinkage towards the residual variance further down in the model, with a PC prior on a variance proportion distributing the total variance between residual and the other the random effects. Note that Jeffreys' prior is only meaningful for a prior where all model components are involved in the same prior tree. Jeffreys' prior is improper, but the joint prior on the variance proportions is proper, and in the case of a single tree and Gaussian data we are ensured a proper posterior. We can also use a PC prior on the total variance to include prior knowledge or induce shrinkage. The interpretation of the total variance in a model with Gaussian likelihood is straight-forward, as we can think on the scale of the data.

When the likelihood is not Gaussian, the interpretation of the total variance is more complicated than with Gaussian data. There is already an induced scale for the random effects through their effect on some measure related to the linear predictor, and a scale-invariant prior is not meaningful. One possible solution is to consider the linear predictor through the link function or on some

other relevant and interpretable scale. For example, odds-ratio for a binomial likelihood with logit link function, or the relative risk in models with a Poisson likelihood with log link function. We can use a PC prior on the total variance of the latent effects to induce shrinkage, meaning the variation in the data will be explained by any fixed effects in the model if there is no support in the data for there to be random effects present.

1.5.3 Priors for variance proportions

A variance proportion is restricted to be between 0 and 1. Several distributions with all the density mass in this interval exist. Just as for variance parameters, any distribution can be used on a scaled variance proportion parameter, for example on the logit-scale, but again this may obfuscate the interpretation of the prior.

The beta distribution is a popular prior choice in for example analysis of clinical trials (e.g., Brophy, 2020; Ye et al., 2020), and is given by

$$\text{Beta}(\omega; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \omega^{\alpha-1} (1 - \omega)^{\beta-1}, \quad 0 \leq \omega \leq 1.$$

$\Gamma(\cdot)$ is the gamma function and $\alpha, \beta > 0$. The uniform distribution is a special case of the beta distribution with $\alpha = \beta = 1$. The beta distribution can be generalized to multiple variables with the Dirichlet distribution, and is then the marginal distribution for each parameter. The Dirichlet distribution is given by:

$$\text{Dirichlet}(\boldsymbol{\omega}; \boldsymbol{\alpha}) = \frac{1}{\text{B}(\boldsymbol{\alpha})} \prod_{i=1}^p \omega_i^{\alpha_i-1}, \quad \boldsymbol{\omega} = (\omega_1, \dots, \omega_p) \text{ and } \sum_{i=1}^p \omega_i = 1, \quad \omega_i \geq 0,$$

where $\text{B}(\boldsymbol{\alpha}) = \prod_{i=1}^K \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^K \alpha_i)$ is the beta function. The Dirichlet and gamma distributions are related. If we have p gamma distributed parameter with equal rate parameter, the ratio of each of them divided by the sum will be Dirichlet distributed:

$$\xi_i \sim \text{Gamma}(\alpha_i, \theta), \quad i = 1, \dots, p \text{ and } \xi_i > 0, \text{ then}$$

$$\left(\frac{\xi_1}{\sum \xi_i}, \dots, \frac{\xi_p}{\sum \xi_i} \right) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_p).$$

The scale-independent Jeffreys' prior can be used for a variance proportion and is then given by

$$\pi(\omega) \propto \frac{1}{\omega(1-\omega)}, \quad 0 \leq \omega \leq 1.$$

As with Jeffreys' prior on variances, this prior does not use any expert knowledge and is improper, and should be used with care to ensure a proper posterior.

A variance proportion can be assigned a penalized complexity (PC) prior. The distribution function is dependent on the covariance matrix structure, and does in general not have an analytic expression. The idea is still the same: choose a value of the proportion that corresponds to the base model, and compute the distance from this base model to the flexible extension. Prior knowledge can be used to tune the prior through hyperparameters, and also here the PC prior provides shrinkage properties. An example for a random intercept model is included in the following section.

1.6 Example: Random intercept model

In general, we have to compute the Kullback-Leibler divergence (KLD) in Equation (1.4) numerically. However, in some special cases we can compute the penalized complexity (PC) prior for a variance proportion analytically. One of these exceptions is the random intercept model:

$$y_{i,j} = \alpha_j + \varepsilon_{i,j}, \quad i = 1 \dots, n_g, \quad j = 1, \dots, n_j, \quad n = \sum_{j=1}^{n_g} n_j. \quad (1.6)$$

$\boldsymbol{\alpha} \sim \mathcal{N}_{n_g}(\mathbf{0}, \sigma_\alpha^2 \mathbf{I}_{n_g})$ is a group effect and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$ is a residual effect. The variance parameters describing this model are σ_α^2 and σ_ε^2 , and we can imagine that a model with a $(\sigma_\alpha^2, \sigma_\varepsilon^2)$ parameterization is visualized as in Figure 1.3. If we instead follow the tree structure in Figure 1.4, we can introduce the parameterization $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$ and $\omega = \sigma_\alpha^2 / \sigma^2$, meaning σ^2 is the total variance in the observations, and ω is the amount of the total variance that is attributed to $\boldsymbol{\alpha}$. This corresponds to $\sigma_\alpha^2 = \omega \sigma^2$ and $\sigma_\varepsilon^2 = (1 - \omega) \sigma^2$. By setting priors for the variance proportion ω and total variance σ^2 instead of for the variances σ_α^2 and σ_ε^2 , we can use prior knowledge we have on the ratio of group effect to residual effect and on total variation in the data.

We choose a prior on ω such that we shrink the group effect unless the data indicate otherwise, i.e., we choose the base model to be $\omega = 0$, which gives



Figure 1.3: Tree structure visualizing the random intercept model (Equation (1.6)) when we use component-wise variance priors for σ_α^2 and σ_ε^2 .

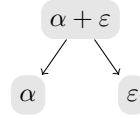


Figure 1.4: Tree structure visualizing the random intercept model (Equation (1.6)) when we use priors on the total variance priors σ^2 and the variance proportion ω .

$\sigma_\alpha^2 = 0$. We compute the distance $d(\omega) = \sqrt{2\text{KLD}}$ (Equation (1.4)) and the corresponding derivative $d'(\omega)$ for the variance proportion:

$$d(\omega) = \sqrt{-[(n - n_g) \log(1 - \omega) + \sum_{j=1}^{n_g} \log(1 + (n_j - 1)\omega)]},$$

$$d'(\omega) = \frac{1}{2d(\omega)} \left(-\frac{n - n_g}{1 - \omega} + \sum_{j=1}^{n_g} \frac{n_j - 1}{1 + (n_j - 1)\omega} \right).$$

If we assume that each group is of the same size, i.e., $n_j = n_p$ for all j , the expressions simplify to:

$$d(\omega) = \sqrt{-[(n - n_g) \log(1 - \omega) + n_g \log(1 + (n_p - 1)\omega)]},$$

$$d'(\omega) = \frac{1}{2d(\omega)} \left(-\frac{n - n_g}{1 - \omega} + n_g \frac{n_p - 1}{1 + (n_p - 1)\omega} \right).$$

We choose a median $\omega = m$, i.e., $\text{Prob}(\omega > m) = 1/2$, and get

$$\lambda = \log(2)/d(m).$$

Note that the latter holds in general for $d(\cdot) \sim \text{Exp}(\lambda)$. Hence, the prior for a variance proportion parameter depends on the choice of base model, the median, and the covariance matrix structure of the random effects involved.

Figure 1.5a shows the prior for ω where $n_g = n_p = 10$ and the median of ω is $m = 0.25$. The spike in $\omega = 1$ is caused by the infinite distance between the base model and $\omega = 1$, which is shown in Figure 1.5b, and does not induce overfitting.

With the (ω, σ^2) parameterization, it is straight-forward to see whether or not the model has learned anything from the data through the parameter ω .

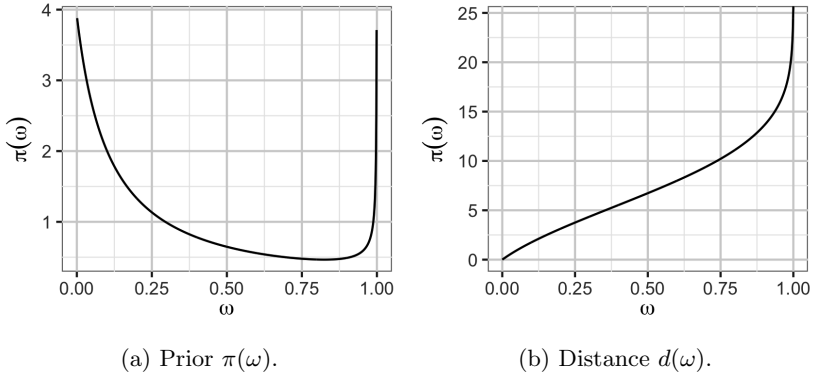


Figure 1.5: a) Prior distribution for the variance proportion ω , and b) the corresponding distance measure, for the random intercept model in Equation (1.6). We use $n_g = n_p = 10$ and $m = 0.25$.

Even though the prior and posterior of the individual variances differ, it does not mean that the model has learned about the ratio of group effect versus residual effect, it has only learned something about the total variance in the data. In this way we can uncover overfitting.

1.7 Bayesian inference

Bayesian inference can be performed in several ways. For models where we have a known posterior due to conjugacy, the inference is simple to carry out. In cases where we do not have conjugate distributions, but we have a low number of parameters, one can analytically obtain the posterior, or do it numerically. However, the models quickly become too big for straight-forward computation, and we need other tools to obtain the posterior distribution.

The bottle neck in the computation is the normalizing constant. In Bayes' theorem (Equation (1.1)), we know the prior $\pi(\boldsymbol{\theta})$, the latent process $\pi(\mathbf{x}|\boldsymbol{\theta})$ and the likelihood $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. However, the normalizing constant $\pi(\mathbf{y})$ is only analytically tractable in special cases. To obtain the posterior distribution, we need this constant, and we need to use numerical methods to compute it.

Markov Chain Monte Carlo (MCMC, see e.g. Gelfand and Smith, 1990) is

a common approach when the problems become too complex for direct computation. One simulates from the known prior distribution, latent model and likelihood to obtain the posterior distribution, without explicitly compute the normalizing constant. This is a very flexible method for inference, and will in the limit of an infinite number of samples be exact. However, it requires good implementation to be efficient, and for large models this can be difficult.

In most of the work in this thesis, Stan (Carpenter et al., 2017) is used to carry out the inference. This is a probabilistic programming language with a sampling algorithm that uses a variant of Hamilton Monte Carlo called the No-U-Turn Sampler (NUTS, Hoffman and Gelman, 2014). NUTS replaces random walks with an exploration strategy based on solutions of differential equations that is more efficient, and requires less tuning compared to other MCMC algorithms. Stan only needs the joint posterior distribution up to proportionality, meaning that we do not need to specify the normalizing constant or the full conditional distributions ourselves; Stan does this for us. This is a huge advantage and makes the model specification easy: the user only needs to write a code in a language similar to C++ that specifies the joint posterior. Through Stan, it is simple to compute desired posterior quantities. Stan is available in many programming languages, such as Python, MATLAB, and R (Stan Development Team, 2021b). We have used the R package `rstan` (Stan Development Team, 2020) to carry out the inference.

Another popular tool for inference is the Integrated Nested Laplace Approximations (INLA, Rue et al., 2009). This is a method where the marginal posterior distributions are approximated in a very efficient way for latent Gaussian models. We omit details on INLA, as we have mainly used Stan for inference and only used INLA for initial tests, and refer to Rue et al. (2009); Blangiardo and Cameletti (2015) and Rue et al. (2017) for a thorough description of INLA.

We listed some software and interfaces for doing Bayesian inference in Section 1.3. They all allow for user-specified prior distributions, and for example the Stan developers have given some recommendations on prior choices (Stan Development Team, 2021a). In most it is easy to implement the prior, but it is not always straight-forward to choose what prior to use, as it is difficult in itself to create a prior that utilize the available prior and expert knowledge.

To be able to visualize and inspect the chosen prior in a simple way can ease the process of choosing the prior we want to use, and it will increase the awareness of what prior distributions are chosen. We have implemented an R package that uses the hierarchical decomposition (HD) prior to ease the process

of using expert knowledge in the construction of in the prior. This software has a graphical user interface where the user easily can look at the chosen prior, and will be confronted with the choice, also when the default is used. The prior should be an active choice, and the R package helps communicating that.

In contrast to component-wise variance priors, which are easy to implement but difficult to choose, it easy to choose priors with the HD prior, but the implementation can be tedious. It may require reparameterization through Jacobians, and the distance measure used for the penalized complexity (PC) prior is complicated to compute. To make the prior framework available and useful in practice, we have automated the computation of the prior in the developed R package. Inference with INLA and Stan can be performed directly after having chosen the desired prior for a selected set of likelihoods and latent models. The software can also be used to select priors that to be used with another software, or to simply verify that chosen prior distributions are indeed the intended ones.

1.8 Discussion

In the first paper of this thesis (Fuglstad et al., 2020), we chose to compute the joint prior following the tree structure from the bottom and up, conditioning on the priors further down in the tree. This can also be done the other way around, by starting at the top and conditioning on priors higher up in the tree as we move downwards, and was merely a design choice. Note that this bottom and up approach does not decide what order one must specify the prior in; you are not restricted to choose priors and hyperparameters for the lowest level first.

To use the HD prior for model averaging could be interesting, by averaging over several prior trees. However, this would require specification of numerous parameters, and is as of today not something we have investigated.

Fixed effects are not included in the framework today. As pointed out by Goel and Degroot (1981) and Gelman et al. (2020), how vague the prior can be depends on the role of the parameter in the model. Parameters close to the data, such as the mean, do not need as strict priors as scale or shape parameters must have. The fixed effects have coefficients that control the mean of the (Gaussian) linear predictor, and we keep them out of the joint prior in the HD prior framework. An issue with fixed effects is that they are often correlated, and the variance explained by one effect is not well defined. It would however be interesting to see how each fixed effect contributes to the data variation, but

as already stated, this was outside the scope of this thesis.

In multiple linear regression models, we measure the amount of variance explained by the model with the coefficient of determination, also known as R^2 . Gelman and Hill (2007) discuss how this can be extended to hierarchical models with random effects, such as for the random intercept model (see Section 1.6), where R^2 is computed at each level in the hierarchical model. For a random intercept model, the variance proportion measuring the amount of variance accounted for by the group effect is equal to the intraclass correlation (ICC, McGraw and Wong, 1996), which again the generalized version of R^2 proposed by Gelman and Hill (2007) is linked to. Zhang et al. (2020) have investigated prior distributions for regression coefficients in high-dimensional linear regression. They propose to use a new class of shrinkage priors where one first specifies a prior on R^2 , which is a prior on a function of parameters, and then this prior is induced on the separate parameters in a natural way. These approaches could be starting points for including fixed effects in the HD prior framework.

Chapter 2

Scientific papers

The papers

Paper I Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*, 15(4):1109–1137.

Paper II Hem, I. G., Selle, M. L., Gorjanc, G., Fuglstad, G.-A., and Riebler, A. (2021). *Genetics*, iyab002. Advance publication.

Paper III Hem, I. G., Fuglstad, G.-A., and Riebler, A. (2021). `makemyprior`: Intuitive construction of joint priors for variance parameters in R. In preparation.

Documentation

Open-access research is important, and all code included in the papers is available for those who wants to reproduce the results. Most of the research has been through simulation studies, and scripts for reproducing the datasets are also available, see Supplementary materials for Papers I and II for code and data. Datasets used in real case studies are available online or upon request from the provider of the data. The R-package developed in Paper III is avail-

able online: https://github.com/ingebogh/makemyprior_0.1.0. See Paper III for details.

Paper I: “Intuitive joint priors for variance parameters”

There is little consensus on what prior distributions to use for variance parameters in Bayesian hierarchical models. Model-specific priors found in literature may be unsuitable for the application at hand, and independent general-purpose priors for variance parameters cannot exploit the model structure.

Paper I presents a new framework for joint prior distributions for variance parameters in latent Gaussian hierarchical models: the hierarchical decomposition (HD) prior framework. The idea is to follow a tree structure describing the model structure and how the total variance is attributed to the different random model components. For each split in the tree, the user can choose to be ignorant through a Dirichlet prior, or informative through a penalized complexity (PC) prior. The results show that the new HD prior approach perform at least as good as current state-of-the-art priors in terms of robust modelling, and are more transparent and intuitive.

The framework enables easy communication between statisticians, and between statisticians and applied scientists. As Bayesian modelling is becoming increasingly more popular and available, there is a corresponding increasing need for frameworks resulting in robust priors and at the same time are easy to understand. This to verify that the prior distribution actually reflects the beliefs of the scientist that holds knowledge about the model. The method yields robust priors in terms of them leading to stable inference.

Paper II: “Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge”

Nonadditive genetic variation is often hard to separate as it is confounded with other model effects. This may result in unstable inference and can in some cases lead to a divergent model where we cannot obtain results at all.

In Paper II we have applied the framework described in Paper I to a genomic model. The main contribution is a Bayesian approach that robustifies genomic modelling by utilizing prior expert knowledge. The hierarchical decomposition prior gives a parameterization on the scale of the expert knowledge, making it

easier for statisticians and geneticists to discuss the prior with minimal statistical jargon. The graphical visualization of the prior using a tree structure makes it intuitive and transparent.

The results show that the proposed prior approach with expert knowledge improves the robustness of genomic modelling over independent component-wise variance priors. It also gives a better variety selection in a simulated case study. In a real case study, the prior increases phenotype prediction accuracy for situations where the standard maximum likelihood approach is not able to find the optimal estimates for the variance parameters.

Paper III: “makemyprior: Intuitive construction of joint priors for variance parameters in R”

The hierarchical decomposition (HD) prior is intuitive and shown to be both useful and perform well, however, it is somewhat tedious to implement, especially in terms of computing the penalized complexity prior in large models. We have automated this process.

Paper III describes the R package `makemyprior`. This is a software where we have gathered the method developed in Paper I (Fuglstad et al., 2020) and applied in Paper II (Hem et al., 2021). The package eases prior specification in latent Gaussian models by utilizing the HD prior framework.

The software extends the idea of the HD prior to be applied to a subset of the model components and use component-wise variance priors and/or another HD priors on the remaining parameters. This makes the HD prior framework applicable in more general settings. The software use a graphical user interface to aid the prior selection process, where the user can choose a tree structure, and then be guided through the tree and choose prior distributions based on prior knowledge the user holds. The package allows the user to feed the chosen prior directly into the Bayesian inference interfaces `INLA` and `rstan` in R.

Bibliography

- Allenby, G. M., Arora, N., and Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, 32(2):152–162.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, 53:370–418.
- Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics*, 37(2):905–938.
- 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.
- Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, West Sussex, United Kingdom.
- Brophy, J. M. (2020). Bayesian interpretation of the excel trial and other randomized clinical trials of left main coronary artery revascularization. *JAMA Internal Medicine*, 180(7):986–992.
- Caldwell, A., Kollar, D., and Kröninger, K. (2009). BAT—The Bayesian analysis toolkit. *Computer Physics Communications*, 180(11):2197–2209.

- 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).
- Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.
- Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Ltd., Hoboken, New Jersey, USA.
- Dallow, N., Best, N., and Montague, T. H. (2018). Better decision making in drug development through adoption of formal prior elicitation. *Pharmaceutical statistics*, 17(4):301–316.
- Dougherty, E. R. and Dalton, L. A. (2013). Scientific knowledge is possible with small-sample classification. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013(1):1–12.
- Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.
- Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*, 15(4):1109–1137.
- Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452.
- Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL.
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multi-level/Hierarchical Models*, volume 1. Cambridge University Press, New York, New York.

- 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.
- 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.
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808 [stat.ME]*.
- Goel, P. K. and Degroot, M. H. (1981). Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association*, 76(373):140–147.
- Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. <https://mc-stan.org/rstanarm>. R package version 2.21.1.
- Grau, V., Mewes, A., Alcaniz, M., Kikinis, R., and Warfield, S. K. (2004). Improved watershed transform for medical image segmentation using prior information. *IEEE transactions on medical imaging*, 23(4):447–458.
- 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.
- Hem, I. G., Selle, M. L., Gorjanc, G., Fuglstad, G.-A., and Riebler, A. (2021). Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge. *Genetics*. iyab002.
- Hobert, J. P. and Casella, G. (1996). The effect of improper priors on gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436):1461–1473.
- Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.
- Holand, A. M., Steinsland, I., Martino, S., and Jensen, H. (2013). Animal models and integrated nested Laplace approximations. *G3: Genes, Genomes, Genetics*, 3(8):1241–1251.

- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.
- Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American statistical Association*, 91(435):1343–1370.
- Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- 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.
- Lorenc, A. C. (1986). Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194.
- 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.
- Mäki-Tanila, A. and Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355–367.
- Martinez-Beneito, M. A. (2013). A general modelling framework for multivariate disease mapping. *Biometrika*, 100(3):539–553.
- McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intra-class correlation coefficients. *Psychological Methods*, 1(1):30.
- Merriam-Webster (2021). Occam’s razor. <https://www.merriam-webster.com/dictionary/Occam%27s%20razor>. Accessed 2021-02-10.
- O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts’ Probabilities*. John Wiley & Sons.

- Plummer, M. (2017). JAGS version 4.3. 0 user manual [Computer software manual]. sourceforge.net/projects/mcmc-jags/files/Manuals/4.x.
- 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.
- 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 (Statistical Methodology)*, 71(2):319–392.
- 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*, 4:395–421.
- 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.
- Smid, S. C. and Winter, S. D. (2020). Dangers of the defaults: A tutorial on the impact of default priors when using Bayesian SEM with small samples. *Frontiers in Psychology*, 11:3536.
- Sørbye, S. H. and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38(6):923–935.
- Sørbye, S. H. and Rue, H. (2018). Fractional Gaussian noise: Prior specification and model comparison. *Environmetrics*, 29(5-6):e2457.
- Sorensen, D. and Gianola, D. (2007). *Likelihood, Bayesian, and MCMC methods in Quantitative Genetics*. Springer Science & Business Media.
- Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). BUGS 0.5* Examples Volume 2 (version ii). *MRC Biostatistics Unit*.
- Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, volume 13. John Wiley & Sons.
- Stan Development Team (2018). Stan modeling language users guide and reference manual, version 2.18.0. <http://mc-stan.org>.

- Stan Development Team (2020). RStan: the R interface to Stan. <http://mc-stan.org/>. R package version 2.21.2.
- Stan Development Team (2021a). Prior choice recommendations. <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>. Accessed 2021-02-25.
- Stan Development Team (2021b). Stan. <https://mc-stan.org/>. Accessed: 2021-02-12.
- StataCorp (2017). *Stata Bayesian analysis, reference manual*. StataCorp LLC, College Station, TX, 15 edition.
- van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26.
- Wakefield, J. (2006). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.
- Ye, J., Reaman, G., De Claro, R. A., and Sridhara, R. (2020). A Bayesian approach in design and analysis of pediatric cancer clinical trials. *Pharmaceutical Statistics*, 19(6):814–826.
- Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). Bayesian regression using a prior on the model fit: The R2-D2 shrinkage prior. *Journal of the American Statistical Association*, 0(0):1–13.
- Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., and de Schoot, R. V. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development*, 14(4):305–320.

Paper I

Intuitive joint priors for variance parameters

Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A.

Bayesian Analysis, 15(4):1109–1137, 2020.

Intuitive joint priors for variance parameters

Geir-Arne Fuglstad¹, Ingeborg Gullikstad Hem¹, Alexander Knight¹,
Håvard Rue², and Andrea Riebler¹

¹Department of Mathematical Sciences, NTNU, Norway

²CEMSE Division, King Abdullah University of Science and Technology,
Saudi Arabia

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 hierarchical 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 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 consensus 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 $\text{InvGamma}(1, 5 \cdot 10^{-5})$ (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 $\text{InvGamma}(0.001, 0.001)$ distributions in their examples (Spiegelhalter et al., 1996; Plummer, 2017), and the Stata manual employs $\text{InvGamma}(0.01, 0.01)$ priors (StataCorp, 2017). Conjugacy provides $\text{InvGamma}(\epsilon, \epsilon)$ 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 $t = \sigma_A^2 + \sigma_B^2 + \sigma_C^2$, where σ_A^2 , σ_B^2 and σ_C^2 are the variances of A, B and C, 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 $(\sigma_A^2, \sigma_B^2, \sigma_C^2)$ 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

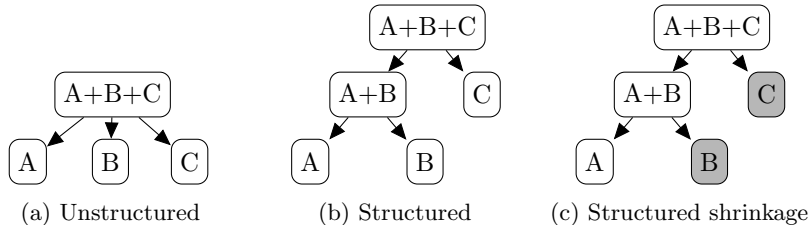


Figure 1: Hierarchical model decomposition. Gray boxes indicate preferred branches.

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 $\omega_1 = (\sigma_A^2 + \sigma_B^2)/t$, and then split $A+B$ into A and B through $\omega_2 = \sigma_A^2/(\sigma_A^2 + \sigma_B^2)$, where $0 \leq \omega_1, \omega_2 \leq 1$. The joint prior for $(\sigma_A^2, \sigma_B^2, \sigma_C^2)$ 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. 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 $\mathbf{y} = (y_1, \dots, y_n)$ be a vector of $n > 0$ observations. We model the expected values $E(y_i) = g^{-1}(\eta_i)$, $i = 1, \dots, n$, through a vector of linear predictors $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)$ and a link function $g : \mathbb{R} \rightarrow \mathbb{R}$. We consider models where the likelihood has parameters $\boldsymbol{\theta}_L$ and factors as $\pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta}_L) = \prod_{i=1}^n \pi(y_i|\eta_i, \boldsymbol{\theta}_L)$. 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

$$\eta_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^N u_{j,k_j[i]}, \quad i = 1, \dots, n, \quad (2.1)$$

where β_0 is the intercept, \mathbf{x}_i is the vector of covariates associated with observation i , $\boldsymbol{\beta}$ is a vector of coefficients, and $\mathbf{u}_j = (u_1, \dots, u_{m_j})$ is a random vector and $k_j[i]$ is the associated element of \mathbf{u}_j for observation i for $j = 1, \dots, N$. 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 \mathbf{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 $\boldsymbol{\theta}_M = (\sigma_1^2, \dots, \sigma_N^2)$. The BHM is completed by specifying the latent model through $\pi(\mathbf{u}_j | \sigma_j^2)$ for $j = 1, \dots, N$, and the prior $\pi(\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_L, \boldsymbol{\theta}_M)$. We follow common practice so that the prior satisfies $\pi(\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_L, \boldsymbol{\theta}_M) = \pi(\beta_0)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\theta}_L)\pi(\boldsymbol{\theta}_M)$. 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 $\pi(\boldsymbol{\theta}_M) = \prod_{j=1}^N \pi(\sigma_j^2)$.

2.2 Hierarchical variance decomposition

The additivity in Equation (2.1) causes the total latent variance $\text{Var}[\eta_i | \beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_M]$ of linear predictor i to decompose as the variance contributed by each random effect $\text{Var}[u_{k_j[i]} | \beta_0, \boldsymbol{\beta}, \sigma_j^2]$, $j = 1, \dots, N$, for $i = 1, \dots, n$. If random effect j is homogeneous, the variance parameter of random effect j will be a marginal variance in the sense that $\text{Var}[u_{k_j[i]} | \beta_0, \boldsymbol{\beta}, \sigma_j^2] = \sigma_j^2$ for $i = 1, \dots, n$. If all random effects are homogeneous, the total latent variance of the linear predictors is homogeneous, $t = \text{Var}[\eta_1 | \beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_M] = \dots = \text{Var}[\eta_n | \beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_M] = \sigma_1^2 + \dots + \sigma_N^2$. If random effect j is heterogeneous so that $\text{Var}[u_{k_j[i]} | \beta_0, \boldsymbol{\beta}, \sigma_j^2]$ 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 $t = \sigma_1^2 + \dots + \sigma_N^2$ 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 $T_0 = \{1, \dots, N\}$ that contains all the random effects, and in the first step we introduce $K_1 > 1$ child nodes T_1, \dots, T_{K_1} that partition T_0 into $T_0 = T_1 \cup \dots \cup T_{K_1}$. 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 $s = 1, \dots, S$. 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, \dots, C_{K_s} and we will define a vector of parameters $\boldsymbol{\omega}_s = (\omega_{s,1}, \dots, \omega_{s,K_s})$, $s = 1, \dots, S$. 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

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

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

For any split $s = 1, \dots, S$, we term a child node and its descendants as a branch of the split. The description of the model structure through a tree structure defines a re-parametrization of $(\sigma_1^2, \dots, \sigma_N^2)$ to $(t, \boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_S)$, 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 $(\sigma_A^2, \sigma_B^2, \sigma_C^2)$. Let the root node be $T_0 = \{A, B, C\}$.

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

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

2.3 Hierarchical 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, we use $\pi(\{\omega_s\}_{s=1}^S) = \prod_{s=1}^S \pi(\omega_s | \{\omega_j\}_{j \in D(s)})$, where $D(s)$ is the set of descendant splits for split $s = 1, \dots, S$.*

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 $\sigma_1^2, \dots, \sigma_N^2$. 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

$$\pi(\sigma_1^2, \dots, \sigma_N^2) = \pi(t | \{\omega_s\}_{s=1}^S) \prod_{s=1}^S \pi(\omega_s | \{\omega_j\}_{j \in D(s)}),$$

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

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 Equation (2.1) in vector form, $\boldsymbol{\eta} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \sum_{j=1}^N \mathbf{A}_j \mathbf{u}_j$, where $\mathbf{1} = (1, \dots, 1)$ 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 $j = 1, \dots, N$. The latent Gaussian structure is achieved by $\beta_0 \sim \mathcal{N}(0, \sigma_1^2)$, $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \sigma_F^2 \mathbf{I}_p)$, and $\mathbf{u}_j | \sigma_j^2 \sim \mathcal{N}_{m_j}(\mathbf{0}, \sigma_j^2 \Sigma_j)$ for $j = 1, \dots, N$. It is common to give σ_1^2 and σ_F^2 suitably vague values, and we will assume that σ_1^2 and σ_F^2 are fixed and focus on the variance parameters $\sigma_1^2, \dots, \sigma_N^2$.

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 \mathbf{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

3.2.1 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., 2019).

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

$$\text{KLD}(\pi(\mathbf{u}|\xi) \parallel \pi(\mathbf{u}|\xi = 0)) = \int \pi(\mathbf{u}|\xi) \log \left(\frac{\pi(\mathbf{u}|\xi)}{\pi(\mathbf{u}|\xi = 0)} \right) d\mathbf{u}, \quad (3.1)$$

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 : $d(f_1 \parallel f_2) = \sqrt{2\text{KLD}(f_1 \parallel f_2)}$. 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.

3.2.2 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 $P(\sigma > U) = 0.05$ 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 $\sigma \sim \text{PC}_{\text{SD}}(U, \alpha)$; see Simpson et al. (2017) for details and derivation.

3.2.3 Shrinking a weight parameter

Consider the situation that the linear predictor only contains two random effects A and B with variances σ_A^2 and σ_B^2 , respectively. The proportion of $t = \sigma_A^2 + \sigma_B^2$

assigned to each random effect is described by $\boldsymbol{\omega} = (1 - \omega, \omega) = (\sigma_A^2, \sigma_B^2)/(\sigma_A^2 + \sigma_B^2)$. 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 Equation (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 \mathbf{u}_1 and \mathbf{u}_2 be random effects of an LGM that enter the linear predictor through $\mathbf{A}_1\mathbf{u}_1 \sim \mathcal{N}_n(\mathbf{0}, \sigma_1^2\tilde{\Sigma}_1)$ and $\mathbf{A}_2\mathbf{u}_2 \sim \mathcal{N}_n(\mathbf{0}, \sigma_2^2\tilde{\Sigma}_2)$. Assume that $\tilde{\Sigma}_1 + \tilde{\Sigma}_2$ is non-singular¹. Let $\omega = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$ and $\Sigma(\omega) = (1 - \omega)\tilde{\Sigma}_1 + \omega\tilde{\Sigma}_2$. 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

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

where $\lambda > 0$ is the hyperparameter. We suggest to set λ so that the median is $\omega_m = 0.25$.

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

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

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

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

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

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

¹If this were not the case, some elements of the sum of $\mathbf{A}_1\mathbf{u}_1$ and $\mathbf{A}_2\mathbf{u}_2$ 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 $\mathbf{A}_1\mathbf{u}_1$ and $\mathbf{A}_2\mathbf{u}_2$.

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).

3.3 Expressing *a priori* ignorance about a split

3.3.1 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 $(\sigma_A^2, \sigma_B^2, \sigma_C^2)$. This can be done, for example, through a Dirichlet prior.

3.3.2 Dirichlet prior

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

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

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 $a_1 = \dots = a_K = a > 0$, 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 vertices of Δ^K , and for $a > 1$ the mode is $\boldsymbol{\omega} = (1, \dots, 1)/K$.

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 $P(\text{logit}(1/4) < \text{logit}(\omega_1) - \text{logit}(\omega_0) < \text{logit}(3/4)) = 1/2$. By symmetry the same marginal properties are satisfied for ω_i , $i = 2, \dots, K$.

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 $\{\boldsymbol{\omega}_1^0, \dots, \boldsymbol{\omega}_S^0\}$, we replace $\pi(\boldsymbol{\omega}_s | \{\boldsymbol{\omega}_j\}_{j \in D(s)})$ with $\pi(\boldsymbol{\omega}_s | \{\boldsymbol{\omega}_j = \boldsymbol{\omega}_j^0\}_{j \in D(s)})$, $s = 1, \dots, S$.*

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 $\sigma_1^2, \dots, \sigma_N^2$ and that the hierarchical decomposition of the variance is described through a tree with S splits. Under base models $\{\boldsymbol{\omega}_1^0, \dots, \boldsymbol{\omega}_S^0\}$, the prior is*

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

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

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 $\pi(t|\{\omega_s\}_{s=1}^S)$ 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, \dots, y_n , described by the Gaussian linear model $y_i|\eta_i \sim \mathcal{N}(\eta_i, \sigma_R^2)$ with

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

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 ω_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 ω_1 , and handle $t|\omega_1$ as discussed in Section 4.2. \triangle

Example 3 (Shrinkage in multilevel models). The latent part of the multilevel model in Section 1 can be written in vector form as $\boldsymbol{\eta} = \mathbf{A}_A \mathbf{u}_A + \mathbf{A}_B \mathbf{u}_B + \mathbf{A}_C \mathbf{u}_C$, where \mathbf{A}_A , \mathbf{A}_B and \mathbf{A}_C are sparse matrices selecting the appropriate group, individual and measurement effects, respectively. Assume we use an LGM, then $\mathbf{u}_1 \sim \mathcal{N}_G(\mathbf{0}, \sigma_A^2 \mathbf{I}_G)$, $\mathbf{u}_2 \sim \mathcal{N}_{GP}(\mathbf{0}, \sigma_B^2 \mathbf{I}_{GP})$ and $\mathbf{u}_3 \sim \mathcal{N}_{GPK}(\mathbf{0}, \sigma_C^2 \mathbf{I}_{GPK})$, 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 $t = \sigma_A^2 + \sigma_B^2 + \sigma_C^2$ through two splits. For the split at the root node, we decompose t according to the proportions $\omega_1 = (\sigma_A^2 + \sigma_B^2, \sigma_C^2)/t$. Then in the second split we decompose $\sigma_A^2 + \sigma_B^2$ according to the proportions $\omega_2 = (\sigma_A^2, \sigma_B^2)/(\sigma_A^2 + \sigma_B^2)$.

We use an HD prior where we apply base models $\omega_1^0 = (0, 1)$, which prefers C over A+B, and $\omega_2^0 = (0, 1)$, which prefers B over A. Due to the desire for shrinkage we apply PC priors and use Theorem 1 with base model ω_2^0 to compute $\pi(\omega_2)$. We

define $\tilde{\mathbf{u}}_1 = \mathbf{A}_A \mathbf{u}_A + \mathbf{A}_B \mathbf{u}_B$ and $\tilde{\mathbf{u}}_2 = \mathbf{A}_C \mathbf{u}_C$. Then if we condition on $\boldsymbol{\omega}_2$, the top split in Figure 1c compares $\tilde{\mathbf{u}}_1 | \boldsymbol{\omega}_2 \sim \mathcal{N}_n(\mathbf{0}, (\sigma_A^2 + \sigma_B^2)(\omega_{2,1} \mathbf{A}_A \mathbf{A}_A^T + \omega_{2,2} \mathbf{A}_B \mathbf{A}_B^T))$ and $\tilde{\mathbf{u}}_2 \sim \mathcal{N}_n(\mathbf{0}, \sigma_3^2 \mathbf{A}_3 \mathbf{A}_3^T)$, and the conditional prior $\pi(\boldsymbol{\omega}_1 | \boldsymbol{\omega}_2 = \boldsymbol{\omega}_2^0)$ can be computed using Theorem 1 with base model $\boldsymbol{\omega}_1^0$ conditional on $\boldsymbol{\omega}_2 = \boldsymbol{\omega}_2^0$. The joint prior is then $\pi(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) = \pi(\boldsymbol{\omega}_1 | \boldsymbol{\omega}_2 = \boldsymbol{\omega}_2^0) \pi(\boldsymbol{\omega}_2)$, and an appropriate prior is chosen for $\pi(t | \boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ as described in Section 4.2. \triangle

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_R^2 is expected to be well-identified, we can define the prior on t relative to σ_R^2 and shrink t by preferring to describe the total variance $V = t + \sigma_R^2$ in the model by σ_R^2 . 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_R^2 . 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 $V = t + \sigma_R^2$. The new top node has one branch with residual variance and the other branch is the subtree describing the latent model. Let $\boldsymbol{\omega}_R = (1 - \sigma_R^2/V, \sigma_R^2/V)$ and assume shrinkage through a PC prior $\pi(\boldsymbol{\omega}_R | \{\boldsymbol{\omega}_s = \boldsymbol{\omega}_s^0\}_{s=1}^S)$ with base model $\boldsymbol{\omega}_R^0 = (0, 1)$.*

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

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

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

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

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 $\pi(t|\{\omega_s\}_{s=1}^S)$ with base model $t_0 = 0$. This results in*

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

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

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

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 $\text{logit}(p) = \mu + x$, 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%, $P(1/2 < \exp(x) < 2) = 0.90$. \triangle

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 $j = 1, \dots, n_i$, $i = 1, \dots, n_g$, 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

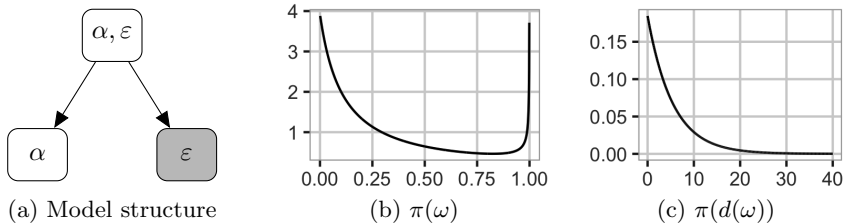


Figure 2: Model structure and prior for ω in the random intercept model with 10 individuals in each group and prior median $\omega_m = 0.25$. The prior is independent of the number of groups. a) Tree structure, b) prior for ω , and c) prior for distance $d(\omega)$.

are i.i.d. Gaussian with variance σ_R^2 . The total latent variance is $t = \sigma_\alpha^2$ and the total variance is $V = \sigma_R^2 + \sigma_\alpha^2$. 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 $\sigma_R^2 = (1-\omega)V$. We desire shrinkage towards the 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 $\omega_m = 0.25$ 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 $\sigma_\alpha^2/(\sigma_R^2 + \sigma_\alpha^2)$, 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 $\omega_m = 0.25$ (P-HD-25), $\omega_m = 0.5$ (P-HD-50) and $\omega_m = 0.75$ (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 $\text{InvGamma}(1, 5 \times 10^{-5})$ (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

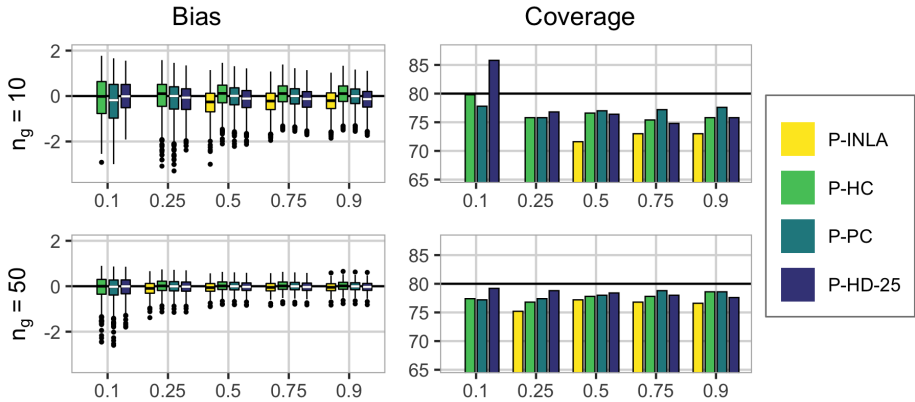


Figure 3: Results for $\text{logit}(\omega)$ for the random intercept simulation study. True value of ω shown on the x -axis, the number of groups is shown on left-hand side, and the group size is 10. Results for P-INLA are only shown when it leads to stable inference.

are simulated from the random intercept model for $n_g \in \{5, 10, 50\}$, 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 MCMC sampler that runs into numerical instabilities (Carpenter et al., 2017). In Figure S3.1 in the Supplementary Materials we 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 $\text{logit}(\omega)$ 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 $\text{logit}(\omega)$ 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 $\text{logit}(\omega)$ also holds for $\text{log}(V)$.

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 (see e.g., 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

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

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

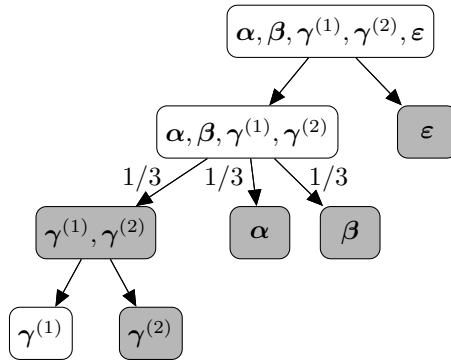
We believe that the effect of the treatment is ordered, and that the treatment

effect consists of a smooth signal of interest $\boldsymbol{\gamma}^{(1)} = (\gamma_1^{(1)}, \dots, \gamma_9^{(1)})$ and random noise $\boldsymbol{\gamma}^{(2)} = (\gamma_1^{(2)}, \dots, \gamma_9^{(2)})$ 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}_{RW2} , and the noise is $\boldsymbol{\gamma}^{(2)} \sim \mathcal{N}_9(\mathbf{0}, \sigma_t^2 \mathbf{I}_9)$. 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.

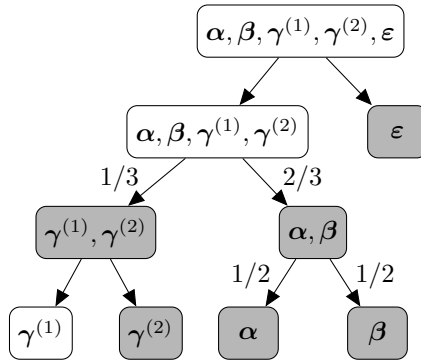
We set the true standard deviations equal, $\sigma_r = \sigma_c = \sigma_t = \sigma_R = 0.1$, and let the true effect of treatment be given by $x_i = C((i-5)^2 - 20/3)$, $i = 1, \dots, 9$. We entertain three scenarios: $C = 0$ for no effect of treatment (S1), $C = 0.05$ for medium effect of treatment (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 the residual variance σ_R^2 combined with $\text{InvGamma}(1, 5 \times 10^{-5})$ for σ_r^2 , σ_c^2 , σ_t^2 and σ_{RW2}^2 (P-INLA), or Half-Cauchy(25) (P-HC) or $\text{PCSD}(3, 0.05)$ (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.

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) $-\frac{1}{81} \sum_{i=1}^{81} \log \pi(y_i | \mathbf{y}_{-i})$. CRPS is a proper scoring rule and given by $\frac{1}{9} \sum_{i=1}^9 \int_{-\infty}^{\infty} (F_i(x) - \mathbb{I}(x \geq x_i))^2 dx$, 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.



(a) Original structure.



(b) Dual-split structure.

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.

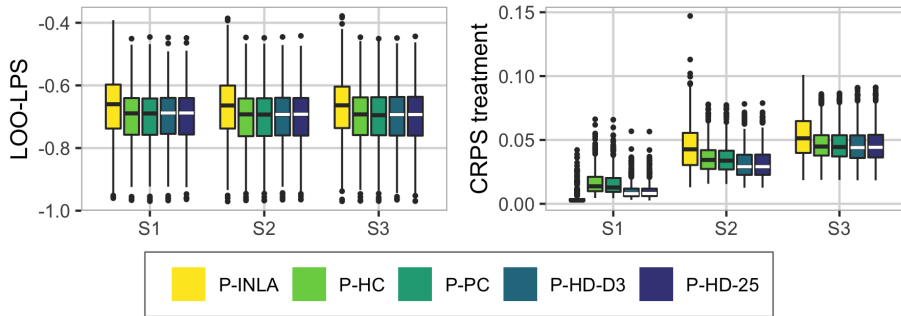


Figure 5: Results from the latin square experiment simulation study.

The 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., 2018; 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

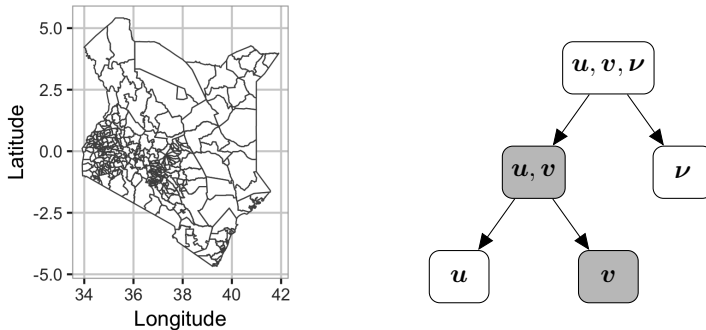
In this section we use the $n = 290$ constituencies shown in Figure 6a². We assume that $m_i = 6$ clusters are visited in constituency i , $i = 1, \dots, n$, and consider births $b_{i,j}$ and neonatal deaths $y_{i,j}$ in cluster j in constituency i . We assume that there are $b_{i,j} = 25$ live births in each cluster and the outcomes are simulated according to the model $y_{i,j}|p_{i,j} \sim \text{Binomial}(b_{i,j}, p_{i,j})$ for

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

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

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, $t = \sigma_B^2 + \sigma_{\text{IID}}^2 + \sigma_C^2$, so that $P(t > 3) = 0.05$. Further, we use $\text{InvGamma}(1, 5 \times 10^{-5})$ for σ_B^2 , σ_{IID}^2 and

²Preliminary 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 simulation study.

σ_C^2 (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_B^2 and σ_{IID}^2 has a PC prior of the type introduced in this paper with $P(\sigma_B^2/(\sigma_B^2 + \sigma_{\text{IID}}^2) < 0.5) = 2/3$ and σ_C^2 is given an independent PC prior $\sigma_C \sim \text{PC}_{\text{SD}}(3, 0.05)$.

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 $\sigma_C^2 = 0.1$ and create five scenarios by combining this with $\sigma_{\text{IID}}^2 = \sigma_B^2 = 0$ (S1), $\sigma_{\text{IID}}^2 = 0.4$ and $\sigma_B^2 = 0$ (S2), $\sigma_{\text{IID}}^2 = \sigma_B^2 = 0.2$ (S3), $\sigma_{\text{IID}}^2 = 0.04$ and $\sigma_B^2 = 0.36$ (S4), and $\sigma_{\text{IID}}^2 = 0$ and $\sigma_B^2 = 0.4$ (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 \mathbf{u} , the degree of structure in the spatial heterogeneity through $\omega^{(2)} = \sigma_B^2(\sigma_B^2 + \sigma_{\text{IID}}^2)^{-1}$, 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 \mathbf{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 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

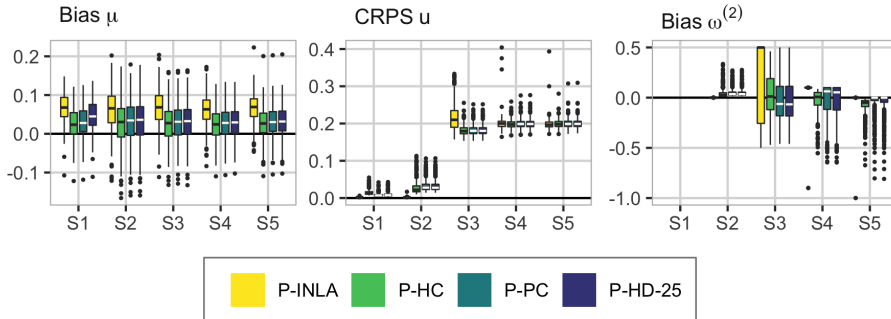


Figure 7: Main results from the Kenya neonatal mortality simulation study. Left to right: bias of the intercept μ , CRPS of \mathbf{u} and bias of $\omega^{(2)}$. Scenario shown on the x-axes.

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

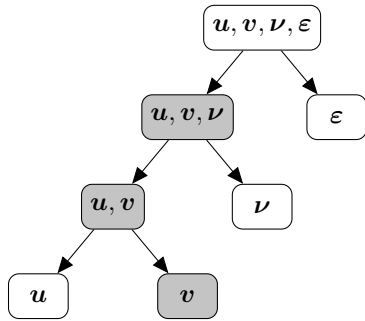
$$\eta_{i,j,k} = \text{logit}(p_{i,j,k}) = \mu + x_{i,j}\beta + u_i + v_i + \nu_{i,j} + \varepsilon_{i,j,k}, \quad (6.1)$$

where μ is an overall intercept, β is the effect of urban, \mathbf{u} is a Besag model with variance σ_{11}^2 , \mathbf{v} is a Gaussian i.i.d. effect of county with variance σ_{12}^2 , $\boldsymbol{\nu}$ is a Gaussian i.i.d. effect of cluster with variance σ_2^2 , and ε is a Gaussian i.i.d.

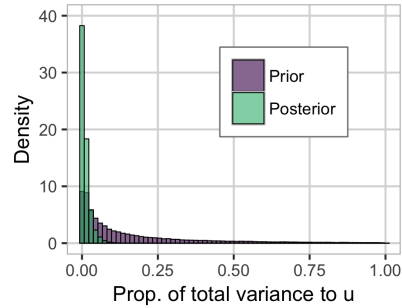
effect of household with variance σ_3^2 . In this model, \mathbf{u} and \mathbf{v} provide structured and unstructured, respectively, between-county variation, $\boldsymbol{\nu}$ provides between-cluster variation, and $\boldsymbol{\varepsilon}$ provides within-cluster variation. The Besag effect has 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 $P(\text{Total variance} > 11.296) = 0.05$. 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.

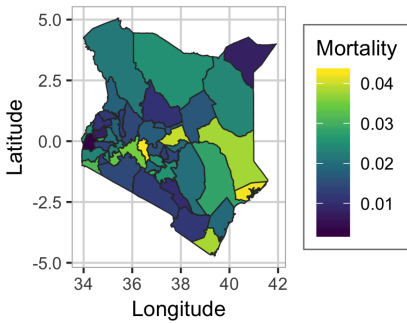
The model is parameterized by total standard deviation σ_T , and proportion of household variance to total variance of the random effects $\omega^{(1)}$, proportion of cluster variance to the sum of cluster and county variance $\omega^{(2)}$, and the proportion of structured spatial variance to county 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.



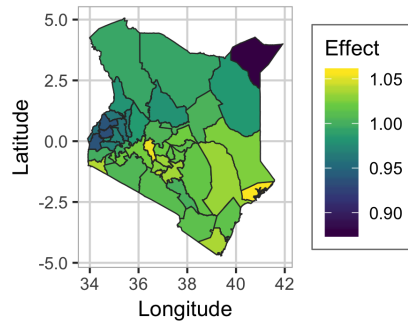
(a) Model structure.



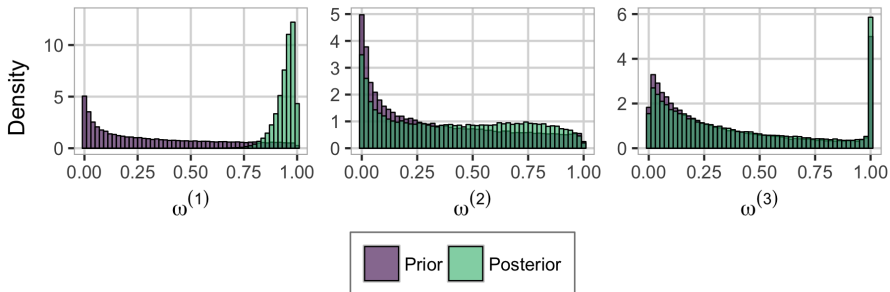
(b) Variance of u relative to total variance.



(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.

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 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 specifies 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., 2019). 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 transparency and clarity in *a priori* assumptions in the scientific community.

Supplement

Supplement to “Intuitive joint priors for variance parameters”. (DOI: [10.1214/19-BA1185SUPP](https://doi.org/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.

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.
- Balakrishnan, N. and Nevzorov, V. B. (2003). *A Primer on Statistical Distributions*. John Wiley & Sons, Hoboken, NJ.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.
- 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.
- 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.
- Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, West Sussex, United Kingdom.
- 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).
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284.
- 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.

- Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL.
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multi-level/Hierarchical Models*, volume 1. Cambridge University Press, New York, New York.
- 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.
- 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.
- General Assembly of the United Nations (2015). Resolution adopted by the General Assembly on 25 September 2015. A/RES/70/1.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- 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.
- 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.
- Hinkelmann, K. and Kempthorne, O. (1994). *Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design*. John Wiley & Sons.
- Holand, A. M., Steinsland, I., Martino, S., and Jensen, H. (2013). Animal models and integrated nested Laplace approximations. *G3: Genes, Genomes, Genetics*, 3(8):1241–1251.

- Jordan, A., Krüger, F., and Lerch, S. (2017). Evaluating probabilistic forecasts with the R package scoringRules. *arXiv preprint arXiv:1709.04743*.
- 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.
- 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*. CRC press, Boca Raton, FL. Github version www.r-inla.org/spde-book.
- 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.
- 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.
- Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63(19):1–25.
- 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.
- 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.
- Martinez-Beneito, M. A. (2013). A general modelling framework for multivariate disease mapping. *Biometrika*, 100(3):539–553.
- McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30.
- Plummer, M. (2017). JAGS version 4.3.0 user manual [Computer software manual]. sourceforge.net/projects/mcmc-jags/files/Manuals/4.x.
- Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538.

- 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.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC press, Boca Raton, Florida.
- Rue, H. and Held, L. (2010). Discrete spatial variation. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, *Handbook of Spatial Statistics*, Handbooks of Modern Statistical Methods, chapter 12, pages 171–200. CRC Press, Boca Raton, FL.
- 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.
- 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.
- 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.
- 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.
- Som, A., Hans, C. M., and MacEachern, S. N. (2014). Block hyper-g priors in Bayesian regression. *arXiv preprint arXiv:1406.6419*.
- 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)*. In press.
- Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51.
- Sørbye, S. H. and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38(6):923–935.
- Sørbye, S. H. and Rue, H. (2018). Fractional gaussian noise: Prior specification and model comparison. *Environmetrics*, 29(5-6):e2457.

- Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). BUGS 0.5* Examples Volume 2 (version ii). *MRC Biostatistics Unit*.
- Stan Development Team (2018a). RStan: the R interface to Stan. <http://mc-stan.org/>. R package version 2.18.1.
- Stan Development Team (2018b). Stan modeling language users guide and reference manual, version 2.18.0. <http://mc-stan.org>.
- StataCorp (2017). *Stata Bayesian analysis, Reference manual*. StataCorp LLC, College Station, TX, 15 edition.
- Wakefield, J. (2006). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.
- Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2018). Estimating under-five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*. In press.

Supplementary materials:

Intuitive joint priors for variance parameters

Geir-Arne Fuglstad¹, Ingeborg Gullikstad Hem¹, Alexander Knight¹,
Håvard Rue², and Andrea Riebler¹

¹Department of Mathematical Sciences, NTNU, Norway

²CEMSE Division, King Abdullah University of Science and Technology,
Saudi Arabia

S1 Proofs

S1.1 Theorem 3.1

Theorem S1.1 (Prior for the case $N = 2$). *Let \mathbf{u}_1 and \mathbf{u}_2 be random effects of an LGM that enter the linear predictor through $\mathbf{A}_1\mathbf{u}_1 \sim \mathcal{N}_n(\mathbf{0}, \sigma_1^2\tilde{\Sigma}_1)$ and $\mathbf{A}_2\mathbf{u}_2 \sim \mathcal{N}_n(\mathbf{0}, \sigma_2^2\tilde{\Sigma}_2)$. Assume that $\tilde{\Sigma}_1 + \tilde{\Sigma}_2$ is non-singular¹. 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

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

where $\lambda > 0$ is the hyperparameter. We suggest to set λ so that the median is $\omega_m = 0.25$.

¹If this were not the case, some elements of the sum of $\mathbf{A}_1\mathbf{u}_1$ and $\mathbf{A}_2\mathbf{u}_2$ 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 $\mathbf{A}_1\mathbf{u}_1$ and $\mathbf{A}_2\mathbf{u}_2$.

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

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

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

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

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

Proof:

First, note that since $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ are positive semi-definite and $\tilde{\Sigma}_1 + \tilde{\Sigma}_2$ is non-singular, $\Sigma(\omega) = (1 - \omega)\tilde{\Sigma}_1 + \omega\tilde{\Sigma}_2$ is positive definite for $0 < \omega < 1$. This follows from the fact that $\tilde{\Sigma}_1 + \tilde{\Sigma}_2$ is non-singular means that $\mathbf{v}^T(\tilde{\Sigma}_1 + \tilde{\Sigma}_2)\mathbf{v} \neq 0$ for $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{v} \neq \mathbf{0}$, where n is the dimension of $\tilde{\Sigma}_1$, which implies that either $\mathbf{v}^T\tilde{\Sigma}_1\mathbf{v} > 0$ or $\mathbf{v}^T\tilde{\Sigma}_2\mathbf{v} > 0$ for each $\mathbf{v} \neq \mathbf{0}$ so that $\mathbf{v}^T[(1 - \omega)\tilde{\Sigma}_1 + \omega\tilde{\Sigma}_2]\mathbf{v} > 0$ for $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{v} \neq \mathbf{0}$.

The proof of the theorem is split into three cases.

S1.1.1 Case 1: $\omega_0 = 0$ and $\tilde{\Sigma}_1$ is non-singular

The Kullback-Leibler divergence (KLD) from $\mathcal{N}_n(\mathbf{0}, \Sigma(\omega))$ to $\mathcal{N}_n(\mathbf{0}, \tilde{\Sigma}_1)$ is given by $\text{KLD}(\omega) = 0.5(\text{tr}(\tilde{\Sigma}_1^{-1}\Sigma(\omega)) - n - \log(|\tilde{\Sigma}_1^{-1}\Sigma(\omega)|))$, where tr denotes the trace of the matrix, and $\text{KLD}(\omega)$ is finite for $0 \leq \omega < 1$ since the KLD between two non-singular multivariate Gaussian distributions is finite. Thus a distance can be defined through

$$d(\omega) = \sqrt{\text{tr}(\tilde{\Sigma}_1^{-1}\Sigma(\omega)) - n - \log(|\tilde{\Sigma}_1^{-1}\Sigma(\omega)|)}, \quad 0 \leq \omega < 1, \quad (\text{S1.1})$$

and we follow Simpson et al. (2017) and use an exponential distribution on the distance so that $\pi(d) = \lambda \exp(-\lambda d)(1 - \exp(-\lambda d(1)))^{-1}$, $0 < d < d(1)$, where $\lambda > 0$, and the possibly truncated density is normalized by $(1 - \exp(-\lambda d(1)))$. A change of variables gives

$$\pi(\omega) = \frac{\lambda |d'(\omega)|}{1 - \exp(-\lambda d(1))} \exp(-\lambda d(\omega)), \quad 0 < \omega < 1. \quad (\text{S1.2})$$

□

S1.1.2 Case 2: $\omega_0 = 0$ and $\tilde{\Sigma}_1$ is singular

If $\tilde{\Sigma}_1$ is singular and $\Sigma(\omega)$, $0 < \omega < 1$, is non-singular, the distance $d(\omega)$ given in Equation (S1.1) is infinite for all $0 < \omega < 1$ and the direct approach for constructing the prior is not possible. We change the notation to $d(\omega; \omega_0)$ to make the dependence on the base model explicit. For any base model $\omega_0 > 0$, $d(\omega; \omega_0)$ is finite for $\omega_0 \leq \omega < 1$, and the prior can be constructed as for Case 1. The distance $d(\omega; \omega_0)$ is scaled by λ in Equation (S1.2) and we seek an expression $\lambda(\omega_0)$ so that $\lambda(\omega_0)d(\omega; \omega_0)$ remains finite for all $\omega_0 \leq \omega < 1$ when $\omega_0 \rightarrow 0^+$.

Since $\tilde{\Sigma}_1 + \tilde{\Sigma}_2$ is positive definite, there exist an $n \times n$ matrix \mathbf{P} so that

$$\mathbf{P}(\tilde{\Sigma}_1 + \tilde{\Sigma}_2)\mathbf{P}^T = \mathbf{I}.$$

This corresponds to a linear transformation of the Gaussian distributions that results in covariance matrices $\mathbf{S}_1 = \mathbf{P}\tilde{\Sigma}_1\mathbf{P}^T$ and $\mathbf{S}_2 = \mathbf{P}\tilde{\Sigma}_2\mathbf{P}^T$. The KLD is invariant to a linear transformation of the variables and the distance in Equation (S1.1) can be calculated by

$$d(\omega; \omega_0)^2 = \text{tr}(\mathbf{S}(\omega_0)^{-1}\mathbf{S}(\omega)) - n - \log(|\mathbf{S}(\omega_0)^{-1}\mathbf{S}(\omega)|),$$

where

$$\mathbf{S}(\omega) = (1 - \omega)\mathbf{S}_1 + \omega\mathbf{S}_2 = \omega(\mathbf{S}_1 + \mathbf{S}_2) + (1 - 2\omega)\mathbf{S}_2 = \omega\mathbf{I} + (1 - 2\omega)\mathbf{S}_1,$$

since $\mathbf{S}_1 + \mathbf{S}_2 = \mathbf{I}$.

\mathbf{S}_1 is symmetric and can be diagonalized so that $\mathbf{S}_1 = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^T$. This gives

$$\mathbf{S}(\omega) = \sum_{i=1}^n [(1 - 2\omega)\lambda_i + \omega] \mathbf{v}_i \mathbf{v}_i^T$$

so that

$$\mathbf{S}(\omega_0)^{-1}\mathbf{S}(\omega) = \sum_{i=1}^n \frac{[(1 - 2\omega)\lambda_i + \omega]}{[(1 - 2\omega_0)\lambda_i + \omega_0]} \mathbf{v}_i \mathbf{v}_i^T.$$

Thus the distance is given by

$$d(\omega; \omega_0)^2 = \sum_{i=1}^n \frac{[(1 - 2\omega)\lambda_i + \omega]}{[(1 - 2\omega_0)\lambda_i + \omega_0]} - n - \sum_{i=1}^n \log \left(\frac{[(1 - 2\omega)\lambda_i + \omega]}{[(1 - 2\omega_0)\lambda_i + \omega_0]} \right).$$

Let l be the rank deficiency of $\tilde{\Sigma}_1$ and assume that the eigenvalues of \mathbf{S}_1 are sorted from largest to smallest, then $\lambda_i > 0$ for $i = 1, \dots, n - l$ and $\lambda_i = 0$ for

$i = n - l + 1, \dots, n$, and the distance can be written as

$$d(\omega; \omega_0)^2 = l \left(\frac{\omega}{\omega_0} - \log \left(\frac{\omega}{\omega_0} \right) \right) + \sum_{i=1}^{n-l} \frac{[(1-2\omega)\lambda_i + \omega]}{[(1-2\omega_0)\lambda_i + \omega_0]} - n - \sum_{i=1}^{n-l} \log \left(\frac{[(1-2\omega)\lambda_i + \omega]}{[(1-2\omega_0)\lambda_i + \omega_0]} \right).$$

The first term blows up as ω_0 tends to zero, whereas the latter terms converges to a finite value. We introduce the scaled distance

$$\tilde{d}(\omega; \omega_0)^2 = \omega_0 d(\omega; \omega_0)^2 = l \left(\omega - \omega_0 \log \left(\frac{\omega}{\omega_0} \right) \right) + \omega_0 C(\omega_0),$$

where $C(\omega_0) = \mathcal{O}(1)$ as $\omega_0 \rightarrow 0^+$, and define $\tilde{d}(\omega; 0) = \lim_{\omega_0 \rightarrow 0^+} \sqrt{\omega_0} d(\omega; \omega_0) = \sqrt{l\omega}$.

Thus by letting $\lambda(\omega_0) = \sqrt{\omega_0/l}\tilde{\lambda}$, we find the density

$$\pi(\omega) = \frac{\tilde{\lambda}}{2\sqrt{\omega}(1 - \exp(-\tilde{\lambda}))} \exp(-\tilde{\lambda}\sqrt{\omega}), \quad 0 < \omega < 1, \quad (\text{S1.3})$$

as $\omega_0 \rightarrow 0^+$.

□

S1.1.3 Case 3: $0 < \omega_0 < 1$

This case proceeds like Case 1 for $0 \leq \omega < \omega_0$ and for $\omega_0 < \omega < 1$. On each side of ω_0 we get a similar expression as in Equation (S1.2). If we want to place the median at ω_0 we must place 1/2 probability on each side of ω_0 by introducing factors of 1/2 in the expressions. The density becomes

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

where $(1 - \exp(-\lambda d(0)))$ makes sure the density in $0 < \omega < \omega_0$ integrates to 1/2 and $(1 - \exp(-\lambda d(1)))$ makes sure the density in $\omega_0 < \omega < 1$ integrates to 1/2. □

S2 Multivariate PC priors for ignorance

The PC prior framework can be applied directly to dual splits since distance can be defined as a function of a single parameter. However, the PC prior framework does not translate to a general approach for distances that are functions of multiple parameters without further assumptions (Simpson et al., 2017, Section 6). Consider a split with $K > 2$ branches, and denote the proportion of variances assigned to each branch as $\omega = (\omega_1, \dots, \omega_K)$. Assume that the base model for the split is equal apportion of variance into the branches. Then the following procedure can be applied to replace the split with a sequence of dual splits.

Assumption S2.1 (Turn a multi-split into dual splits). *Consider a split in the tree structure that has $K > 2$ branches and assume that the variance in each branch is $\tilde{\sigma}_i^2$, for $i = 1, \dots, K$. We sequentially split out random effect 1, 2, and so on, through $K - 1$ dual splits. The proportion of variance assigned to random effect i of the total variance $\sum_{j=i}^K \tilde{\sigma}_j^2$ is $\omega^{(i)} = \tilde{\sigma}_i^2 / \sum_{j=i}^K \tilde{\sigma}_j^2$ for $i = 1, \dots, K - 1$. The base models are $\omega_0^{(i)} = 1/(K + 1 - i)$, and ensures that conditioning on the base models results in a proportion of $1/K$ of the total variance to each child node.*

The priors for each dual split can be precomputed before inference. The prior depends on the ordering of the $K - 1$ dual splits, but when the hyperparameters are set according to the suggested values for dual splits in the main article, we do not expect the ordering of the child nodes within each multisplit to greatly affect inference because the conditional priors are weakly informative in the sense that they put most mass around the base models, but also ensure that large deviations from the base model are plausible. The base models are chosen so that the variance is split equally between the child nodes.

S3 Gaussian responses: Random intercept model

In this section we include additional background, theory and results for the random intercept model simulation study from Section 5.1 in the main article.

S3.1 Additional background

The *random intercept model* is given by

$$y_{i,j} = \alpha_i + \varepsilon_{i,j}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n_g, \quad N = \sum_{i=1}^{n_g} n_i,$$

where $y_{i,j}$ is the j -th observation in group i , $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{n_g})^\top \sim \mathcal{N}_{n_g}(\mathbf{0}, \sigma_\alpha^2 \mathbf{I}_{n_g})$ is a vector with the random intercepts (group effect), and the residual noise (individual effect) is $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}, \varepsilon_{1,2}, \dots, \varepsilon_{n_g, n_{n_g}})^\top \sim \mathcal{N}_N(\mathbf{0}, \sigma_R^2 \mathbf{I}_N)$. We denote the N -dimensional vector of observations $\mathbf{y} = (y_{1,1}, y_{1,2}, \dots, y_{n_g, n_{n_g}})^\top$ and let \mathbf{A} be a block matrix of size $N \times n_g$ connecting the correct entries of $\boldsymbol{\alpha}$ to each observation in \mathbf{y} . Reparameterizing the model with total variance $V = \sigma_R^2 + \sigma_\alpha^2$ and $\omega = \sigma_\alpha^2/V$, the model can be written in vector form as

$$\mathbf{y} = \sqrt{V} (\sqrt{\omega} \mathbf{A} \boldsymbol{\alpha} + \sqrt{1 - \omega} \boldsymbol{\varepsilon}), \quad (\boldsymbol{\alpha}, \boldsymbol{\varepsilon}) \sim \mathcal{N}_{n_g + N}(\mathbf{0}, \mathbf{I}_{n_g + N}).$$

We use the R package **RStan** (Stan Development Team, 2018b) to perform the inference for all the three simulation studies in the paper. More specifically, we use the function `stan` from this package, where we use the following settings for the random intercept model simulation study: burn-in of length 25 000, total sample length of 125 000 (i.e., 100 000 samples after burn-in), one chain, we thin the chain to every fifth sample, initialize all parameters to zero, and we set the value `adapt_delta` to 0.95. `adapt_delta` is the average proposal acceptance probability Stan aims for during the adaption (burn-in) period, and a larger value will give a smaller step size (Stan Development Team, 2018a). For all other inputs we use the default values. We ran the simulation study on a computing cluster, where the full study runs in between a day and a week, depending on the available memory on the cluster.

RStan reports a *divergent transition* for each iteration of the MCMC sampler that runs into numerical instabilities (Carpenter et al., 2017). The divergent transitions are typically caused by an inappropriately large step size in the sampler or a poorly parameterized model, and may indicate that the results are biased since the sampler had trouble exploring the posterior (Stan Development Team, 2018a). It is difficult to completely avoid divergent transitions across all datasets, but to

avoid reporting biased results, we removed dataset and prior combinations that resulted in 0.1% or more divergent transitions during the inference for $n_g = 10$ or 50. For $n_g = 5$ we remove the dataset from the study if at least one prior results in too many divergent transitions. We report the proportion of datasets that resulted in at most 0.1% divergent transitions for each prior and scenario and use this as a measure of stability of the inference scheme for each prior.

S3.2 Connection to R^2

The coefficient of determination, commonly known as R^2 , is a measure on how much of the data variance is explained by a given linear regression model (Gelman and Hill, 2007). In frequentistic statistics, the R^2 is used to assess model fit by comparing the variance in the residuals to the variance in the data. Gelman and Hill (2007) generalise the R^2 to also make sense for multilevel models, such as the random intercept model. In this approach the R^2 is computed at each level of the model, which means we can assess the model fit at each level. In the case of the random intercept model, we have two levels in the model. The classical R^2 can be written as

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

where y_i , $i = 1, \dots, N$, are observations, $\bar{y} = N^{-1} \sum_{i=1}^N y_i$, and \hat{y}_i are the fitted values. Originally, the R^2 compares the model fit of any given linear regression model with covariates to a regression model with only an intercept. Gelman and Hill (2007) define the generalised R^2 at each level k in the model to be a comparison of the errors $\varepsilon_i^{(k)}$ at level k and the total linear predictor $\eta_i^{(k)}$ at the same level of the model. The total linear predictor $\eta_i^{(k)}$ is the covariates and predictors at level k in addition to the errors at the level, which means that $\eta_i^{(k)} \geq \varepsilon_i^{(k)}$ for all k . We write the generalised R^2 as

$$R_{\text{gen}}^{2,(k)} = 1 - \frac{\text{E} \left(\frac{1}{n_k} \sum_i \left(\varepsilon_i^{(k)} - \bar{\varepsilon}_i^{(k)} \right)^2 \right)}{\text{E} \left(\frac{1}{n_k} \sum_i \left(\eta_i^{(k)} - \bar{\eta}_i^{(k)} \right)^2 \right)}$$

where n_k is the number of observations/groups at level k . The random intercept model has two levels, so $k \in \{1, 2\}$. In the main article we have standardised the data and omitted the intercept from the random intercept model we use, and we have no covariates. This means that $\varepsilon_i^{(1)} = \varepsilon_i$, $\eta_i^{(1)} = y_i$, $\varepsilon_i^{(2)} = \alpha_i$ and $\eta_i^{(2)} = \alpha_i$,

and we have that

$$\begin{aligned} \mathbb{E} \left(\frac{1}{n_g} \sum_i (\alpha_i - \bar{\alpha}_i)^2 \right) &\stackrel{n_g \rightarrow \infty}{=} \mathbb{E}(\text{Var}(\boldsymbol{\alpha})) = \sigma_\alpha^2, \\ \mathbb{E} \left(\frac{1}{N} \sum_i (\varepsilon_i - \bar{\varepsilon}_i)^2 \right) &\stackrel{N \rightarrow \infty}{=} \mathbb{E}(\text{Var}(\boldsymbol{\varepsilon})) = \sigma_R^2, \\ \mathbb{E} \left(\frac{1}{N} \sum_i (y_i - \bar{y}_i)^2 \right) &\stackrel{N \rightarrow \infty}{=} \mathbb{E}(\text{Var}(\mathbf{y})) = \sigma_\alpha^2 + \sigma_R^2. \end{aligned}$$

The generalised R^2 at the group level ($k = 2$) for our model is zero (in the limit $n_g \rightarrow \infty$), which makes sense as there is nothing more in the linear predictor than the errors at the lowest level when we have no covariates in the model. For the data level, the generalised R^2 is given by $1 - \sigma_R^2/(\sigma_\alpha^2 + \sigma_R^2) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_R^2)$, which is the weight ω in the parametrization presented in this paper. Thus this weight is the asymptotic $R_{\text{gen}}^{2,(1)}$, which is also equal to the intra-class correlation.

S3.3 Results

We present all the results from the random intercept model simulation study. The priors used in the study are the HD prior with median $\omega_m = 0.25$ (P-HD-25), $\omega_m = 0.5$ (P-HD-50) and $\omega_m = 0.75$ (P-HD-75), the HD prior with a symmetric Dirichlet prior on the weight (P-HD-D), and the three commonly used priors P-INLA (Jeffreys' prior on residual variance and $\text{InvGamma}(1, 5 \times 10^{-5})$ on group variance), P-HC (Jeffreys' prior on residual variance and $\text{Half-Cauchy}(25)$ on group variance) and P-PC (Jeffreys' prior on residual variance and $\text{PC}_{\text{SD}}(3, 0.05)$ on group variance). The different scenarios we have used are the true weight $\omega \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, $n_g \in \{5, 10, 50\}$, $n_i = 10 \forall i$, and $n_i = 50 \forall i$, and 10 groups with varying group size where the group size is sampled from a $\text{Poisson}(10)$ -distribution, and samples equal to 0 or 1 is set to 10 so no group is of size smaller than 2. As performance measures we use the bias (estimated median minus true value) and 80% coverage (found by counting the number of times the true value lies in the 80% credible interval) of $\log(V)$ and $\text{logit}(\omega)$, and the number of datasets that leads to more than 0.1% divergent transitions during the inference as a measure of stability. All the box-plots show the median, the first and third quartile, 1.5 times the inter-quartile range (distance between first and third quartile), and outliers, if any.

From Figure S3.1 we see that P-INLA is less stable than the other priors, except for datasets with five groups where also P-HC leads to inference with too many divergent transitions. If a dataset leads to more than 0.1% divergent

transitions for a given prior, we remove the dataset from the study for this prior. For the scenarios with $n_g = 5$, P-INLA and P-HC are more affected by divergent transitions than the other priors. In this case we remove the dataset from the study for all priors. This means that the results for P-INLA is based on fewer simulations than the other priors for $n_g = 10$ or 50.

Figure S3.2 shows the posterior distribution of the logarithm of the group variance ($\log(\sigma_\alpha^2)$) when the priors of σ_R^2 and σ_α^2 are Jeffreys' and $\text{InvGamma}(1, 5 \times 10^{-5})$ (i.e. the INLA default prior), respectively. This is the true posterior, calculated using numerical integration, with a dataset where the maximum likelihood (ML) estimates of the group and residual variances are exactly equal to ω and $1 - \omega$, respectively. We vary the value of ω , and have 10 groups with 10 persons in each. When the true $\omega = 0.1$, and most of the variance in the model is residual variance, the posterior is highly influenced by the prior and we have close to no mass at the ML estimate (which is 0.1). When $\omega = 0.25$, the posterior is bimodal, and when $\omega = 0.5$ almost all the mass is at the ML estimate. This explains the bad results from P-INLA for datasets with true $\omega \leq 0.5$.

Figures S3.3-S3.7 show all the bias and coverage results from the random intercept model simulation study. Note that the coverage of ω is only shown for values larger than 65%. The order of the priors is the same in the legend and for each scenario in all plots, so P-INLA is the leftmost, so comes P-HC and so on. For a given number of groups and group size, the magnitude of the bias for $\log(V)$ increases and for $\text{logit}(\omega)$ decreases when the true value of ω increases. This is expected as a larger value of ω means that the group variance is larger relative to the residual variance and the dataset provides more information about the ω than would be the case when group variance is small relative to residual variance. On the other hand, a larger ω means the group variance dominates the total variance V more and there is less information about the group effect, which only has 5, 10 or 50 replicates, than the residual effect, which has 10 or 50 replicates for each group. This means less information about the V .

In the following we list the main results from each figure. It is clear from Figure S3.3 that the choice of ω_m does not have a large impact on the results. For an HD prior with a Dirichlet prior on the weight ω (P-HD-D), the results are similar for the scenario with equal group and residual variance (true $\omega = 0.5$), and worse for the other scenarios. This is true for all dataset sizes. Figure S3.4 shows that also for varying group sizes the HD prior with a PC prior on ω behaves as well as or better than the other priors in terms of bias and coverage, and again the value of ω_m does not influence the results noticeably. Figure S3.5 shows that larger groups improves the results in terms of low bias and accurate coverage, especially for P-INLA, but not as much as larger number of groups improves the results. In Figures S3.6 and S3.7 we include results for fewer groups, $n_g = 5$,

and 10 and 50 persons in each group, respectively. It is difficult to estimate the group variance with a low number of groups, and the results show that P-INLA is performing badly in terms of both bias and coverage for V and ω . For a given scenario with the HD prior, the bias and the coverage both increases for increasing values of ω_m . P-HC leads to the least stable inference for $n_g = 5$, and the other five priors give about equally stable inference. Note that for a given scenario we have removed the same datasets from the results for all priors, and the results may be slightly biased because of this.

S3.4 Simulation study for small group sizes

We explore the properties of the HD prior when applied to problems with small datasets with only few observations in each group. Here the amount of information about the parameters is low and the risk of overfitting is high. We define overfitting as overestimating the value of ω , and thus estimating spurious signals in the group effect; and define underfitting as underestimating the value of ω . Specifically, we use a small simulation study with two observations per group, and group size $n_g \in \{10, 50, 100\}$. We include an additional prior denoted P-HD-10 not included in the main article, which is the HD prior with PC prior on weight with median $\omega_m = 0.1$. P-HD-10 is added to explore the option of higher shrinkage in small data settings. The remaining HD priors are introduced in the main article.

From Figure S3.8 one can see that the inference for total variance V is stable in terms of bias and coverage. This indicates that the Jefferey's prior on V works well also in low information settings. From Figure S3.9, one can see that the inference for the weight ω depends on the chosen prior. Using the recommended P-HD-25, we are slightly overfitting for the scenario where the true weight is 0.1, and we are slightly underfitting in the other scenarios. Using stronger shrinkage through P-HD-10 avoids overfitting for true weight equal to 0.1, but results in a stronger bias for higher values of the true weight, and the resulting coverage varies from 100% to 0% in the scenarios. P-HD-50, P-HD-75 and P-HD-D result in overfitting also for true weight equal to 0.25 for $n_g = 10$. The results indicate that the recommended prior P-HD-25 is also appropriate for small group sizes. None of the priors displayed lead to inference with more than 0.1% divergent transitions.

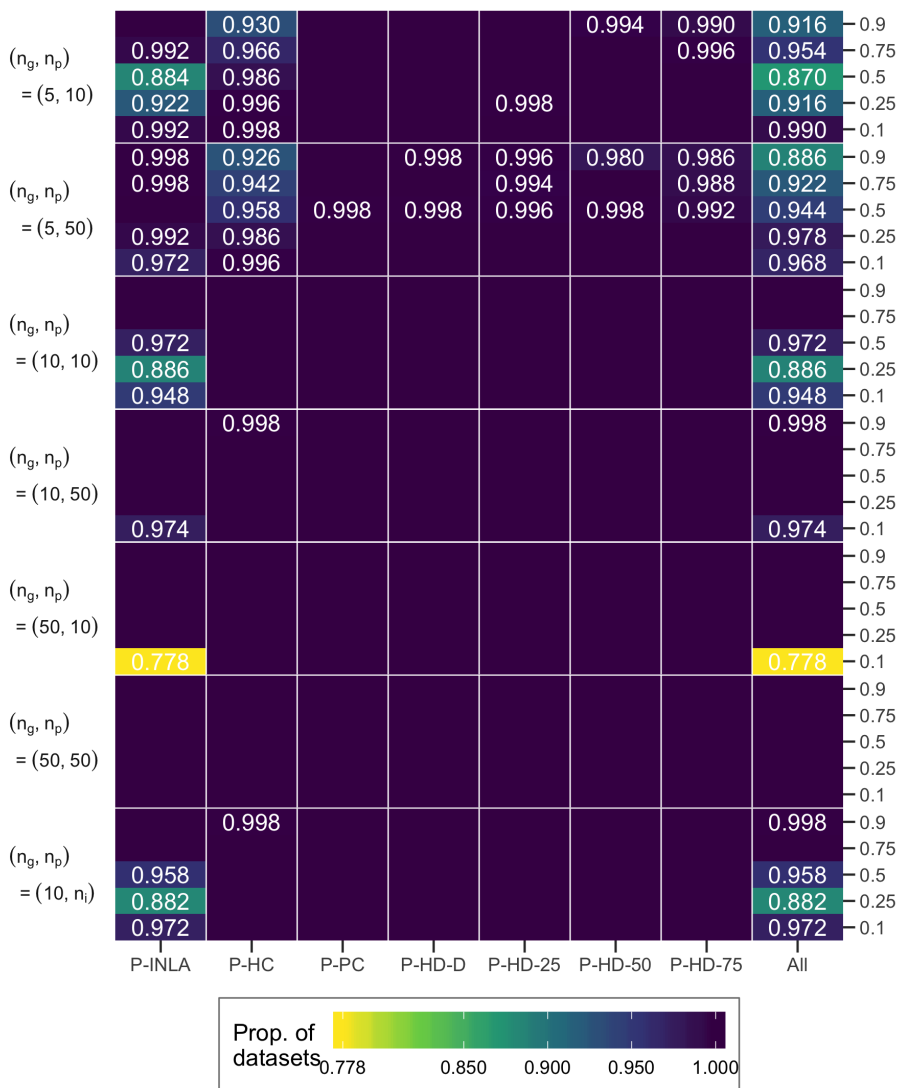


Figure S3.1: The proportion of datasets for each scenario and prior leading to at most 0.1% divergent transitions during the inference in the random intercept model simulation study. We say that the stability is 1.0 if all datasets for a given prior and scenario lead to no more than 0.1% divergent transitions. No number means that the stability is 1.0. The rightmost column, denoted “All”, shows how many datasets must be removed from the study so all priors lead to at most 0.1% divergent transitions for the remaining datasets.

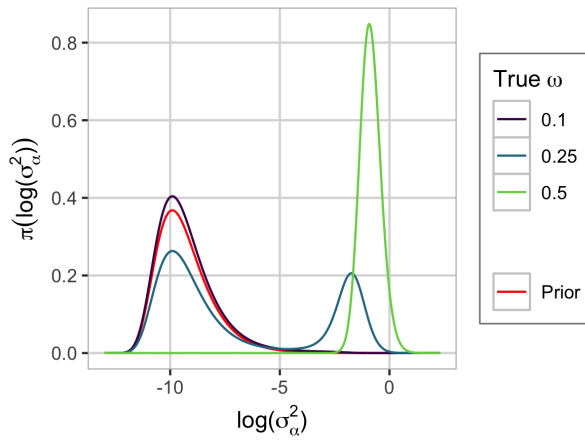


Figure S3.2: The posterior distribution of the logarithm of the group variance σ_α^2 when using Jeffreys' prior on the residual variance and $\text{InvGamma}(1, 5 \times 10^{-5})$ on the group variance (P-INLA). The prior on the group variance is included in the plot. We have $n_g = 10$ and $n_i = 10 \forall i$.

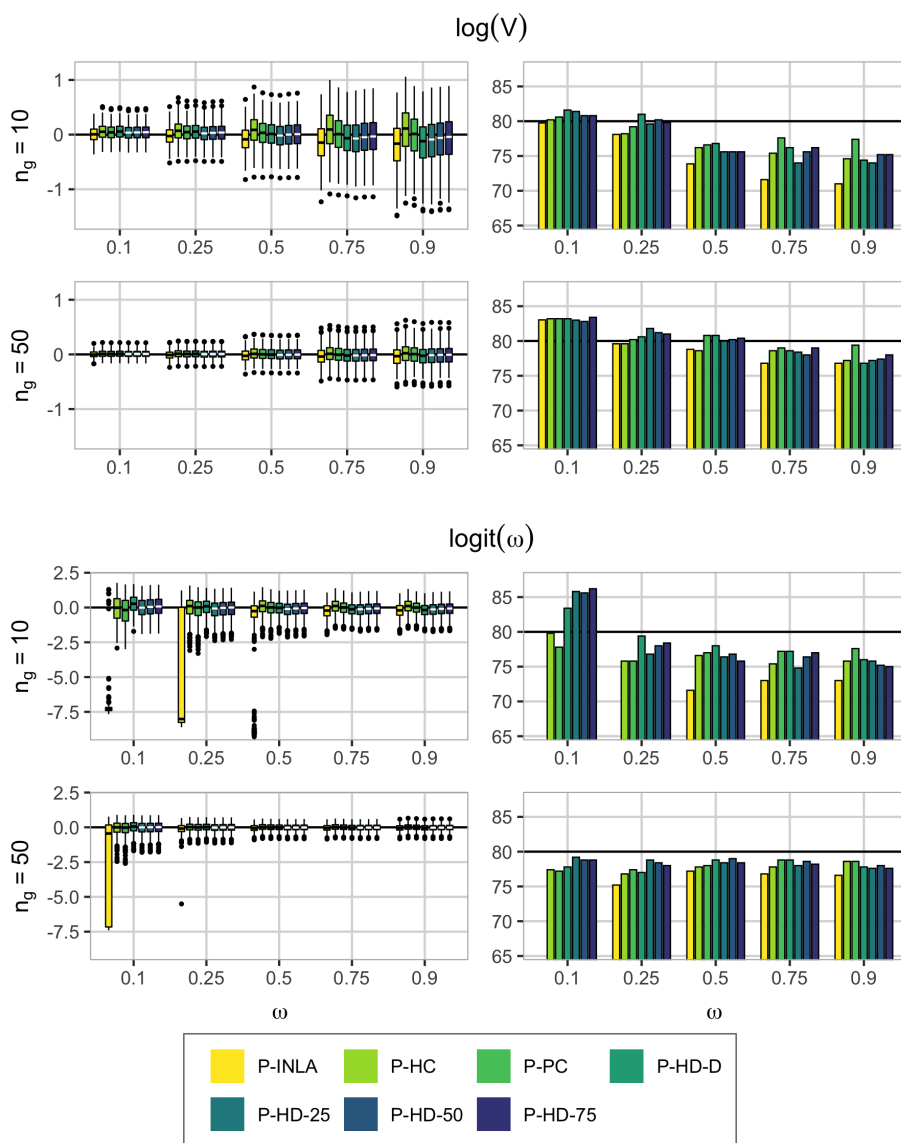


Figure S3.3: The true value of ω is on the x-axis in all graphs, the two upper rows contain the posterior diagnostics for the log total variance, and the two lower rows for logit weight. Bias in the left column, coverage in the right. The number of groups is indicated at the beginning of each row, either 10 or 50, and the group size $n_i = 10 \forall i$. The order of the priors is the same in the legend and for each scenario. The coverage for P-INLA is sometimes below the 65% and not shown in the figure.

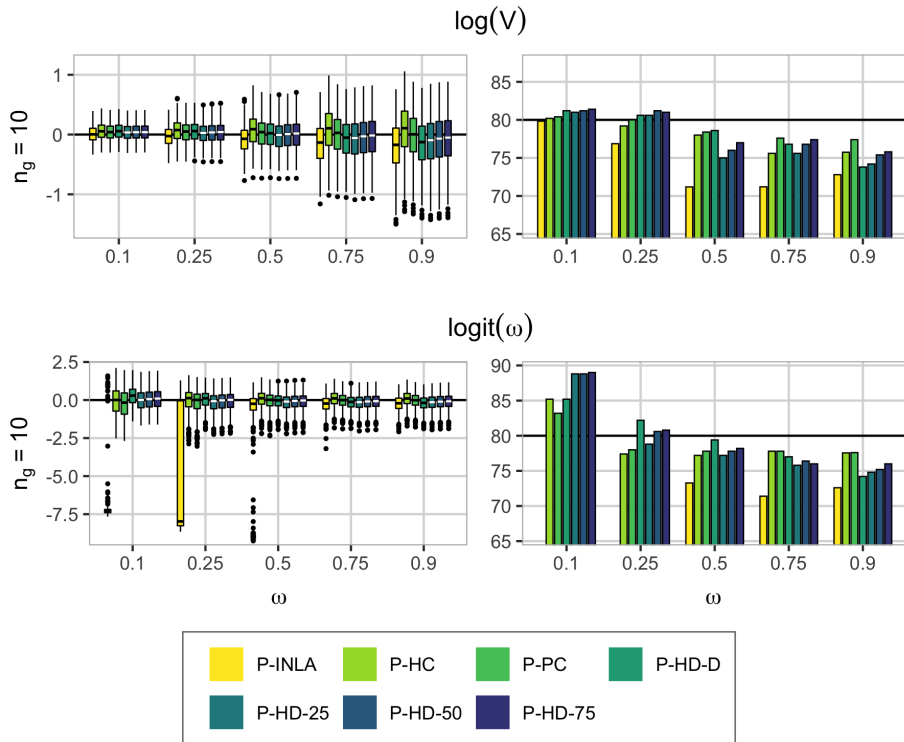


Figure S3.4: The true value of ω is on the x-axis in all graphs, the upper row contains the posterior diagnostics for the log total variance, and the lower row for logit weight. Bias in the left column, coverage in the right. The number of groups is 10 and the group size n_i varies. The order of the priors is the same in the legend and for each scenario. The coverage for P-INLA is sometimes below the 65% and not shown in the figure.

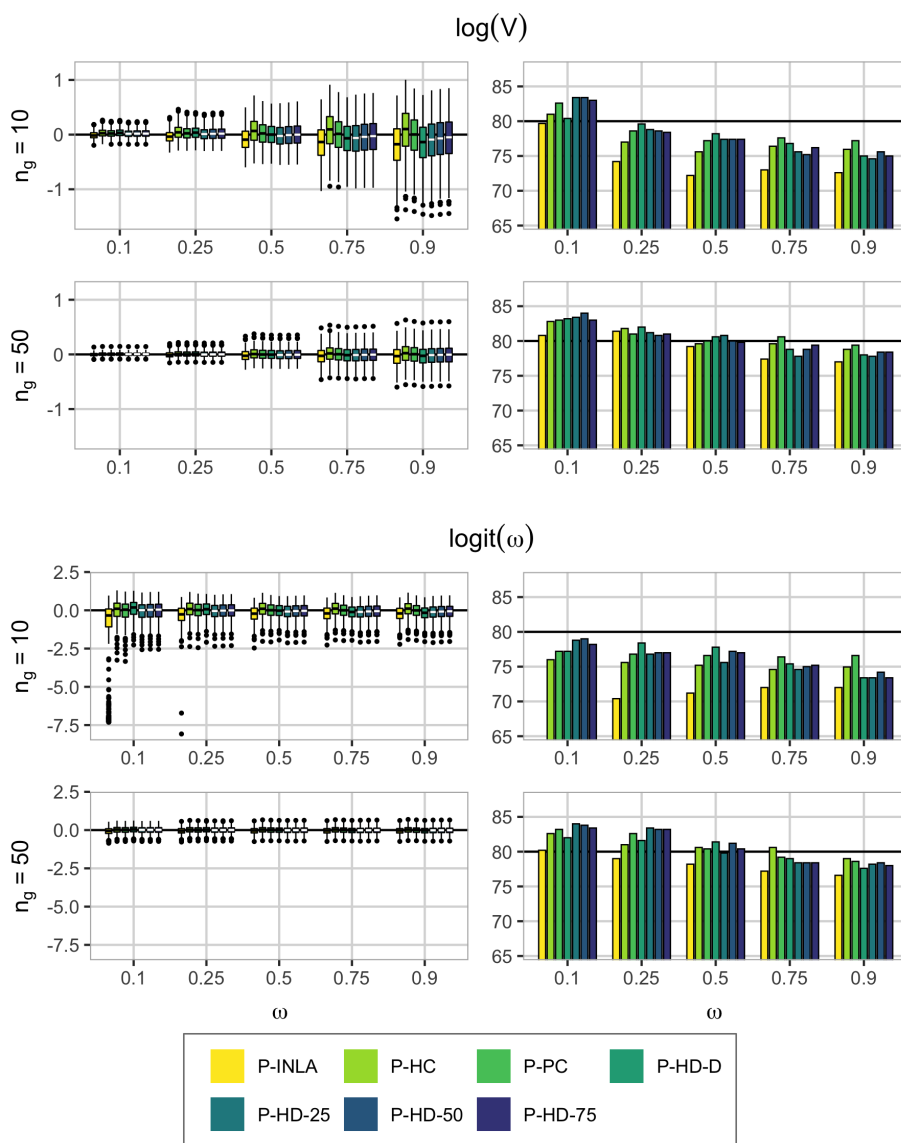


Figure S3.5: The true value of ω is on the x-axis in all graphs, the two upper rows contain the posterior diagnostics for the log total variance, and the two lower rows for logit weight. Bias in the left column, coverage in the right. The number of groups is indicated at the beginning of each row, either 10 or 50, and the group size $n_i = 50 \forall i$. The order of the priors is the same in the legend and for each scenario. The coverage for P-INLA is sometimes below the 65% and not shown in the figure.

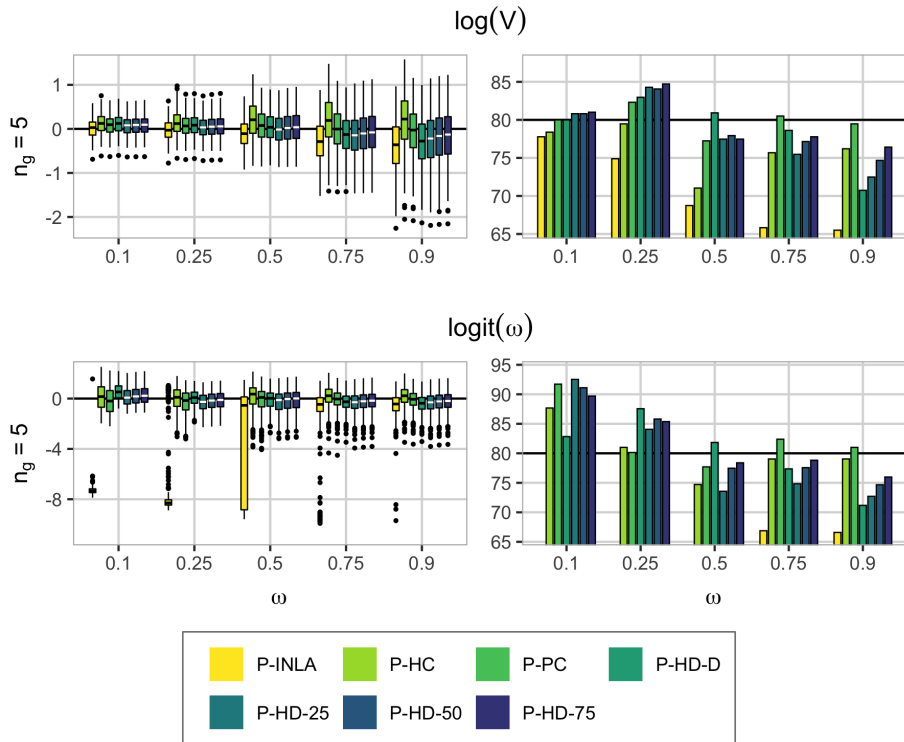


Figure S3.6: The true value of ω is on the x-axis in all graphs, the upper row contains the posterior diagnostics for the log total variance, and the lower row for logit weight. Bias in the left column, coverage in the right. The number of groups $n_g = 5$, and the group size $n_i = 10 \forall i$. The order of the priors is the same in the legend and for each scenario. The coverage for P-INLA is sometimes below the 65% and not shown in the figure.

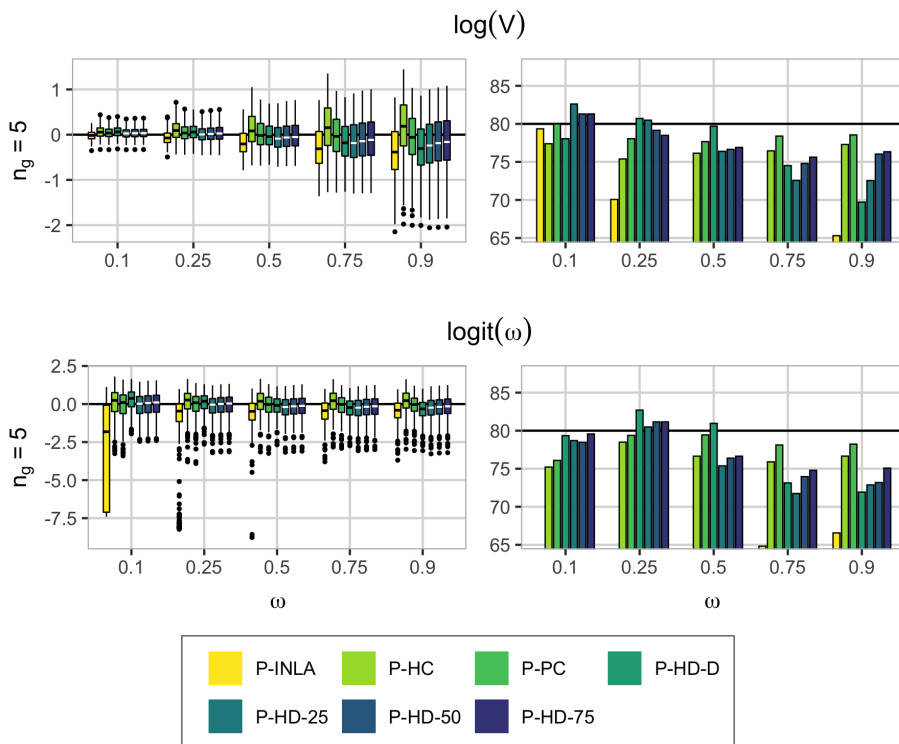


Figure S3.7: The true value of ω is on the x-axis in all graphs, the upper row contains the posterior diagnostics for the log total variance, and the lower row for logit weight. Bias in the left column, coverage in the right. The number of groups $n_g = 5$, and the group size $n_i = 50 \forall i$. The order of the priors is the same in the legend and for each scenario. The coverage for P-INLA is sometimes below the 65% and not shown in the figure.

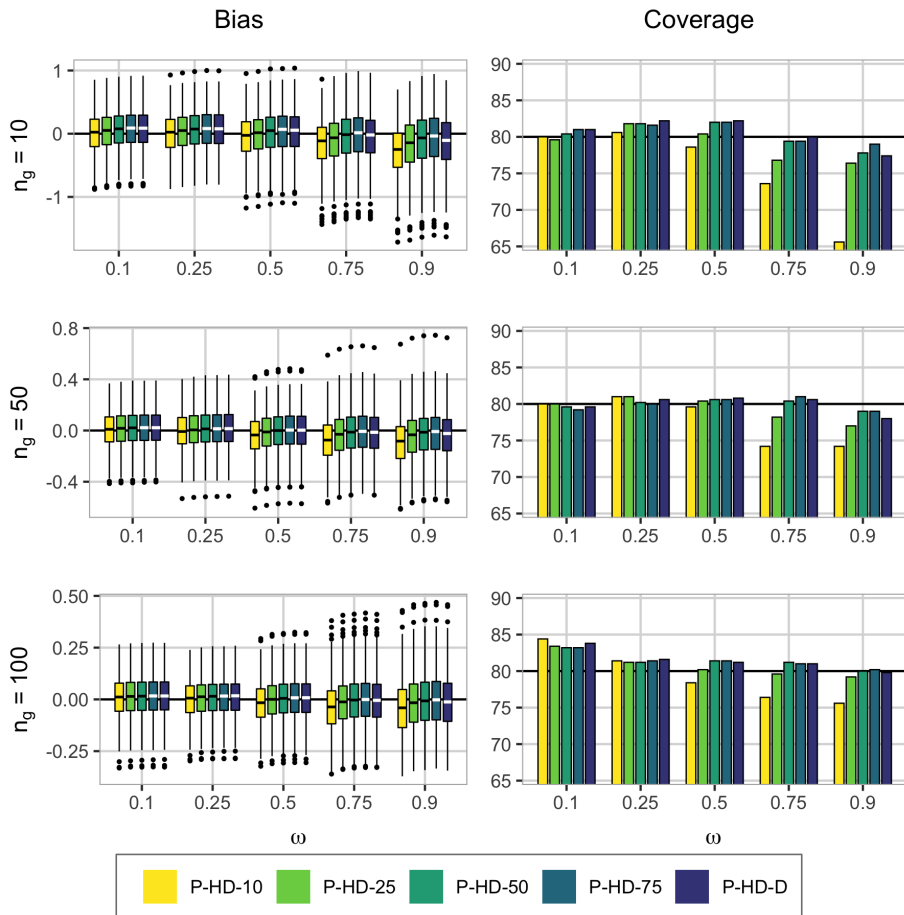


Figure S3.8: Results for $\log(V)$. The true value of ω is on the x-axis in all graphs, bias is shown in the left column, coverage in the right. The number of groups is indicated at the beginning of each row, and there are two persons in each group. The order of the priors is the same in the legend and for each scenario.

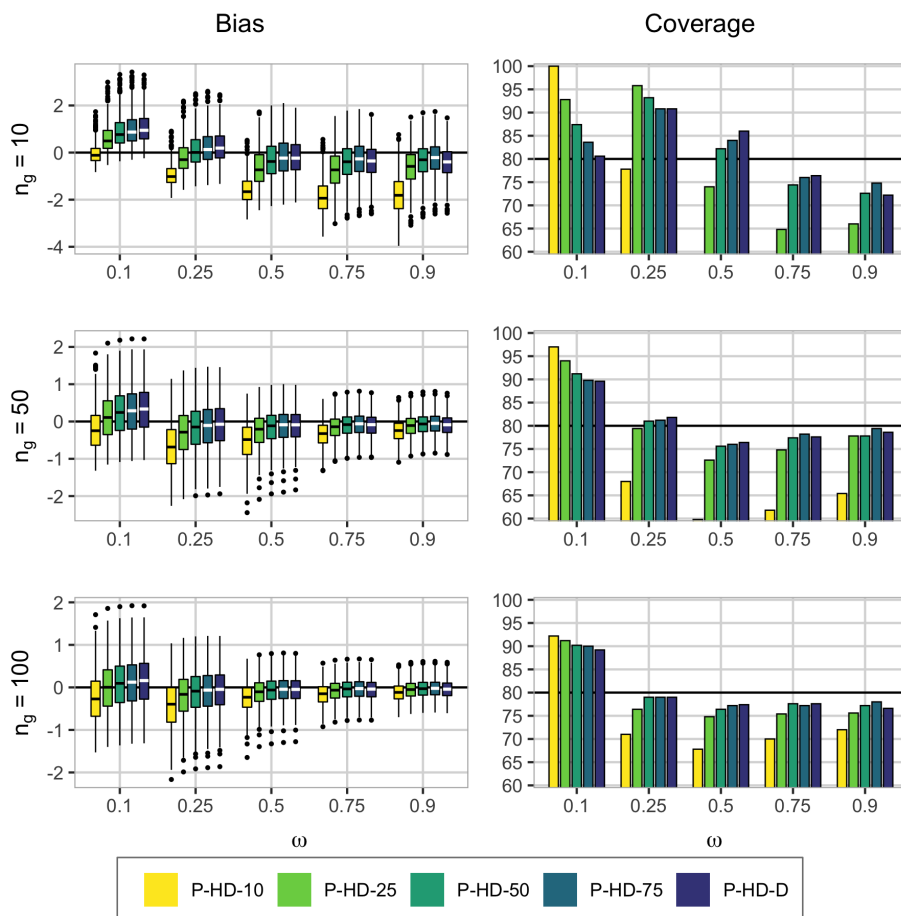


Figure S3.9: Results for $\text{logit}(\omega)$. The true value of ω is on the x-axis in all graphs, bias is shown in the left column, coverage in the right. The number of groups is indicated at the beginning of each row, and there are two persons in each group. The order of the priors is the same in the legend and for each scenario. The coverage for P-HD-10 is sometimes below the 65% and not shown in the figure.

S4 Gaussian responses: Latin square

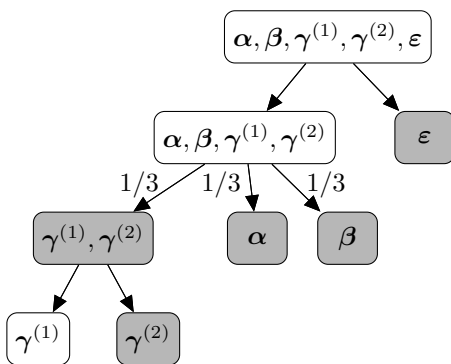
We include additional background and all results from the latin square simulation study from Section 5.2 in the main article.

S4.1 Additional background

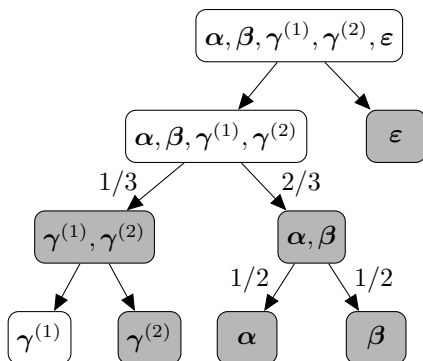
The reasoning behind the tree structure for the prior in the latin square simulation study displayed in Figure S4.1a is as follows: At the first level (top level) the prior shrinks the latent part of the model, at the second level the total latent variance is distributed with equal preference to the row effect, the column effect and the treatment effect, and at the third level the treatment effect is shrunk towards the unstructured effect. We select an HD prior using the model structure in Figure S4.1a. We also implement the triple split as explained in Section S2. The original order chosen in the main article is denoted Order1 (S4.1b), and the permuted orders Order2 (S4.1c) and Order3 (S4.1d). The total variance of the latent model is split into $\omega^{(1)}$, $\omega^{(2)}$ and $\omega^{(3)}$, which are the proportions of the latent variance going to the row effect, column effect and the treatment effect, respectively. Figure S4.3 shows the difference in marginal priors for $\omega^{(1)}$, $\omega^{(2)}$ and $\omega^{(3)}$ for Order1 and Order2, on weight scale and on logit weight scale. Figure S4.4 shows the difference in the same marginal priors for Order1 and a Dirichlet prior on the triple split, where the latter is the default choice in the HD prior framework.

The true treatment effect $\mathbf{x} = (x_1, \dots, x_9)$ we use in the latin square simulation study is given by $x_i = C((i - 5)^2 - 20/3)$, $i = 1, \dots, 9$ where $C = 0$ for scenario S1, $C = 0.05$ for scenario S2, and $C = 0.2$ for scenario S3. These corresponds to signal to noise ratios (SNRs) of 0%, 48% and 94% for S1, S2 and S3, respectively, as computed by $\text{SNR} = S_{xx}/(S_{xx} + \sigma_t^2)$, where $S_{xx} = \sum_{i=1}^9 (x_i - \bar{x})^2$. Figure S4.2 shows the true treatment effect for the three scenarios.

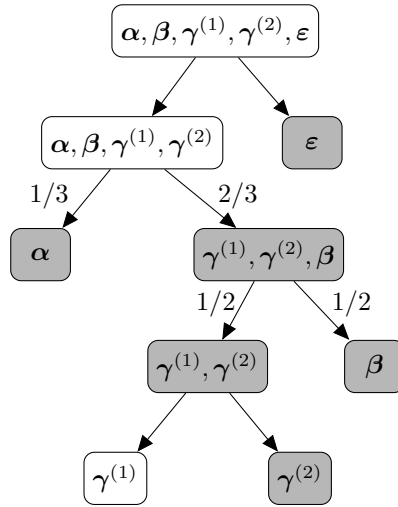
In the latin square experiment we use the following settings in the R-function `stan`: a burn-in of length 25 000, a total sample number (including burn-in) of 125 000, one chain which we thin to every fifth sample, we initialize all parameters to zero, and use `adapt_delta` equal to 0.95. We use default values for the rest of the settings. For the leave-one-out log predictive score (LOO-LPS), we use 1000 simulations for warm-up and 2000 samples in total, which yields a low estimated variance of the LOO-LPS. The simulation study ran on a computer cluster and takes no more than a couple of days, depending on the activity on the cluster.



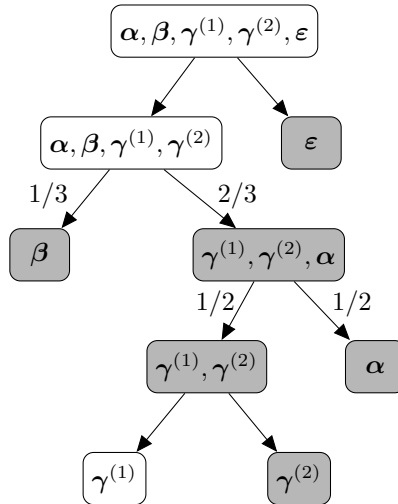
(a) Multi-split structure.



(b) Dual-split structure, original order (Order1)



(c) Dual-split structure, permuted order (Order2)



(d) Dual-split structure, permuted order (Order3)

Figure S4.1: Two of the possible orderings for turning the triple split into a dual split. a) The multi-split structure of the HD prior, b) the original order used in simulation study in paper (Order1), c) one permuted order (Order2), and d) the other permuted order (Order2)

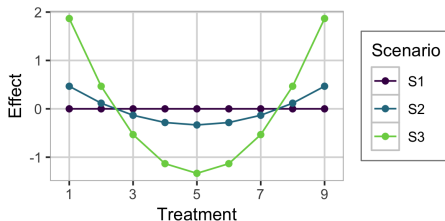


Figure S4.2: The true treatment effect for the simulated datasets in the latin square simulation study.

S4.2 Results

We have investigated the properties of the HD prior when the principles of the framework are tweaked. What we investigate is varying values of the median ω_m of the prior on the weight indicating the proportion of treatment variance going to the structured effect, varying distributions on the distance in the original PC prior framework, varying the value of λ for the multi-split, varying the type and ordering of the multi-split (see Figure S4.1), and we also study a joint prior where we use a Dirichlet prior on all effects except the residuals, and on all five effects. We compare the HD prior to the following default priors, where all have Jeffreys' prior on the residual variance and the following priors on the remaining variances or standard deviations: InvGamma($1, 5 \times 10^{-5}$) (P-INLA), Half-Cauchy(25) (P-HC), and PC_{SD}(3, 0.05) (P-PC).

For each scenario, we have removed the datasets that lead to more than 0.1% divergent transitions for at least one of the priors, so all the results for a given scenario are based on the same datasets for all priors. We use the proportion of datasets leading to at most 0.1% divergent transitions during the inference as a measure of stability, for each prior and scenario. Figure S4.5 displays these proportions for the latin square simulation study, and we see that it is not a big difference between P-INLA, P-HC, P-PC, and P-HD-25. However, when we lower the value of the shape parameter in the distribution we use on the distance (tweaking the third principle of the PC prior), the number of divergent transitions occurring during the inference increases, which indicates a more difficult posterior to draw samples from. When we change the values of ω_m , λ , or the way we implement the triple split (see Figure S4.1) the stability of the inference does not suffer.

Figures S4.6-S4.11 show all results from the latin square simulation study. The box-plots include the median, the first and third quartile, 1.5 times the inter-quartile range (distance between first and third quartile), and outliers, if

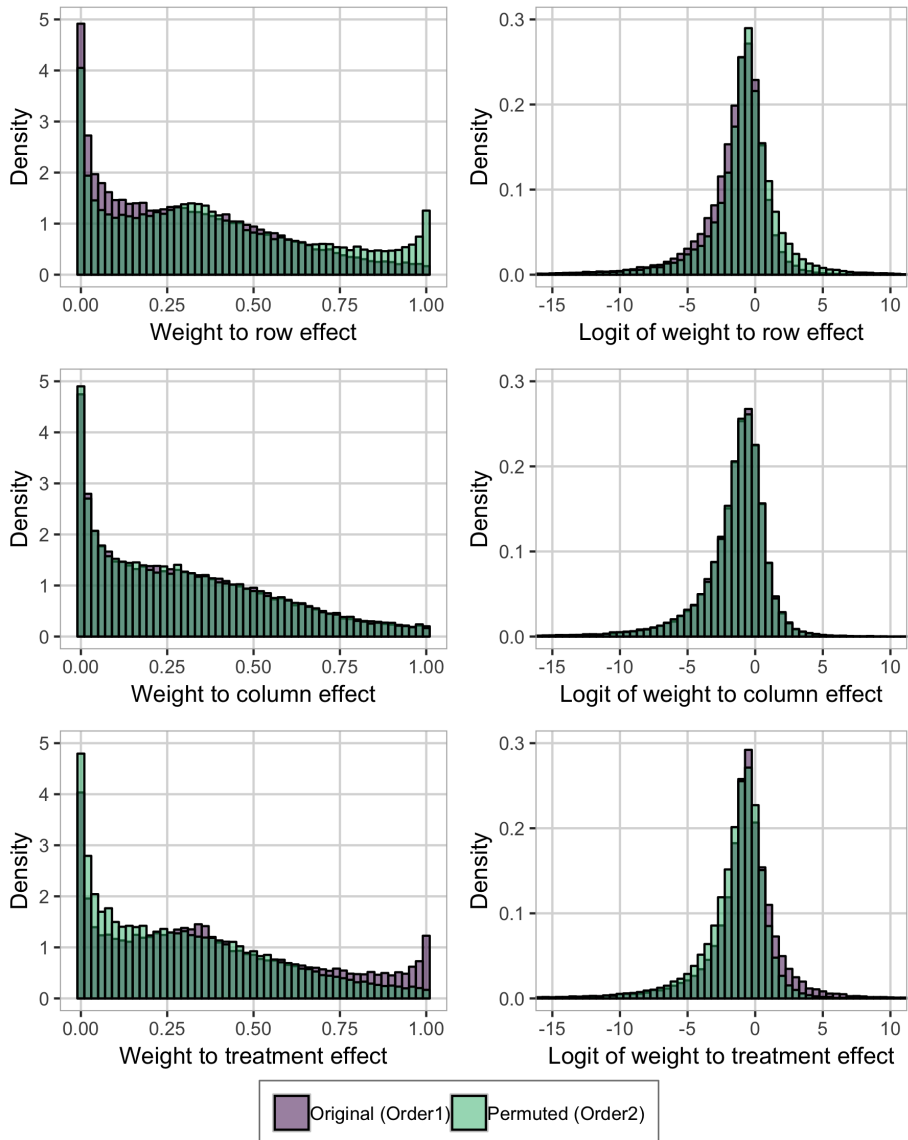


Figure S4.3: Comparison of priors on distribution of total latent variance to row effect, column effect and treatment effect for the original order Order1 and the permuted order Order2. The distributions of the weights to the left, and of the logit weights on the right.

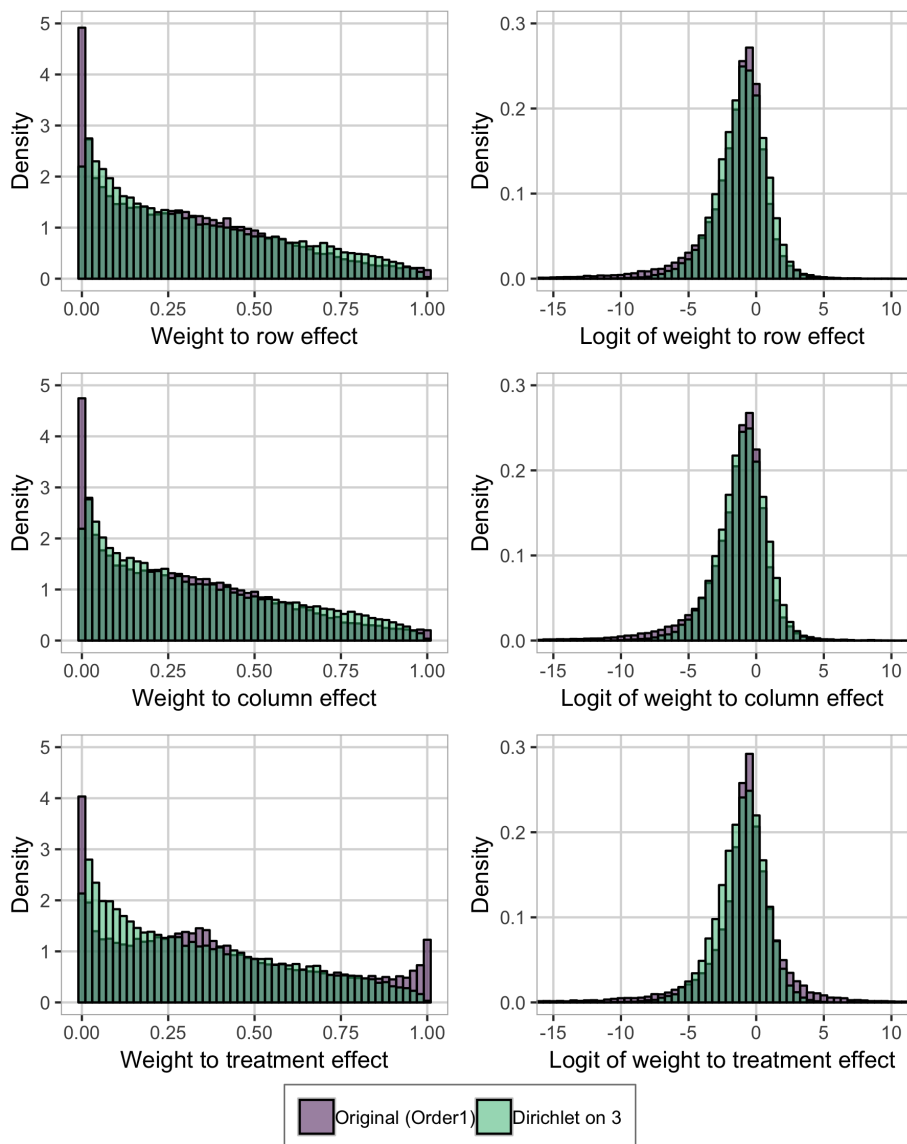


Figure S4.4: Comparison of priors on distribution of total latent variance to row effect, column effect and treatment effect for the original order Order1 and a Dirichlet prior on the triple split. The distributions of the weights to the left, and of the logit weights on the right.

any. The six graphs all show the continuous rank probability score (CRPS) of the structured treatment effect $\gamma^{(1)}$ and the leave-one-out log predictive score (LOO-LPS). In each plot, we have removed the datasets leading to too many (i.e., more than 0.1%) divergent transitions in the inference for at least one of the three priors displayed. The order of the priors is the same in the legend and for each scenario in all plots, so P-INLA is the leftmost, so comes P-HC and so on.

Figure S4.6 shows the results that are also displayed in the main paper: P-INLA gives a lower LOO-LPS, i.e. a poorer model fit, than the other priors. The CRPS is lowest for the HD prior with either triple split implementation for scenarios S2 and S3. Figure S4.7 shows results for varying values of the median ω_m for the prior for selecting between $\gamma^{(1)}$ and $\gamma^{(2)}$ has little effect on the results, and we see that a lower value of the median is slightly better when the true treatment effect is weak, and a higher value is slightly better when the true treatment effect is strong. The difference is however small. Figure S4.8 shows the results when we change the distribution we use on the distance between $\gamma^{(1)}$ and $\gamma^{(2)}$. Changing the exponential prior on the distance between $\gamma^{(1)}$ and $\gamma^{(2)}$ to a gamma prior with shape parameter 0.5 or 0.25, which has a stronger peak at 0, improves results for S1 (see Figure S4.8), but induces more instability in the inference (Figure S4.5). The results are also stable to changes in the hyperparameter for the two dual-splits (Figure S4.9) and changes in the way that the triple-split is implemented; either decomposed into dual-splits in different ways (Figure S4.10) or using a Dirichlet distribution (Figure S4.11).

We have compared the HD prior with a Dirichlet prior on the triple split (P-HD-D3) to HD priors with a Dirichlet prior on a quadruple split between α , β , $\gamma^{(1)}$ and $\gamma^{(2)}$ (P-HD-D4) and between all five effects (P-HD-D5). The two latter perform worse than P-HD-D3 when the treatment effect has no structured contribution, scenario S1, in terms of CRPS (Figure S4.11). Using P-HD-D4 and P-HD-D5 we lose the shrinkage properties between the unstructured and structured treatment effect, so we expect them to perform worse for S1. For S2 and S3 they perform slightly better. The LOO-LPS is not affected noticeably by the implementation of the triple split.

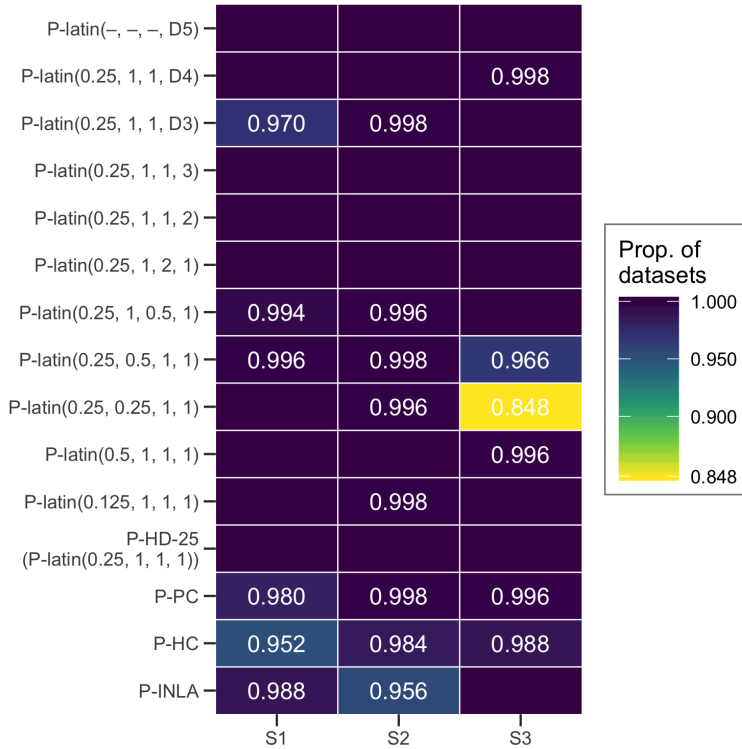


Figure S4.5: The proportion of datasets for each scenario and prior leading to at most 0.1% divergent transitions during the inference in the latin square experiment simulation study. We say that the stability is 1.0 if all datasets for a given prior and scenario lead to no more than 0.1% divergent transitions. No number means that the stability is 1.0. The bottom four priors are the main focus of the study, the top three are the Dirichlet priors, while the middle eight are the HD prior with varying values of ω_m , amount of shrinkage, varying values of λ , and varying ordering of the implementation of the triple split. The notation for the HD prior is P-latin(ω_m , shape parameter, λ , order number/type).

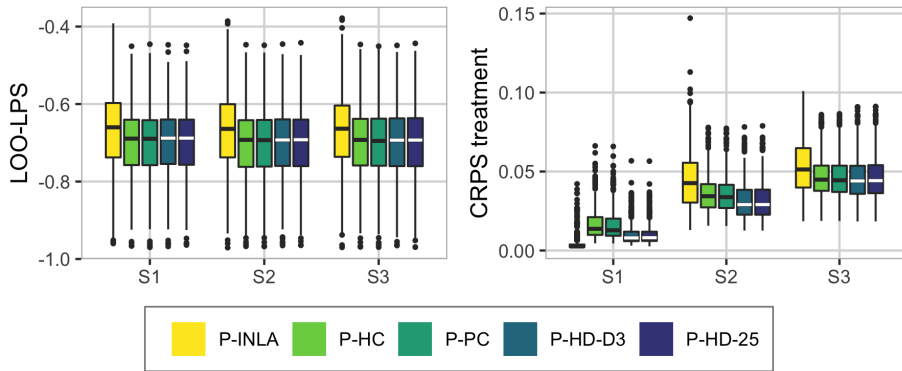


Figure S4.6: Results from the latin square simulation study.

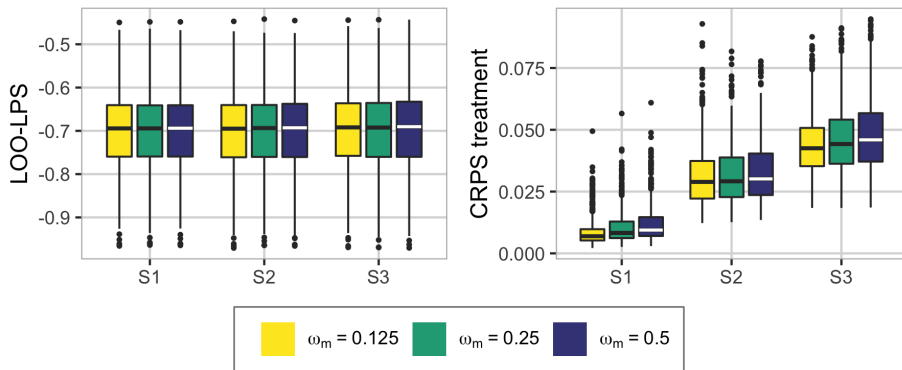


Figure S4.7: Results from the latin square simulation study when varying the position of the median ω_m in the PC prior on the distance between $\gamma^{(1)}$ and $\gamma^{(2)}$. $\omega_m = 0.25$ gives P-HD-25.

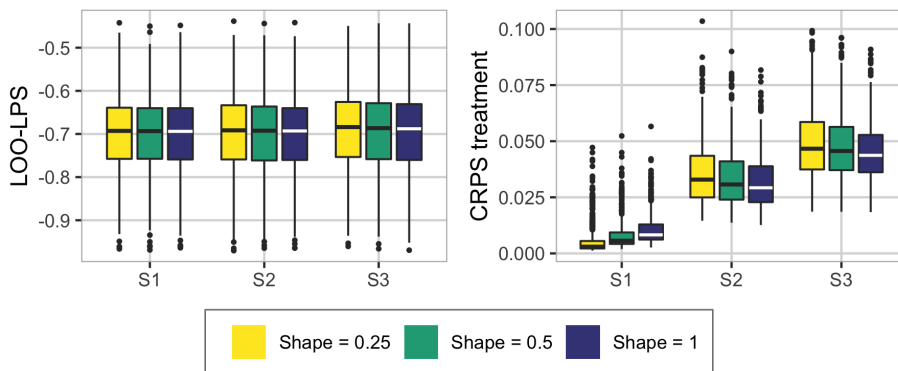


Figure S4.8: Results from the latin square simulation study when varying the shape parameter in the distribution on the distance in the PC prior for the split between unstructured and structured treatment effect. Shape parameter 1 gives the exponential distribution, which gives P-HD-25.

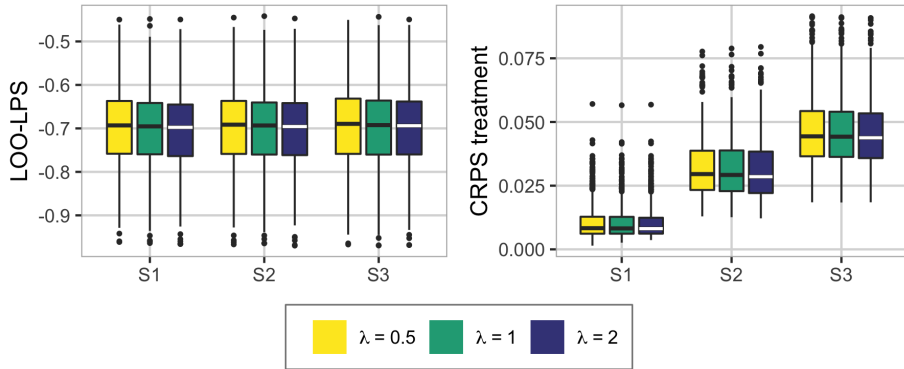


Figure S4.9: Results from the latin square simulation study when varying the value of λ in the PC prior for the multi split. $\lambda = 1$ gives P-HD-25.

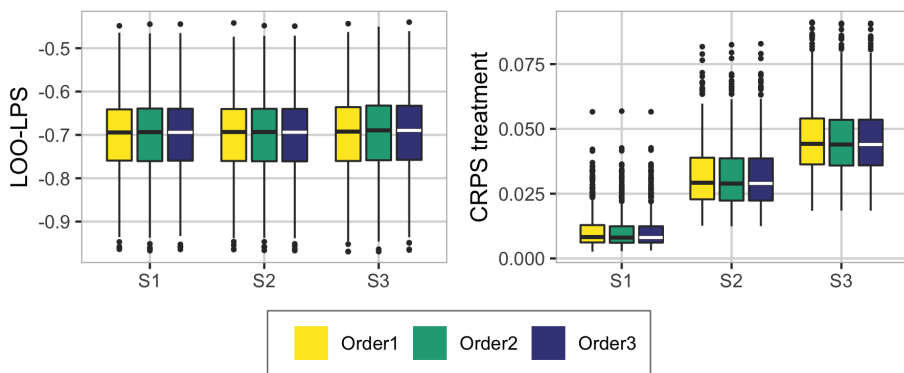


Figure S4.10: Results from the latin square simulation study when varying order of the implementation of the triple split in the PC prior. Order1 gives P-HD-25.

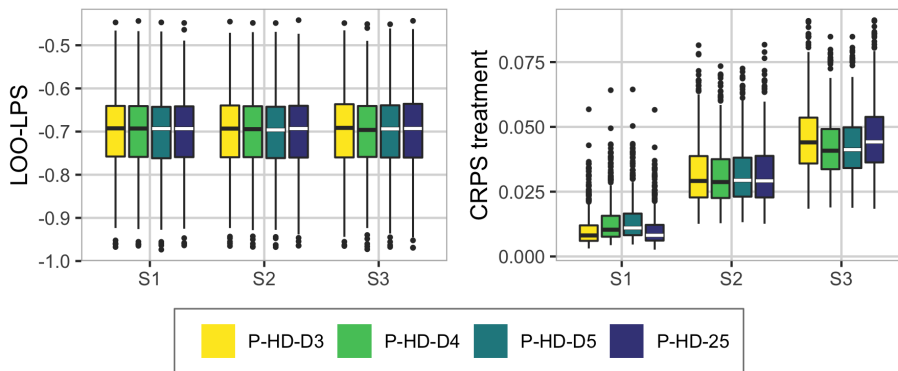


Figure S4.11: Results from the latin square simulation study for the Dirichlet prior. P-HD-D3 has a Dirichlet prior on the split between the row, column and treatment effects, P-HD-D4 has a Dirichlet prior between all effects except the residuals, and P-HD-D5 has a Dirichlet prior on all five random effects (including the residuals). The other weights has PC priors as in P-HD-25.

S4.3 Example

We provide a script for **R** that can be used to simulate data and fit the latin square model. The script is available as part of the Supplementary Materials. This script can be used to look at differences between the priors and the resulting posteriors.

The following priors from the simulation study can be chosen:

1. INLA default (P-INLA)
2. Half-Cauchy (P-HC)
3. Component-wise PC priors (P-PC)
4. HD prior with PC priors on all splits (for example, P-HD-25). Here you can choose to change
 - the median ω_m for the proportion of treatment variance going to the structured effect [0.25 is default],
 - the shape parameter for the gamma distribution on the distance between the unstructured and structured treatment effect [1 is default],
 - A scaling factor for the value of λ used in the multi-splits [1 is default],
 - the ordering of the triple split [1 is default, 2 and 3 are the other orderings].
5. HD prior with a combination of PC and Dirichlet priors (for example, P-HD-D3). Here you can choose to change
 - the number of effects involved in the Dirichlet prior in the HD prior [3 is default, 4 and 5 are the other options].

After the prior has been chosen, the **scenario** can be selected: scenario S1 (no treatment effect), S2 (medium treatment effect) or S3 (strong treatment effect). See Section S4.1 for details. A dataset of the same size as the ones in the simulation study is simulated, and the dataset can be reproduced using a seed value.

Rstan is used for the inference, and you can choose between the following number of samples: "low" (250 (warmup) + 1000, only for testing, this will not give enough samples), "medium" (2500 (warmup) + 10000) or "high" (25000 (warmup) + 100000, this is used in the simulation study in the paper).

The sampler can be run without the likelihood to sample from the prior. A plot of the prior on total weight (the amount of the total variance) for each of the

five effects in the model is available. The prior on total variance and the separate variances for the effects are not shown as they do not have proper priors under the scale-invariant HD priors or Jeffreys' prior on the residual variance.

For the posterior, the following scores and plots are provided:

- The number of divergent transitions that occurred during the inference (see e.g. Section S3.1).
- The posterior total weights for the five model effects and the posterior total variance.
- The posterior standard deviations for the five model effects.
- The posterior mean of the structured treatment effect, with standard deviations, compared to the true effect.
- The average CRPS of the structured treatment effect, see Section 5.2 in the main article for details.
- The LOO-LPS (see Section 5.2 in the main article for details), with corresponding variance of the estimate, and the number of the 81 inferences with more than 1% divergent transitions.

S5 Binomial responses

We include additional background and results from the Kenyan neonatal mortality simulation study and real application presented in Section 6 in the main article.

S5.1 Additional background

The DHS survey from 2014 is stratified by county and urban/rural and has two levels of clustering. Since the counties Nairobi and Mombasa are fully urban, there are in total 92 strata. The households were selected within each stratum through a two-stage clustered sampling design. Kenya was divided into 96251 enumeration areas (EAs) based on the 2009 national census, and the first stage of the sampling design consists of sampling clusters from the list of EAs in the stratum and the second stage consists of sampling households within the selected clusters. Within the selected households all women aged 15–49 who spent the last night in the household are interviewed.

In Section 6.2 in the main paper, we simulate from a model consisting of spatially structured and unstructured random effects and an i.i.d. effect of cluster. Further, preliminary investigations showed that the design with 47 counties provides little information about how the variance should be distributed between the structured and the unstructured spatial effect. Therefore, we use the 290 constituencies of Kenya with 6 clusters per constituency to replicate the size of the survey, but provide a spatial design where the data is more informative about the relative sizes of the unstructured and structured spatial effects. In Section 6.3 in the main paper we analyse the original data on the county-level and include a random effect of household. The key focus of the application is to display how to use and select the new prior, and how the interpretability and transparency of the prior is helpful for assessing and criticising the results.

S5.2 Simulation study

We use the following input values to the function `stan` for the simulation study with neonatal mortality in Kenya: 25 000 samples for burn-in, in total 75 000 samples, one chain thinned to every fifth sample, all parameters initialized to zero, `adapt_delta` equal to 0.95, and default settings for all other input values. The simulation study ran on a computer cluster and takes less than a week, depending on the activity on the cluster.

We include additional results from the Kenya neonatal mortality simulation study. Figure S5.1 shows the proportion of datasets leading to no more than 0.1% divergent transitions during the inference. P-INLA is the only prior which leads to a large number of datasets giving divergent transitions, and mainly in scenario S3, the other three priors give stable inference for all scenarios. Figure S5.2 shows the bias and coverage of μ , the the bias of $\omega^{(1)}$ and $\omega^{(2)}$ and the CRPS of \mathbf{u} , for the five priors we have used in the simulation study. It is only for scenario S3, when the Dirichlet prior is closest to the truth, that P-HD-D is performing better than P-HD-25, in the other scenarios it is doing worse.

Figure S5.2 shows that P-INLA gives way too low coverage for μ , while the other priors leads to a better and similar coverage. For scenarios S2-S5 the true value of the weight is 0.2, P-INLA is for most datasets estimating $\omega^{(1)}$ to be 0, giving a bias of -0.2. The other four priors are all slightly underestimating the weight in S2-S5. P-HD-D is as good as (only scenario S3) or worse than P-HD-25. In scenario S1, the true weight is equal to 1 while the base model is 0, and all priors are underestimating the weight. P-INLA is doing worst with a bias around -0.75 for most datasets, while P-HD-25 is doing a bit better with a bias of around -0.5, and P-HC and P-PC are also underestimating the weight. This may be an indication that we get the prior back, and that the likelihood does not contribute much in the inference.

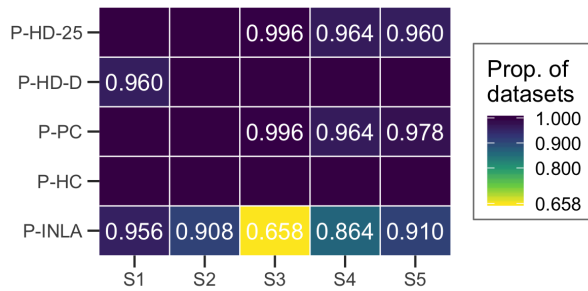


Figure S5.1: The proportion of datasets for each scenario and prior leading to at most 0.1% divergent transitions during the inference in the neonatal mortality in Kenya simulation study. We say that the stability is 1.0 if all datasets for a given prior and scenario lead to no more than 0.1% divergent transitions. No number means that the stability is 1.0.

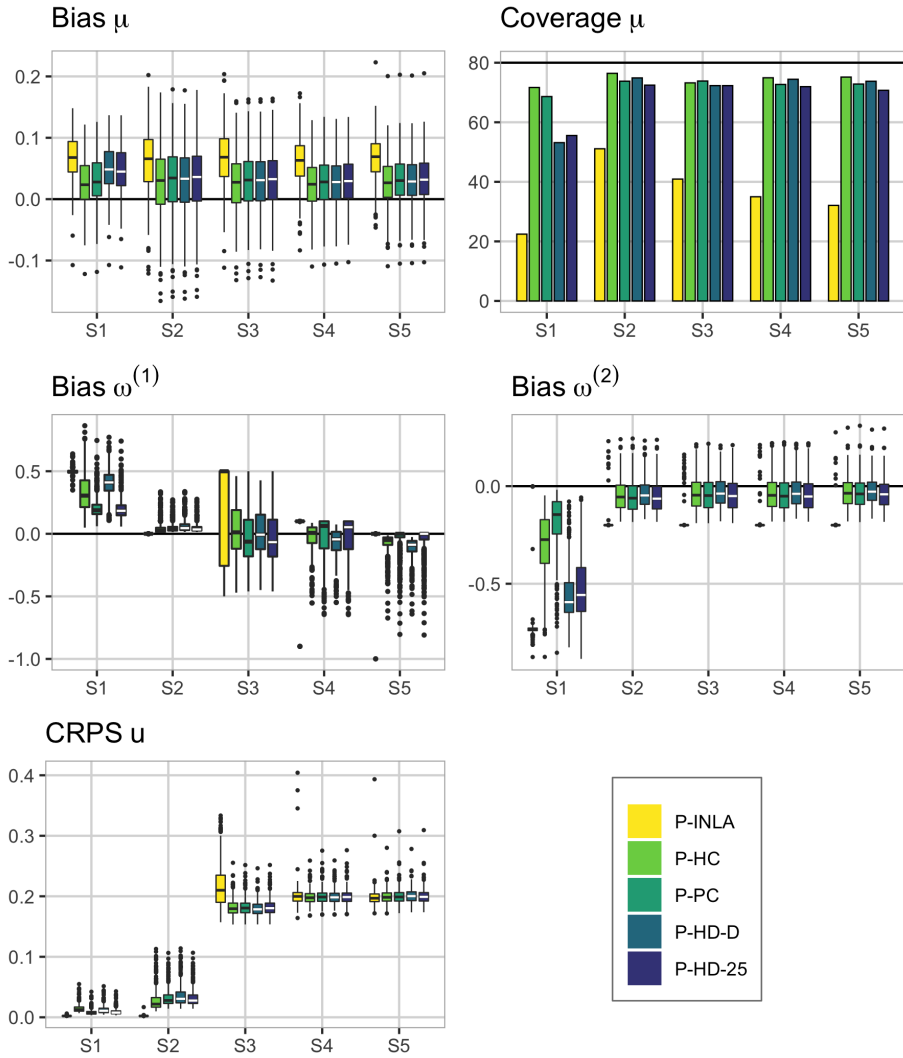


Figure S5.2: Upper left: bias of the intercept μ , upper right: the coverage of μ , mid left: the bias of $\omega^{(1)}$, mid right: the bias of $\omega^{(2)}$, and lower left: CRPS of u . Scenario is indicated at the x-axes. The order of the priors is the same in the legend and for each scenario, so P-INLA is the leftmost, then comes P-HC and so on. The biases are calculated using the estimated median minus the true value, and the coverage is found by counting the number of times the true value lies in the 80% credible interval.

S5.3 Application

The prior and posterior of the total standard deviation from the Kenya neonatal mortality dataset analysis can be seen in Figure S5.3.

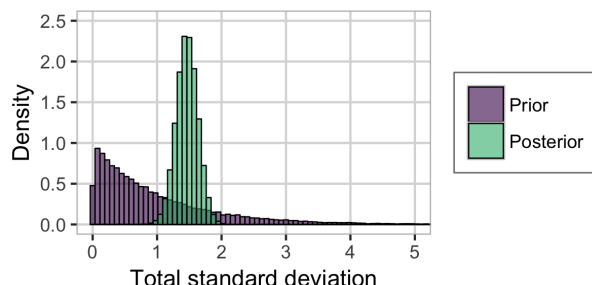


Figure S5.3: The prior and posterior of the total standard deviation σ_T from the analysis of the neonatal mortality in Kenya dataset.

The prior and posterior distributions of the total weight of the unstructured random effects v (unstructured county effect), ν (unstructured cluster effect) and ε (unstructured household effect) can be seen in Figure S5.4. The total weight is $\omega^{(1)}$ for ε , $\omega^{(2)}(1 - \omega^{(1)})$ for ν , and $(1 - \omega^{(3)})(1 - \omega^{(2)})(1 - \omega^{(1)})$ for v . The medians of these three are 0.955, 0.014 and 0.011, respectively. It is clear that the household effect ε explains most of the variance, the cluster effect ν explains some, and the unstructured county effect v explains the least of the three.

Figure S5.5 shows how far a value of 0 is from the posterior median of \mathbf{u} expressed by the posterior tail probability of getting 0 or further away from the median. We see that for many counties the posterior median of u is close to 0 as expressed by the value 0.5 in the figure, and 0 is at the most barely outside the interquartile range as expressed by a value of 0.25.

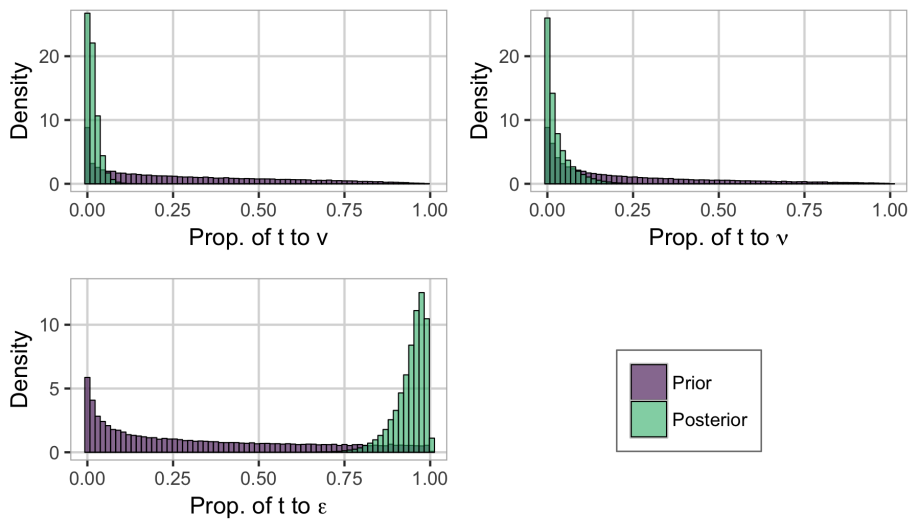


Figure S5.4: The priors and posteriors of the proportion of the total latent variance assigned to the household effect, the cluster effect, and the unstructured spatial effect.

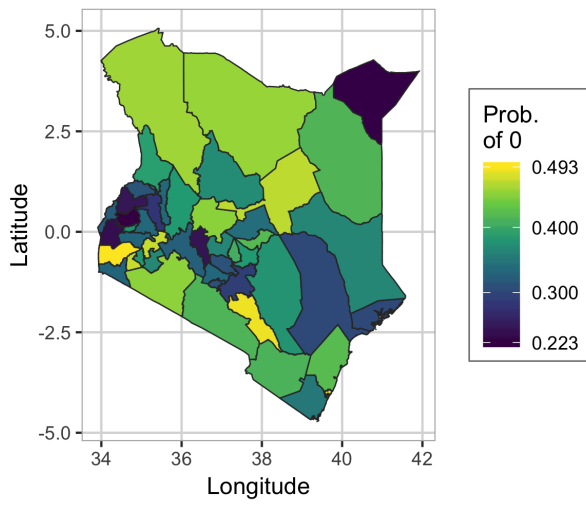


Figure S5.5: The significance of the spatial effect u visualized through the tail probabilities $\text{Prob}(u_i > 0)$ for the counties where the median of u is smaller than 0, and $\text{Prob}(u_i < 0)$ for the counties where the median of u is larger than 0.

References

- 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).
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multi-level/Hierarchical Models*, volume 1. Cambridge University Press, New York, New York.
- 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.
- Stan Development Team (2018a). Brief Guide to Stan’s Warnings. <http://mc-stan.org/misc/warnings.html>. Accessed 2018-12-07.
- Stan Development Team (2018b). RStan: the R interface to Stan. <http://mc-stan.org/>. R package version 2.18.1.

Paper II

Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge

Hem, I. G., Selle, M. L., Gorjanc, G., Fuglstad, G.-A., and Riebler, A.

Genetics, iyab002, 2021. Advance publication.

Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge

Ingeborg Gullikstad Hem¹, Maria Lie Selle¹, Gregor Gorjanc², Geir-Arne Fuglstad¹, and Andrea Riebler¹

¹Department of Mathematical Sciences, NTNU, Norway

²The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK

Abstract

We propose a novel Bayesian approach that robustifies genomic modeling by leveraging expert knowledge through prior distributions. The central component is the hierarchical decomposition of phenotypic variation into additive and nonadditive genetic variation, which leads to an intuitive model parameterization that can be visualised as a tree. The edges of the tree represent ratios of variances, for example broad-sense heritability, which are quantities for which expert knowledge is natural to exist. Penalized complexity priors are defined for all edges of the tree in a bottom-up procedure that respects the model structure and incorporates expert knowledge through all levels. We investigate models with different sources of variation and compare the performance of different priors implementing varying amounts of expert knowledge in the context of plant breeding. A simulation study shows that the proposed priors implementing expert knowledge improve the robustness of genomic modeling and the selection of the genetically best individuals in a breeding program. We observe this improvement in both variety selection on genetic values and parent selection on additive values; the variety selection benefited the most. In a real case study expert knowledge increases phenotype prediction accuracy for cases in which the standard maximum likelihood approach did not find optimal estimates for the variance components. Finally, we discuss the importance of expert knowledge priors for genomic modeling and breeding, and point to future research areas of easy-to-use and parsimonious priors in genomic modeling.

Keywords: Bayesian analysis, expert knowledge, genomic selection, hierarchical variance decomposition, nonadditive genetic variation.

1 Introduction

Plant breeding programs are improving productivity of a range of crops and with this addressing the global and rising hunger problem that impacts 820 million people across the world (FAO et al., 2019). One of the most important food sources in the world is wheat (Shewry and Hey, 2015), however, recent improvements in wheat yield are smaller than the projected requirements (Ray et al., 2013) and might become more variable or even decrease due to climate change (Asseng et al., 2015). This trend is in stark contrast to the United Nation's Sustainable Development Goals that aim to end hunger and malnutrition by 2030 (General Assembly of the United Nations, 2015). Breeding has contributed significantly to the improvement of wheat yields in the past (e.g., Mackay et al., 2011; Rife et al., 2019), and the recent adoption of genomic selection could enable further significant improvements (Gaynor et al., 2017; Belamkar et al., 2018; Sweeney et al., 2019).

Breeding programs generate and evaluate new genotypes with the aim to improve key characteristics such as plant height, disease resistance and yield. Nowadays, a key component in breeding is genomic modeling, where we aim to reduce environmental noise in phenotypic observations and associate the remaining variation with variation in individual genomes. We use these associations to estimate genetic values for phenotyped or even non-phenotyped individuals and with this identify the genetically best individuals (Meuwissen et al., 2001). Improving this process involves improving the methods for disentangling genetic variation from environmental variation.

Genetic variation can be decomposed into additive and nonadditive components (Fisher, 1918; Falconer and Mackay, 1996; Lynch et al., 1998; Mäki-Tanila and Hill, 2014). Additive variation is defined as variation of additive values, which are sums of allele substitution effects over the unobserved genotypes of causal loci. Statistically, the allele substitution effects are coefficients of multiple linear regression of phenotypic values on causal genotypes coded in an additive manner. Nonadditive variation is defined as the remaining genetic variation not captured by the additive values. It is commonly partitioned into dominance and epistasis values. Dominance values capture deviations from additive values at individual loci. Epistasis values capture deviations from additive and dominance values at combinations of loci. Statistically, dominance and epistasis values capture deviations due to allele interactions at individual loci and combinations of loci, respectively. Modeling interactions between two loci at a time gives additive-by-additive, additive-by-dominance and dominance-by-dominance epistasis. Modeling interactions between a larger number of loci increases the number of interactions.

Estimates of genetic values and their additive and nonadditive components have different applications in breeding (Acquaah, 2009). Breeders use estimates of additive values to identify parents of the next generation, because additive values indicate the expected change in mean genetic value in the next generation under the assumption that allele frequencies will not change. Breeders use estimates of genetic values to identify individuals for commercial production, because genetic values indicate the expected phenotypic value. Estimates of genetic values are particularly valuable in plant breeding where individual genotypes can be effectively cloned. While genomic modeling currently focuses on additive values (Meuwissen et al., 2001; Varona et al., 2018), the literature on modeling nonadditive variation is growing (Oakey et al., 2006; Wittenburg et al., 2011; Muñoz et al., 2014; Bouvet et al., 2016; Martini et al., 2017; Vitezica et al., 2017; Varona et al., 2018; de Almeida Filho et al., 2019; Santantonio et al., 2019; Tolhurst et al., 2019; Martini et al., 2020). Notably, modeling nonadditive variation has been shown to improve the estimation of additive values in certain cases (Varona et al., 2018).

However, modeling nonadditive variation is challenging because it is difficult to separate nonadditive variation from additive and environmental variation even when large datasets are available (e.g., Misztal, 1997; Zhu et al., 2015; de los Campos et al., 2019). Further, pervasive linkage and linkage disequilibrium are challenging the decomposition of genetic variance into its components (Gianola et al., 2013; Morota et al., 2014; Morota and Gianola, 2014). This suggests that genomic modeling needs *robust* methods that do not estimate spurious nonadditive values and whose inference is *stable* for all sample sizes.

One way to handle partially confounded sources of variation is to take advantage of expert knowledge on their absolute or relative sizes. Information about the relative magnitude of the sources of phenotypic variation has been collated since the seminal work of Fisher (1918). The magnitude of genetic variation for a range of traits is well known (e.g., Houle, 1992; Falconer and Mackay, 1996; Lynch et al., 1998). Data and theory indicate that the majority of genetic variation is captured by additive values (Hill et al., 2008; Mäki-Tanila and Hill, 2014), while the magnitude of variation in dominance and epistasis values varies considerably due to a range of factors. For example, there is no dominance variation between inbred individuals by definition. Further, model specification has a strong effect on the resulting estimates (e.g., Huang and Mackay, 2016; Martini et al., 2017; Vitezica et al., 2017; Martini et al., 2020). With common model specifications, additive values capture most of the genetic variation because they capture the main effects (in the statistical sense), while dominance and epistasis values capture interaction deviations from the main effects (Hill et al., 2008; Mäki-Tanila and Hill, 2014; Hill and Mäki-Tanila, 2015; Huang and Mackay, 2016). This ex-

pert knowledge does not need to come directly from the literature, it can also be formed based on internal estimates for a similar population in the past, or be a combination of both.

In a Bayesian setting we can take advantage of such expert knowledge through prior distributions; see Gianola and Fernando (1986); Sorensen and Gianola (2007) for an introduction to Bayesian methods in animal breeding and quantitative genetics, respectively. Prior distributions reflect beliefs and uncertainties about unknown quantities of a model and should be elicited from an expert in the field of interest (O’Hagan et al., 2006; Farrow, 2013). Intuitively, prior distributions allow expert knowledge to act as additional observations, and make the current analysis more robust by borrowing strength from past analyses. Priors can improve weak identifiability of the sources of variation by guiding inference towards expert knowledge when the information in the sample is low. However, quantification of the effective number of samples added by a prior is only available in specific situations (Morita et al., 2008).

We propose an easy-to-use, intuitive, and robust Bayesian approach that builds on two recent innovations in Bayesian statistics: 1) the hierarchical decomposition prior framework (Fuglstad et al., 2020) to provide an hierarchical description of the decomposition of phenotypic variation into different types of variation, and 2) the penalized complexity prior framework (Simpson et al., 2017) to facilitate robust genomic modeling. The key ideas of the approach are that (i) visualization eases model specification and communication about the model (see Figure 1), (ii) hierarchical decomposition of variation makes it easy to incorporate expert knowledge on e.g. heritability, (iii) leveraging expert knowledge provides robust methods, and (iv) comparison of posterior distributions and prior distributions reveal the amount of information the data provided on the decomposition of variation.

The aim of this paper is to demonstrate the new approach and to evaluate the potential impact of using the approach along with expert knowledge in plant breeding. We first describe the genomic model and how to incorporate the expert knowledge in this model. To test the proposed approach, we first use a simulated wheat breeding program and evaluate inference stability, estimation of genetic values and variance components with different priors and with the standard maximum likelihood estimation. We also investigate the impact of dataset size. Then we apply the approach to a publicly available wheat yield dataset with 1,739 individuals from 11 different trials in 6 locations in Germany with varying amounts of observed phenotypes from Gowda et al. (2014) and Zhao et al. (2015). We use cross-validation to assess the accuracy of phenotype prediction when using the proposed priors in the model. A description of the simulated and real wheat breeding case studies, model fitting and analysis follows. Our key

focus is to demonstrate how an analyst can take advantage of expert knowledge from literature or domain experts to enable robust genomic modeling of additive and nonadditive variation. This focus involves specifying and visualizing the expert knowledge in an intuitive way. We then present the results and discuss the relevance of our work.

2 Materials and Methods

2.1 Genomic model

We model observed phenotypic values of n individuals $\mathbf{y} = (y_1, \dots, y_n)$ with the aim to estimate their genetic values and their additive and nonadditive components. To this end we use the genomic information about the individuals contained in the observed single nucleotide polymorphism (SNP) matrix \mathbf{Z} , where row i contains SNP marker genotypes for individual i coded additively with 0, 1, 2. We let \mathbf{Z}_a be the column-centered \mathbf{Z} where we have removed markers with low minor allele frequency, and let \mathbf{Z}_d be the column-centered matrix obtained from \mathbf{Z} after setting heterozygote genotypes to 1 and homozygote genotypes to 0.

We model the phenotypic observation y_i of individual i as

$$y_i = \mu + g_i + e_i, \quad i = 1, \dots, n, \quad (1)$$

where μ is an intercept, g_i is the genetic value and e_i the environmental residual for individual i . We model the environmental residual as an independently and identically distributed Gaussian random variable, $\mathbf{e} = (e_1, \dots, e_n) \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$, where σ_e^2 is the environmental variance and \mathbf{I}_n is the $n \times n$ identity matrix. The intercept is typically well-identified from the data, and we specify the nearly translation-invariant prior $\mu \sim \mathcal{N}(0, 1000)$.

We consider the simple additive model with $g_i = a_i$ (*Model A*), and non-additive extension with dominance $g_i = a_i + d_i$ (*Model AD*), and epistasis $g_i = a_i + d_i + x_i$ (*Model ADX*). Here, $\mathbf{a} = (a_1, \dots, a_n)$, $\mathbf{d} = (d_1, \dots, d_n)$ and $\mathbf{x} = (x_1, \dots, x_n)$ respectively denote vectors of the additive, the dominance and the epistasis values for the individuals. Figure 1 shows the model structure for all three models, where every added component extends the model tree by one level. Moving from the root downwards, Model A is defined by the first split. Here only the additive value represents the genetic value. Model AD is defined by the first two splits, and as such has one level more. The genetic value splits into additive and nonadditive values, where only the dominance value represents the nonadditive value. Model ADX is defined by the complete tree and the nonadditive value consists of both dominance and epistasis values.

We model the genetic values as $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 \mathbf{A})$, $\mathbf{d} \sim \mathcal{N}(\mathbf{0}, \sigma_d^2 \mathbf{D})$ and $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbf{X})$, where σ_a^2 , σ_d^2 and σ_x^2 are the additive, dominance and epistasis variances, respectively. We specify the covariance matrices as $\mathbf{A} = \mathbf{Z}_a \mathbf{Z}_a^T / S_a$, $\mathbf{D} = \mathbf{Z}_d \mathbf{Z}_d^T / S_d$ and $\mathbf{X} = \mathbf{A} \odot \mathbf{A} / S_x$ (we consider only additive-by-additive epistasis), where \odot is the Hadamard product (Henderson, 1985; Horn, 1990; Gianola and de los Campos, 2008; Vitezica et al., 2017). To incorporate our expert knowledge in a unified way, we scale the covariance matrices with S_a , S_d , and S_x according to Sørbye and Rue (2018). The idea of such scaling is not new, see Legarra (2016), Vitezica et al. (2017) and Fuglstad et al. (2020) for details. Finally, the phenotypic variance is $\sigma_P^2 = \sigma_g^2 + \sigma_e^2 = \sigma_a^2 + \sigma_d^2 + \sigma_x^2 + \sigma_e^2$.

2.2 Expert knowledge about variance components

As highlighted in the introduction, there is prior information about the relative magnitude of the genetic and environmental variation and the relative magnitude of the additive, dominance and epistasis variation that can guide the construction of prior distributions. We specify this expert knowledge (EK) in a hierarchical manner:

EK-pheno

informs on the split of phenotypic variation into genetic and environmental variation. The proportion of genetic to phenotypic variation is denoted as

$$R_{\frac{g}{g+e}} = \frac{\sigma_g^2}{\sigma_P^2} = h_g^2, \text{ where } h_g^2 \text{ is the broad-sense heritability.}$$

EK-genetic

informs on the split of genetic variation into additive and nonadditive variation. The proportion of additive to genetic variation is denoted as

$$R_{\frac{a}{g}} = \frac{\sigma_a^2}{\sigma_g^2} = \frac{h_a^2}{h_g^2}, \text{ where } h_a^2 \text{ is the narrow-sense heritability.}$$

EK-nonadd

informs on the split of nonadditive variation into dominance and epistasis variation. The proportion of dominance to nonadditive variation is denoted as

$$R_{\frac{d}{d+x}} = \frac{\sigma_d^2}{\sigma_g^2 - \sigma_a^2} = \frac{h_d^2}{h_g^2 - h_a^2}, \text{ where } h_d^2 \text{ is the proportion of dominance to phenotypic variation.}$$

Figure 1 illustrates where the respective expert knowledge in the form of relative magnitudes R_* applies. Of note, for Model A only EK-pheno is used, and EK-genetic is one ($R_{\frac{a}{g}} = 1$) as nonadditive effects are not considered in this model. Similarly, for Model AD only EK-pheno and EK-genetic are used as EK-nonadd is one ($R_{\frac{d}{d+x}} = 1$).

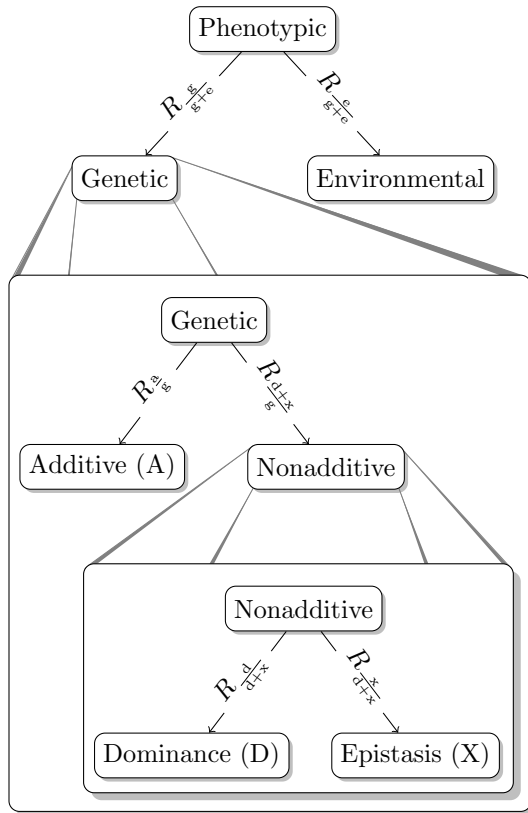


Figure 1: Tree structure visualizing the three possible model formulations A, AD and ADX. Edge labels illustrate where expert knowledge applies, namely on the relative magnitude of the genetic and environmental variation and the relative magnitude of the additive, dominance and epistasis variation.

Values for the relative magnitudes R_* will vary between study systems and traits in line with the expert knowledge. In this study our knowledge is based on the cited literature in the introduction and practical experience with the analysis of a range of datasets. We follow the fact that many complex traits in agriculture are under sizeable environmental effect and that additive effects capture most genetic variation by standard quantitative model construction. With this in mind we assume EK-pheno to be $R_{\frac{g}{g+e}} = 0.25$, EK-genetic to be $R_{\frac{a}{g}} = 0.85$ and EK-nonadd to be $R_{\frac{d}{d+x}} = 0.67$. This implies $R_{\frac{d}{g}} = 0.15 \cdot 0.67 \approx 0.10$ and $R_{\frac{x}{g}} = 0.15 \cdot 0.33 \approx 0.05$. We emphasize that we use this information to construct prior distributions, i.e., these relative magnitudes are only taken as a guide and not as the exact magnitude of variance components. Fuglstad et al. (2020) show that the prior for the first partition, the broad-sense heritability h_{g}^2 , is not very influential.

We present two approaches for constructing a prior based on EK-pheno, EK-genetic and EK-nonadd: 1) a component-wise (comp) prior, that is placed independently on each variance parameter; and 2) a tree-based (tree) model-wise prior, that is defined jointly for all variance parameters. Both approaches are motivated by the concept of penalized complexity priors (Simpson et al., 2017).

2.3 Penalized complexity priors

A penalized complexity (PC) prior for a parameter θ is typically controlled by: 1) a preferred parameter value θ_0 which is intuitive or leads to a simpler model; and 2) an idea on how strongly we believe in θ_0 . The PC prior shrinks towards θ_0 , unless the the data indicate otherwise. This is achieved by constructing the prior based on a set of well-defined principles, for details we refer to Simpson et al. (2017). PC priors can be applied to a standard deviation or variance, a proportion of variances, or other parameters such as correlations (Guo et al., 2017).

The PC prior for a standard deviation (σ) of a random effect will shrink the standard deviation towards zero, that is, towards a simpler model without the corresponding random effect (assuming the prior mean of the effect is zero). This prior is denoted as $\sigma \sim \text{PC}_0(\sqrt{V}, \alpha)$ and results in an exponential distribution with rate parameter $-\ln(\alpha)/\sqrt{V}$. The subscript 0 in $\text{PC}_0(\cdot)$ indicates that the prior shrinks towards $\sigma = 0$. To define the prior the analyst has to specify an upper value \sqrt{V} and a tail probability α such that the upper-tail probability $P(\sigma > \sqrt{V}) = \alpha$. Here, we use $\alpha = 0.25$ so the prior distribution is weakly-informative towards \sqrt{V} , but shrinks to zero unless the data informs otherwise.

For a variance proportion $p \in [0, 1]$ we denote the PC prior as $p \sim \text{PC}_0(R)$.

The numerical value $R \in [0, 1]$ encodes the available expert knowledge about the proportion and is set as the median of the prior, i.e. $P(p > R) = 0.5$. The subscript 0 indicates that the prior shrinks towards $p = 0$. Shrinkage towards the median is achieved by the PC prior $p \sim \text{PC}_M(R)$, where R has the same interpretation as for $\text{PC}_0(R)$. For $\text{PC}_M(R)$, we need to specify how concentrated the distribution is on logit-scale in the interval $[\text{logit}(R) - 1, \text{logit}(R) + 1]$ around the median (see Fuglstad et al. (2020) for details). We allocated 75% probability to this interval.

The PC prior for a variance proportion depends on the structure of the two random components that are involved through their covariance matrices. We omit this in the notation for simplicity, and to emphasize that we chose to make the marginal priors on the proportions independent of each other. As the PC prior on proportions depends on the covariance matrix structure, it is application specific, and the priors do not correspond to common families of distributions such as the exponential or normal distributions (see Riebler et al. (2016); Fuglstad et al. (2020) for more details).

2.4 Component-wise prior

In the component-wise setting we use a PC prior for each standard deviation parameter σ_* . The PC prior on σ_* requires an upper value $\sqrt{V_*}$, so in addition to the relative magnitudes specified through EK-pheno, EK-genetic and EK-nonadd we need information on the magnitude of the phenotypic variance to set up the component-wise priors. For this purpose we could calculate the empirical phenotypic variance V_P from a separate dataset, which is a trial study or a study believed to exhibit similar phenotypic variance as the study at hand. From this we can define the upper values for the individual PC priors. For example, to formulate priors for Model A we use EK-pheno to find $\sigma_a \sim \text{PC}_0\left(\sqrt{h_g^2 V_P}, 0.25\right)$ and $\sigma_e \sim \text{PC}_0\left(\sqrt{(1 - h_g^2) V_P}, 0.25\right)$. For Model AD we need EK-pheno and EK-genetic to formulate the priors, and for Model ADX, the most complex model, we take into account all available expert knowledge.

We follow the tree-structure shown in Figure 1 downwards to define the upper values, and multiply the relative magnitudes on the edges leading to the respective leaf nodes. For Model ADX this leads us to:

- $\sigma_e \sim \text{PC}_0\left(\sqrt{(1 - h_g^2) V_P}, 0.25\right)$,
- $\sigma_a \sim \text{PC}_0\left(\sqrt{h_a^2 V_P}, 0.25\right)$,

- $\sigma_d \sim \text{PC}_0 \left(\sqrt{h_d^2 V_P}, 0.25 \right)$, and
- $\sigma_x \sim \text{PC}_0 \left(\sqrt{(h_g^2 - h_a^2 - h_d^2) V_P}, 0.25 \right)$.

Combining the available expert knowledge procedure with the three different genomic models gave us settings we denote as A-comp*, AD-comp* and ADX-comp*. We have contrasted these settings with a *default* component-wise PC prior proposed by Simpson et al. (2017) with $\sqrt{V} = 0.968$ and $\alpha = 0.01$ on all variance parameters, which gave us settings denoted as A-comp, AD-comp and ADX-comp. This default prior is a prior without any expert knowledge. Preliminary analyses showed that the inferences for AD-comp, AD-comp*, ADX-comp and ADX-comp* are not stable, i.e. the methods are not robust in the sense that they did not avoid estimating spurious nonadditive effects, and we do not present results from these settings. The priors for A-comp* and A-comp are plotted in Figure S1 in File S1 in the Supplemental materials. using $h_g^2 = 0.25$ and $V_P = 1$. If V_P takes another value, we simply rescale the x - and y -axes; the shape of the prior stays the same. In the simulated case study, we will use $V_P = 1.86$. The priors are equal on all standard deviations for A-comp, AD-comp and ADX-comp. The priors for AD-comp* and ADX-comp* can be seen in Figures S2 and S3 in File S1. See Note S1 in File S1 for a detailed description of the component-wise prior and posterior distributions for Model A and Model AD.

2.5 Tree-based model-wise prior

In the model-wise setting, we shift the focus in Figure 1 from the leaf nodes to the splits. In other words, a shift from the component-wise perspective of variances associated with different sources of variation to a model-wise perspective of how these variances arise as a hierarchical decomposition of the phenotypic variance. This provides a complementary way to construct priors where EK-pheno, EK-genetic and EK-nonadd are directly incorporated at the appropriate levels in the tree structure. We achieve this by applying the hierarchical decomposition (HD) prior framework of Fuglstad et al. (2020). We focus the presentation on the essential ideas for understanding and successfully applying the priors, and provide the comprehensive and mathematical description in Note S1 in File S1. We emphasize that in the following p_* denotes an actual variance proportion that we will infer (along with variances), while R_* denotes expert knowledge for this proportion.

We first assign a marginal prior for the decomposition of variances in the lowest split, and then move step-wise up the tree assigning a prior to the de-

composition of variance in each split conditional on the splits below it. The bottom-up process ends with the assignment of a prior to the root split, and the result is a joint prior for the decomposition of phenotypic variance into the different sources of variance. In the final step, we assign a prior for phenotypic variance σ_P^2 that is conditionally independent of the prior on the decomposition of the phenotypic variance.

We follow Fuglstad et al. (2020) and simplify the prior at each split by conditioning on expert knowledge from the lower splits. For example, the prior for $p_{\frac{a}{g}}$ is constructed under the assumption that $p_{\frac{d}{d+x}} = R_{\frac{d}{d+x}}$; that is, $\pi(p_{\frac{a}{g}} | p_{\frac{d}{d+x}})$ is replaced with $\pi(p_{\frac{a}{g}} | p_{\frac{d}{d+x}} = R_{\frac{d}{d+x}})$. Note that even though we construct the prior from the bottom and up, the arrows in the tree indicate how the phenotypic variance is distributed in the model from the top down. This means that the amount of, for example, dominance variance σ_d^2 depends on the variance partitions further up, since $\sigma_d^2 = \sigma_P^2 p_{\frac{g}{g+e}} (1 - p_{\frac{a}{g}}) p_{\frac{d}{d+x}}$ following the tree structure (Figure 1).

In this study, we assumed that at the lower levels the model shrinks towards the expert knowledge EK-nonadd and EK-genetic by using $PC_M(\cdot)$ priors. Further, at the top level we use a $PC_0(\cdot)$ prior to shrink towards the environmental effect unless the data indicates otherwise to reduce overfitting. Note that we could have chosen different assumptions. To obtain a prior fulfilling our assumptions, we follow four steps:

1. we use a $PC_M(\cdot)$ prior for the proportion of dominance to nonadditive variance with median $R_{\frac{d}{d+x}} = \frac{h_d^2}{h_g^2 - h_a^2}$ (EK-nonadd),
2. we use a $PC_M(\cdot)$ prior for the proportion of additive to genetic variance with median $R_{\frac{a}{g}} = \frac{h_a^2}{h_g^2}$ (EK-genetic),
3. we use a $PC_0(\cdot)$ prior for the proportion of genetic to phenotypic variance with median $R_{\frac{g}{g+e}} = h_g^2$ (EK-pheno), and
4. we achieve scale-independence through the non-informative, scale-invariant Jeffreys' prior for the phenotypic variance $\sigma_P^2 \sim 1/\sigma_P^2$.

This construction gives the joint prior

$$\pi(\sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}, p_{\frac{d}{d+x}}) = \pi(\sigma_P^2) \pi(p_{\frac{g}{g+e}}) \pi(p_{\frac{a}{g}}) \pi(p_{\frac{d}{d+x}})$$

for Model ADX, where the conditioning on expert knowledge from lower splits is omitted to simplify notation. We denote this setting as ADX-tree* and show this prior in Figure 2 for $R_{\frac{g}{g+e}} = 0.25$, $R_{\frac{a}{g}} = 0.85$ and $R_{\frac{d}{d+x}} = 0.67$. Note that the model-wise priors with expert knowledge are dependent on the covariance

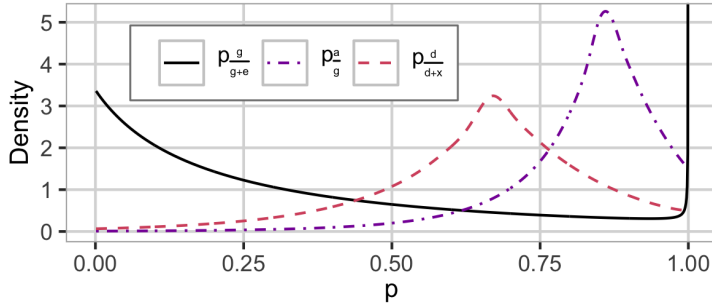


Figure 2: The HD prior used in the ADX-tree^{*a} setting with the proportion of genetic to phenotypic variance $p \frac{g}{g+e}$, additive to genetic variance $p \frac{a}{g}$ and dominance to nonadditive variance $p \frac{d}{d+x}$. We use $R_{\frac{g}{g+e}} = 0.25$, $R_{\frac{a}{g}} = 0.85$, and $R_{\frac{d}{d+x}} = 0.67$. This is a dataset specific prior.

^aAdditive and nonadditive model with model-wise expert knowledge prior.

matrices of the modelled effects and are therefore dataset specific (Fuglstad et al., 2020), and the plots of these priors thus pertain to one specific dataset. The spike at $p = 1$ for $p \frac{g}{g+e}$ in Figure 2 is an artifact of the parameterization chosen for visualization and does not induce overfitting; see Fuglstad et al. (2020) for details. See Note S1 in File S1 for a detailed description of the model-wise prior and posterior distributions for Model A and Model AD.

We explored the influence of alternative expert knowledge. In addition to the previously stated values for EK-pheno, EK-genetic and EK-nonadd we also tested $R_{\frac{g}{g+e}} = 0.25$, $R_{\frac{a}{g}} = 0.05$, and $R_{\frac{d}{d+x}} \approx 0.11$ (so $R_{\frac{d}{g}} \approx 0.95 \cdot 0.11 \approx 0.10$ and $R_{\frac{x}{g}} \approx 0.95 \cdot 0.89 \approx 0.85$). The constructions follows the description above but with these relative magnitudes instead. We denote this setting as ADX-tree^{opp*}, as it expresses almost opposite or ”wrong” beliefs compared to ADX-tree^{*} setting, and show the prior in Figure S4 in File S1 in the Supplemental materials.

For Model AD the nonadditive effect only consists of dominance, and the variance is attributed to the different effects as visualized by the top and middle split in Figure 1. We construct a prior using EK-pheno and EK-genetic with $R_{\frac{g}{g+e}} = 0.25$ and $R_{\frac{a}{g}} = 0.85$ and denote this setting AD-tree^{*}. The prior is shown in Figure 3.

For Model A the genetic variance is not decomposed to different sources and the distribution of the phenotypic variance can be visualized using the top split in Figure 1. We use EK-pheno with $R_{\frac{g}{g+e}} = 0.25$ to construct a prior for the

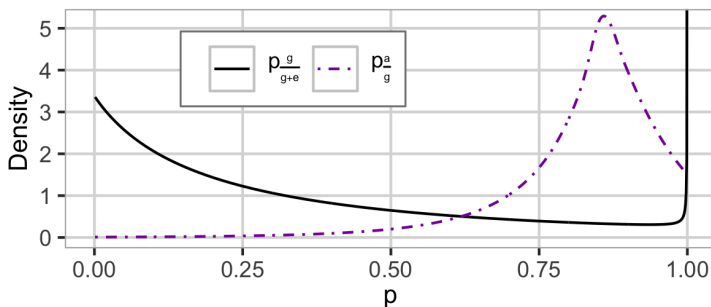


Figure 3: The HD prior used in the AD-tree*^a setting with the proportion of genetic to phenotypic variance $p_{\frac{g}{g+c}}$ and additive to genetic variance $p_{\frac{a}{g}}$. We use $R_{\frac{g}{g+c}} = 0.25$ and $R_{\frac{a}{g}} = 0.85$. This is a dataset specific prior.

^aAdditive and dominance model with model-wise expert knowledge prior.

proportion of genetic to phenotypic variance and denote this setting as A-tree*. We show this prior in Figure 4.

We compared the model-wise prior with expert knowledge to a default prior with no expert knowledge by constructing an HD prior using the exchangeable Dirichlet prior on the proportion of phenotypic variance attributed to each of the sources of variance following Fuglstad et al. (2020). For Model A we use a uniform prior, which is a special case of the symmetric Dirichlet(m) prior with $m = 2$, on the proportion of genetic to phenotypic variance $p_{\frac{g}{g+c}}$ and denote this setting as A-tree (see Figure 4). For Models AD and ADX we use Dirichlet(3) and Dirichlet(4) priors on the proportions, respectively, and denote these settings AD-tree and ADX-tree. These priors do not induce shrinkage towards any of the effects, and assume that each model effect contributes equally to the phenotypic variance, which due to the lack of expert knowledge did not lead to stable inference for Models AD and ADX. We do not show results from these settings. The tree structure and prior for AD-tree and ADX-tree can be seen in Figures S5 and S6 in File S1, respectively. We summarize the model-wise priors that will be used in the following in Table 1.

2.6 Simulated case study

We applied the described genomic model (1) with the above mentioned priors to a simulated case study that mimics a wheat breeding program to investigate the

Table 1: Summary of the model-wise (tree-based) prior distributions on proportions^{a, b} and total phenotypic variance.

Additive (Model A, $g_i = a_i$)		Additive and dominance (Model AD, $g_i = a_i + d_i$)	Additive and nonadditive (Model ADX, $g_i = a_i + d_i + x_i$)
<i>Default</i>	<i>Expert</i>	<i>Expert</i>	<i>Expert</i>
A-tree: $p_{\frac{g}{g+e}} \sim \text{Dirichlet}(2)$ $\sigma_p^2 \sim \text{Jeffreys}'$	A-tree*: $p_{\frac{g}{g+e}} \sim \text{PC}_0\left(R_{\frac{g}{g+e}}\right)$ $\sigma_p^2 \sim \text{Jeffreys}'$	AD-tree*: $p_{\frac{g}{g+e}} \sim \text{PC}_0\left(R_{\frac{g}{g+e}}\right)$ $p_{\frac{d}{g}} \sim \text{PC}_M\left(R_{\frac{d}{g}}\right)$ $\sigma_p^2 \sim \text{Jeffreys}'$	ADX-tree*: $p_{\frac{g}{g+e}} \sim \text{PC}_0\left(R_{\frac{g}{g+e}}\right)$ $p_{\frac{d}{g}} \sim \text{PC}_M\left(R_{\frac{d}{g}}\right)$ $p_{\frac{d}{d+x}} \sim \text{PC}_M\left(R_{\frac{d}{d+x}}\right)$ $\sigma_p^2 \sim \text{Jeffreys}'$

^a $p \sim \text{PC}_0(R)$ describes a PC prior for a variance proportion that has median equal to R and a preference for the variance proportion being equal to 0.

^b $p \sim \text{PC}_M(R)$ describes a PC prior for a variance proportion with median R and a preference for the variance proportion being equal to the median R , with 75% probability in $[\text{logit}(R) - 1, \text{logit}(R) + 1]$ around the median on logit-scale.

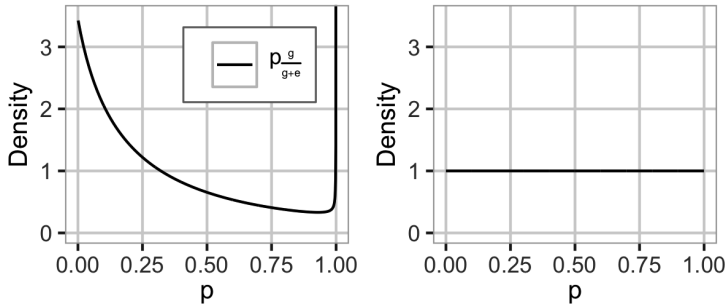


Figure 4: The prior for the proportion of genetic to phenotypic variance $p_{\frac{g}{g+e}}$ for the A-tree*^a (left) and A-tree^b (right) settings. We use $R_{\frac{g}{g+e}} = 0.25$. A-tree* is a dataset specific prior.

^aAdditive model with model-wise expert knowledge prior.

^bAdditive model with model-wise default prior.

properties of the different settings. We simulated the breeding program using the R package AlphaSimR (Faux et al., 2016; Gaynor, 2019) and closely followed the breeding program descriptions of Gaynor et al. (2017) (see their Figure 3) and Selle et al. (2019).

Specifically, we simulated a wheat-like genome with 21 chromosomes, selected at random, 1,000 SNP markers and 1,000 causal loci from each chromosome and used this genome to initiate a breeding program with breeding individuals. Every year we have used 50 fully inbred parents to initiate a new breeding cycle with 100 random crosses. In each cross we have generated 10 progenies and selfed them to generate 1,000 F2 (second filial) individuals, which were selfed again to generate 10,000 F3 (third filial) individuals. We reduced the 10,000 F3 individuals in four successive selection stages (headrow, preliminary yield trial, advanced yield trial and elite yield trial) with 10% selection intensity in each stage. In the headrow stage, we simulated a visual selection on a phenotype with the heritability of 0.03. For the preliminary, advanced and elite yield trials we respectively simulated selection on phenotype with heritability 0.25, 0.45 and 0.62. We used the 50 individuals with the highest phenotype values from the last three selection stages as parents for the next breeding cycle.

We ran the simulation for 30 years. At year 1, we set the following variances for the population of the 50 parents: additive variance of 1.0, dominance variance of 0.5, and epistasis variance of 0.1. We set the environmental variance to 6.0 for all stages and years. We ran the simulation for 20 years as a "burn-in" to obtain

realistic breeding data under selection. We then ran the simulation for another 10 years with selection on phenotype. We removed the SNP markers with minor allele frequency below 5%. We did not use the models for selection.

2.7 Real case study

We also applied the described genomic model (1) to the publicly available Central European wheat grain yield data from Gowda et al. (2014) and Zhao et al. (2015). In short, the data consists of 120 female and 15 male parent lines, which were crossed to create 1,604 hybrids. The parents and hybrids were phenotyped for grain yield in 11 different trials in 6 locations in Germany. The number of observed phenotypes for the parents and hybrids vary between the trials, i.e., some datasets have more observed phenotypes than others, ranging from 834 to 1,739 (see Table S1 in File S1 in the Supplemental materials). The parents and hybrids have genotype data for 17,372 high-quality SNP markers.

In the real case study we analyzed the performance of the tree-based priors using expert knowledge (tree*) for the additive model (A), the additive and dominance model (AD), and the additive and nonadditive model (ADX). We used the same as in the simulation study: $R_{\frac{a}{g}} = 0.85$ and $R_{\frac{d}{d+x}} = 0.67$. We have however used a higher value in EK-pheno, $R_{\frac{g}{g+e}} = 0.75$, in line with Reif et al. (2011) - later stage trials tend to have higher heritability than early stage trials. Again, we emphasize that these values are only used to construct prior distributions and are not taken as literal proportions. The resulting priors can be seen in Figure S7 in File S1.

2.8 Implementation details

We fitted the models with a Bayesian approach through the R package `rstan` (Carpenter et al., 2017; Stan Development Team, 2019). This package provides a sampling algorithm that uses the No-U-Turn sampler, a variant of Hamiltonian Monte Carlo, and only requires that the user specifies the joint posterior distribution up to proportionality, without having to write a sampling algorithm. See Note S1 in File S1 in the Supplemental materials for details. Sampling methods such as Markov Chain Monte Carlo and Hamiltonian Monte Carlo have guaranteed asymptotic accuracy as the number of drawn samples go to infinity. However, in an applied context with finite computational resources, it is hard to assess this accuracy. Betancourt (2016) developed a diagnostic metric for the Hamiltonian Monte Carlo, called divergence, that indicates whether the sampler is able to transition through the posterior space effectively or not, where in the latter case

the results might be biased (we show an example on this in the results).

We also fitted Models A, AD and ADX with the maximum likelihood (ML) approach using the low-storage BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm through the R package `nloptr` (Nocedal, 1980; Liu and Nocedal, 1989; Johnson, 2020). This approach does not use priors. We denote them as A-ML, AD-ML and ADX-ML and use them as a base-line for comparison because maximum likelihood is a common approach in the literature. Inference for ADX-ML was not robust, and we do not present results from this setting. At last, we compared the model results to the performance of selection based solely on phenotype where we treat the phenotype as a point estimate of the genetic value.

2.9 Performance assessment

For the simulated case study, we ran the breeding program simulation 4,000 times and fitted the model and prior settings in each of the last 10 years of simulation (40,000 model fits in total) at the third selection stage (advanced yield trial) in the program. Here we had 100 individuals each with five replicates and the goal was to select the 10 genetically best individuals for the fourth, last, stage. For each model fit we evaluated: (i) robustness of method, (ii) the accuracy of selecting the genetically best individuals, (iii) the accuracy of estimating the different genetic values and (iv) the accuracy of estimating the variance parameters. We evaluated the fits against the true (simulated) values.

We measure how *robust* the method (model and inference approach) is, i.e., to which degree it avoids estimating spurious nonadditive effects, in *stability of inference*. For the stability of inference of the Bayesian approach with Stan we used the proportion of analyses that had stable inference (which we define as at least 99% samples where no divergent transitions were observed) for each model and prior setting. For the stability of inference of the maximum likelihood approach we used the proportion of analyses where the optimizer algorithm converged.

For the accuracy of selecting the genetically best individuals we ranked the best 10 individuals based on the estimated genetic value or estimated additive value, and counted how many were among the true genetically best 10 individuals based on the true genetic value or true additive value. We used the posterior mean of the effects as estimated values for ranking. Selection on the genetic value indicated selection of new varieties, while the selection on the additive value indicated selection of new parents.

For the accuracy of estimating the different genetic values (total genetic, additive, dominance and epistasis values) we used Pearson correlation and continuous rank probability score (CRPS, Gneiting and Raftery, 2007). With the correla-

tion we measured how well posterior means of genetic values correlated with true values (high value is desired). This metric works with point estimates and ignores uncertainty of inferred posterior distributions of each individual genetic value. The CRPS is a proper scoring rule and as such measures a combination of bias and sharpness of the posterior distribution compared to true values (low value is desired). Specifically, CRPS integrates squared difference between the cumulative estimated posterior distribution and the true value over the whole posterior distribution (Gneiting and Raftery, 2007). See Selle et al. (2019) for a detailed explanation of CRPS used in a breeding context. In the case of phenotypic selection, we have a phenotype value for selection candidates, which is a point estimate of the genetic value, and its CRPS then reduces to the mean absolute error between the true genetic values and the phenotype.

The accuracy of the estimates of the variance parameters was assessed by dividing them by the true genetic variances for each of the 10 years from the simulated breeding program (a value close to 1 is desired). This is not done for phenotype selection.

To test the effect of dataset size on inference, we ran the breeding program an additional 1,000 times and fitted the models to $n = 700, 600, \dots, 100$ individuals in the preliminary stage (instead to 100 individuals in the advanced stage) at year 21. We used the settings with tree-based expert knowledge priors and the maximum likelihood approach and investigated the performance of the methods for increasing number of observations by evaluating the robustness, and the accuracy of estimating the different genetic values and variance parameters.

We analyzed the real case study with the same models and tree-based expert knowledge priors and focused on the ability of predicting observed phenotypes in a cross-validation scheme. We performed 5-fold cross-validations five times for each of the 11 trials independently. For each fold in each cross-validation, we predicted the held-out phenotypes (their posterior distribution involving intercept, genetic value and environmental variation), and calculated the correlation between the point predictions and the observed phenotypes, and the CRPS using the phenotype posterior prediction distributions and the observed phenotypes available for each trial. We note that phenotype posterior predictions involve environmental variation, which does not affect point predictions and correlations, but affects the CRPS as the whole distribution of the prediction is involved in the calculations. We also looked at the posterior medians of the model variances. Of note, in contrast to the simulated case study the genetic effects are unknown for real data, so that we cannot assess the estimation accuracy of the effects.

2.10 Data and code availability

We provide code to simulate the data described in the simulated case study (File S2 in the Supplemental materials). We also provide example code to fit the models presented in this paper together with an example dataset (File S3). In the real case study we used data from Gowda et al. (2014) (SNP genotypes) and Zhao et al. (2015) (phenotypes), and provide code for fitting the models in File S4, including the folds used in the cross-validation. The Supplemental materials are available at figshare: <https://doi.org/10.6084/m9.figshare.12040716>.

3 Results

3.1 Simulated case study

In the simulated case study the model-wise priors and expert knowledge improved the inference stability of the nonadditive models and the selection of the genetically best individuals, but did not significantly improve the accuracy of estimating different genetic values for all individuals or for variance components.

3.1.1 Robustness and stability:

Table 2 shows the proportion of stable model fits by model and prior setting. The model-wise priors combined with expert knowledge improved the inference stability of the additive and dominant (AD) model and the nonadditive (ADX) model to the level of stability of the additive (A) model and phenotypic selection. Phenotypic selection does not depend on a model fit to a dataset and therefore had the highest method robustness by definition. This maximum level of robustness was matched by the simple additive model with the model-wise prior with and without using expert knowledge (A-tree* and A-tree) and with the standard maximum likelihood approach (A-ML). This high model robustness was followed closely by fitting the more complicated nonadditive and additive and dominance models with model-wise prior and expert knowledge (ADX-tree* and AD-tree*). The Bayesian approach using component-wise priors with expert knowledge (A-comp*), the additive and dominance model with the maximum likelihood approach (AD-ML), the component-wise priors without expert knowledge (A-comp), and the model-wise prior with wrong/opposite expert knowledge (ADX-tree-opp*) also resulted in satisfactory robustness, but then the proportion of model fits with stable inference started to decrease. The robustness of the additive and dominance model and the nonadditive model with default component-

Table 2: Method robustness measured in stability of inference^a by model and prior setting.

Setting (abbreviation)	Stability
Phenotype selection	1.00
Add. tree expert (A-tree*)	1.00
Add. tree default (A-tree)	0.99
Add. maximum likelihood (A-ML)	0.99
Nonadd. tree expert (ADX-tree*)	0.98
Add. + dom. tree expert (AD-tree*)	0.97
Add. comp. expert (A-comp*)	0.94
Add. + dom. maximum likelihood (AD-ML)	0.88
Add. comp. default (A-comp)	0.86
Nonadd. tree expert opposite (ADX-tree-opp*)	0.86
Nonadd. comp. expert (ADX-comp*)	0.80
Nonadd. maximum likelihood (ADX-ML)	0.79
Add. + dom. comp. expert (AD-comp*)	0.69
Add. + dom. tree default (AD-tree)	0.51
Nonadd. tree default (ADX-tree)	0.23
Add. + dom. comp. default (AD-comp)	0.13
Nonadd. comp. default (ADX-comp)	0.04

^aas a proportion of analyses with less than 1% divergences for the Bayesian approach and as a proportion of analyses with convergence for the maximum likelihood approach.

wise priors (AD-comp and ADX-comp) was improved by using the model-wise priors (AD-tree and ADX-tree), and even further by expert knowledge (AD-comp* and ADX-comp*), but neither they nor the nonadditive model fitted with maximum likelihood (ADX-ML) had more than 80% stable model fits.

The reason for deteriorated robustness of some model and prior settings is that genetic (especially the nonadditive) and environmental effects can be partially confounded, which limits the exploration of the posterior when using the Bayesian approach or limits convergence of mode-seeking algorithms when using the maximum likelihood approach. We show the partial confounding with images of the covariance matrices for additive, dominance, epistasis and environmental sources of variation for one dataset in Figure S8 in File S1 in the Supplemental materials, and scatterplots of the pairwise elements on and off the diagonal of the same matrices in Figure S9. Figure S10 shows joint posterior samples for the epistasis and environmental variance for model ADX with model-wise priors with and without expert knowledge (ADX-tree* and ADX-tree) for one dataset. Without a robust method (this includes both the model and inference approach),

the posterior distribution becomes difficult to explore, and this is also supported by the divergence diagnostics (Table 2). The posterior of the ADX-tree setting is bimodal and the sampler has not been able to sample with convergence due to confounding.

We do not present results from the settings with 80% or less stable model fits (see Table 2) in the following. Note that Table 2 includes all model abbreviations used. For each setting, the breeding programs that did not result in stable inference were removed from the results.

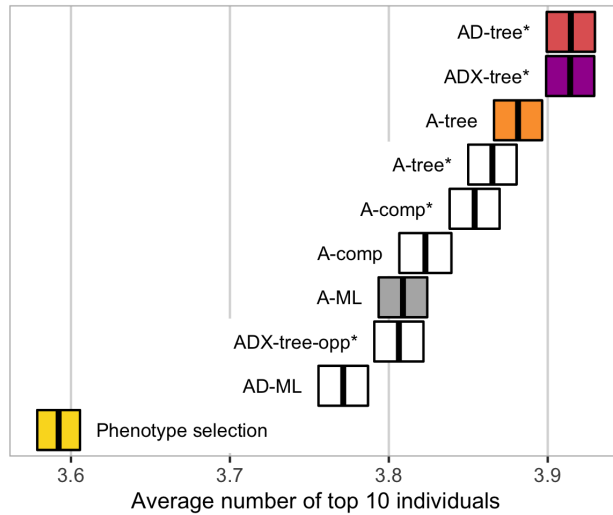
3.1.2 Selecting best individuals:

Figure 5 shows the accuracy of selecting individuals with the highest genetic value (variety selection, Figure 5a) and with the highest additive value (parent selection, Figure 5b). The model-wise priors exploiting expert knowledge improved the selection of the genetically best individuals significantly, and the model choice was important for different breeding aims. The settings with the additive and dominance model and the nonadditive model with model-wise expert knowledge (AD-tree* and ADX-tree*) performed significantly better in selection of new varieties than the others, which followed in order from A-tree, A-tree*, A-comp*, A-comp, A-ML, ADX-tree-opp* and AD-ML (see Table 2 for abbreviations). The differences between the settings were small, but they all performed significantly better than sole phenotype selection, which is sensitive to environmental noise. For the selection of new parents the simpler additive model performed the best, and the model-wise priors improved its performance (A-tree, A-tree* and A-comp*). Wrong expert knowledge harmed the parent selection (ADX-tree-opp*), but it still outperformed sole phenotype selection.

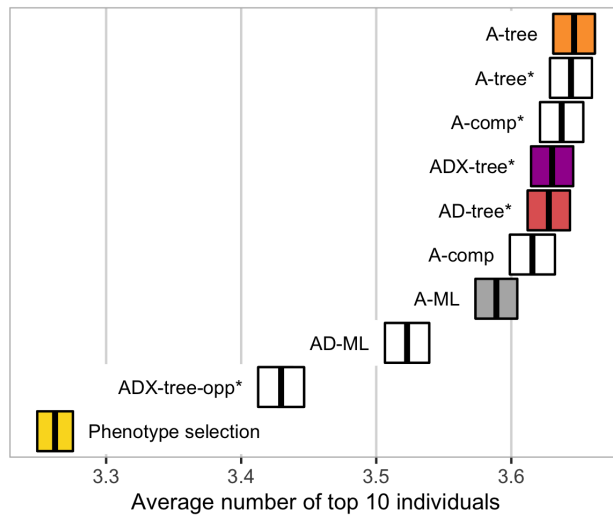
3.1.3 Estimation:

We summarize the remaining results here, and include a detailed description of the results for the additive model with model-wise default prior (A-tree) and the maximum likelihood approach (A-ML), the additive and dominance model and the nonadditive model with model-wise expert knowledge prior (AD-tree* and ADX-tree*), in addition to phenotype selection, in Note S2, and provide the complete results for all settings in Figures S11-S16 in File S1 in the Supplemental materials.

While using the model-wise priors and expert knowledge significantly improved the selection of the genetically best individuals compared to the maximum-likelihood approach, it did not significantly improve the accuracy of estimating



(a) Genetic value (variety selection).



(b) Additive value (parent selection).

Figure 5: Accuracy of selecting individuals with the highest (a) genetic value (for variety selection) and (b) additive value (for parent selection) by model and prior setting - measured with the number of the top 10 true best individuals among the top 10 selected individuals (average \pm two standard errors over replicates).

different genetic values across all individuals (Figures S11 and S12). There was a tendency for the Bayesian models to perform better than the models fitted with the maximum likelihood approach, but the variation between replicates was larger than the variation between the settings. All models outperformed phenotype selection, where we treat the phenotype as a point estimate of the genetic value.

Figure S13 shows that the variance component estimates varied considerably around the true values for all models and prior settings. The estimates from the Bayesian inference showed slightly larger biases and smaller variances than maximum likelihood estimates. Estimates for epistasis variance were considerably more underestimated than for the dominance variance.

The inference stability did not increase with increasing number of observations for any of the models fitted with the maximum likelihood approach. The Bayesian models with model-wise expert knowledge priors had the same high inference stability as in Table 2. The variation between replicates decreased for the variance estimates (Figure S14) and the correlation and continuous rank probability score (CRPS) of the model effects improved for all models for increasing number of observations (Figures S15 and S16). 700 observations was not enough for the maximum likelihood approach to obtain a bias in dominance and epistasis variance estimates as low as the Bayesian approach (Figure S14), indicating that the need for good prior distributions is still there, but decreases with increasing number of observations.

3.2 Real case study

The Bayesian approach with model-wise expert knowledge priors performed at least as good as or better than the maximum likelihood (ML) approach. Figure 6 shows the predictive ability of phenotypes measured with correlation and CRPS from three trials in Seligenstadt (Sel13 and Sel12) and Hadmersleben (Had12) over the five 5-fold cross-validations. These trials had phenotype observations for 1,739 (Sel13), 834 (Sel12) and 1,738 (Had12) parents and hybrids, and represent three different groups of trials: Sel13 represents the trials Ade13, Boh13, Hhof12, Hoh12, Hoh13 and Sel13 where few observations are missing and the Bayesian and ML approaches perform equally good. Sel12 represents the trials Boh12 and Sel12 where we have many missing observations and the ML approach is diverging. Had12 represents the trials Had12, Had13 and Hhof13 where few observations are missing but the ML approach leads to overfitting of the nonadditive effects. Inside each group the results give similar conclusions, and we show results for only one trial in each group here. We include correlation and CRPS for all 11 trials in Figures S17 and S18 in File S1 in the Supplemental materials. The

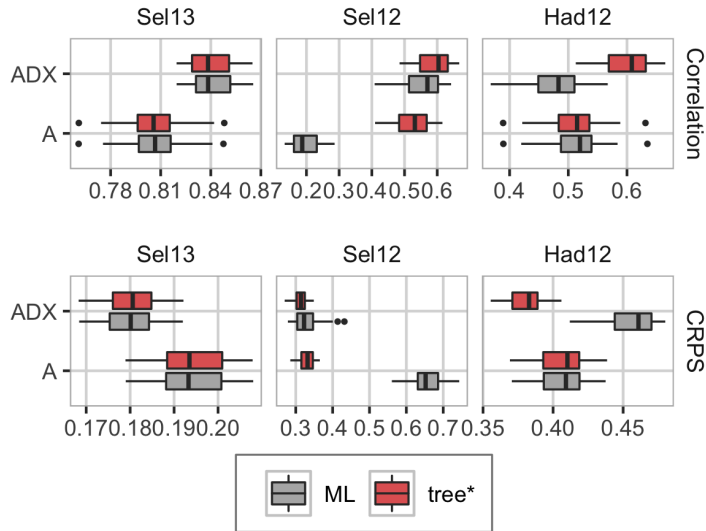


Figure 6: Phenotype prediction ability measured with correlation (top; high value desired), and continuous rank probability score (bottom; low value desired) from three of the trials in the real case study (boxplots show variation over the cross-validations and folds). Left: Sel13 (1,739 observations), middle: Sel12 (834 observations), right: Had12 (1,738 observations).

maximum likelihood approach was as good as the Bayesian approach in the Sel13 trial where all phenotypes were observed for the parents and hybrids, but in the Sel12 trial, which consists of only 834 out of 1,739 observed phenotypes, the maximum likelihood approach had worse predictive ability for the additive model (A), and slightly worse for the nonadditive model (ADX). In the Had12 trial with practically no unobserved phenotypes, the maximum likelihood approach is outperformed by the Bayesian approach for the nonadditive model due to overfitting through overestimation of the epistasis variance (see Figure 7). The results from the additive and dominance (AD) model did not differ from the results from the additive and nonadditive model, and we do not discuss them here, but include the results from AD-tree* and AD-ML in File S1 (Figures S17-S19).

We explored reasons for the bad performance of A-ML in the Sel12 trial (representing trials with many missing observations). The maximum likelihood optimizer returned a converge error message for two of the total 25 folds (we removed these model fits from all the results). However, the severe overestimation

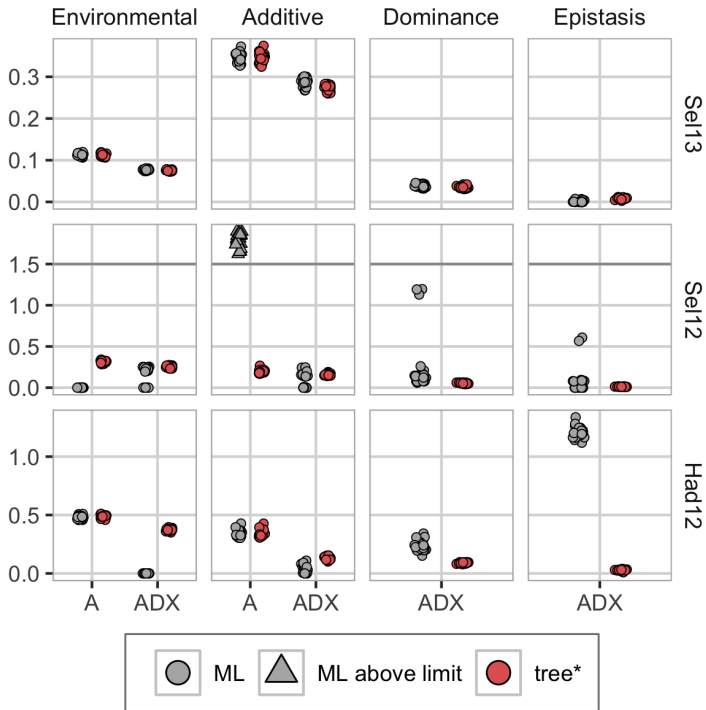


Figure 7: Posterior median variances from the real case study for three of the trials for the five 5-fold cross-validations. Top: Sel13 (1,739 observations), middle: Sel12 (834 observations), bottom: Had12 (1,738 observations). For Sel12, the A-ML is overestimating the additive variance so badly (values over 400) that we have truncated the y-axes at 1.5 to highlight the other results.

of the additive variance shown in Figure 7 indicates that the optimizer did not find the global maximum, but rather a local one. A closer investigation of the variance estimates showed that the optimizer got "stuck" at the lower boundary values (-20 for the environmental and -50 for the other variances on a logarithmic scale). We gave 0 as initial value for the intercept and log-variances for both the Bayesian and maximum likelihood approach, however, the latter did not converge.

In Figure 7 we see that for Sel13 the approaches are in agreement on the variance estimates. With a dataset with many unobserved phenotypes (represented by Sel12), the additive model fitted with the maximum likelihood approach (A-ML) estimated the environmental log-variance at -20 , and in compensation severely overestimated the additive variance. The nonadditive model fitted with maximum likelihood (ADX-ML) had the same underestimation of the environmental variance for some folds, but compensated with nonadditive effects. This indicates overfitting and means that predictions from such are based solely on genetic values, and no environmental effects, which gives misleading predictions. ADX-ML was also underestimating the environmental variance for the data from Had12, Had13 and Hhof13, and compensated this variance with the dominance and epistasis effects. We reran the maximum likelihood optimizer with initial values set to posterior medians from the corresponding Bayesian models. In this case, the maximum likelihood approach was not outperformed by the Bayesian approach (see Figures S17 and S18 in File S1). The variance estimates for all environments can be seen in Figure S19.

In Figures S17 and S18 we see that the trend is the same across the trials; for datasets where we have observed most of the phenotypes for the parents and hybrids, the maximum likelihood and Bayesian approaches are in general performing equally, and we gain predictive accuracy by including nonadditive effects, but as soon as there are many unobserved phenotypes, such as for Boh12 and Sel12 (see Table S1 for information about all trials), the maximum likelihood approach is deteriorating. For the Had12, Had13 and Hhof13 trials, which has few unobserved phenotypes but still has poor predictive abilities for the nonadditive model (ADX), the maximum likelihood approach has problems with overfitting (see Figure S19). The model underestimates the environmental variance and attributes this variation to the dominance and epistasis effects.

4 Discussion

In this study we have introduced new priors for robust genomic modeling of additive and nonadditive variation based on the penalized complexity prior (Simpson et al., 2017) and hierarchical decomposition prior (Fuglstad et al., 2020) frame-

works. In the simulated case study, the new priors enabled straightforward use of expert knowledge, which in turn improved the robustness of genomic modeling and the selection of the genetically best individuals in a wheat breeding program. However, it did not improve the overall accuracy of estimating genetic values for all individuals or for variance components. In the real case study, the new priors improved the prediction ability, especially for trials with fewer observations, and they reduced overfitting. These results highlight three points for discussion: (i) expert-knowledge priors for genomic modeling and prediction, (ii) the importance of priors for breeding and (iii) limitations of our work.

4.1 Expert-knowledge priors for genomic modeling and prediction

Genomic modeling is challenging due to inherent high-dimensionality and pervasive correlations between loci and therefore requires substantial amounts of information for robust estimation. Most genomes harbour millions of segregating loci that are highly or mildly correlated. While estimating additive effects at these loci is a challenging task in itself (e.g., Visscher et al., 2017; Young, 2019), estimating dominance and epistasis effects is an even greater challenge (e.g., Miształ, 1997; Zhu et al., 2015; de los Campos et al., 2019). One challenge in estimating the interactive dominance and epistasis effects is that they are correlated with the main additive effects and all these effects are further correlated across nearby loci (Mäki-Tanila and Hill, 2014; Hill and Mäki-Tanila, 2015; Vitezica et al., 2017). Information to estimate all these locus effects and corresponding individual values has to inherently come from the data, but could also come in a limited extent from the expert knowledge. There is a wealth of expert knowledge in genetics (e.g., Houle, 1992; Falconer and Mackay, 1996; Lynch et al., 1998), however, this expert knowledge is seldom used because it is not clear how to use it in a credible and a consistent manner.

We showed how to use the expert knowledge about the magnitude of different sources of variation by leveraging two recently introduced prior frameworks (Simpson et al., 2017; Fuglstad et al., 2020). While the penalized complexity priors are parsimonious and intuitive, they require absolute prior statements when used in a component-wise approach, which are challenging to elicit for multiple effects. The hierarchical decomposition framework imposes a tree structure according to a domain model, and the intuitive penalized complexity prior can be used in the hierarchical decomposition prior framework to ensure robust modeling. This model-wise approach enables the use of *relative* prior statements, which are less challenging to elicit than the absolute prior statements, because we tend to have good knowledge of the broad sense heritability for most traits and

by the standard quantitative genetic model construction we know that additive effects capture majority of genetic variance (Hill et al., 2008; Mäki-Tanila and Hill, 2014; Hill and Mäki-Tanila, 2015; Huang and Mackay, 2016). The presented priors therefore pave a way for a fruitful elicitation dialogue between a geneticist and a statistician (Farrow, 2013). In particular, the hierarchical decomposition prior framework provides both a method for prior construction and a platform for communication among geneticists and statisticians. The model-wise expert knowledge prior must naturally be adapted to each model, as it depends on the model structure, but using the tree structures makes this adaptation intuitive and should as such help with prior elicitation (O’Hagan et al., 2006; Farrow, 2013). Further, the graphical representation allows a description of a joint prior in a visual way with minimal statistical jargon (Figure 1).

An example of using such expert knowledge was the choices of a median for the broad-sense heritability of 0.25 in the simulated and 0.75 in the real case study. However, as Figures 2 and S7 show, the priors do not differ tremendously. This shows that the prior proposed in this study is vague and not restricted by the value chosen for the median. Perhaps there is even scope for more concentrated priors, should such information be available.

The hierarchical decomposition prior framework enabled us to use expert knowledge on relative additive and nonadditive variation. If nonadditive effects are to be added to the model, expert knowledge is necessary for the inference to be stable and the results reliable, and the simulation study shows that the expert knowledge must be added in such a way that the magnitude of the variances are not restricted by the prior, i.e., the model-wise approach instead of the component-wise approach. In the simulated case study the expert knowledge improved the stability of inference of the Bayesian approach over the maximum likelihood approach and improved the selection of the genetically best individuals. This improvement was due the additional information that alleviated the strong confounding between the nonadditive (particularly epistasis) and environmental variation.

The hierarchical decomposition prior framework is also useful when expert knowledge is only available on parts of the model. For example, an expert may not have a good intuition about the level of broad-sense heritability, say for a new trait, but will likely have a good intuition on how the genetic variance *relatively* decomposes into additive, dominance and epistasis components, simply due to the model specification (Hill et al., 2008; Mäki-Tanila and Hill, 2014; Hill and Mäki-Tanila, 2015; Huang and Mackay, 2016). In those cases, we can use weakly-informative default priors on the parts of the model where expert knowledge is missing, and priors based on expert knowledge for the rest of the model. The component-wise specification of expert knowledge with the standard

(Sorensen and Gianola, 2007) or the penalized complexity (Simpson et al., 2017) priors is infeasible in this context, and does not admit a simple visualization of the implied assumptions on the decomposition of the phenotypic variance. Further, the component-wise specification of expert knowledge is particularly challenging when phenotypic variance is unknown or when collected observations are influenced by a range of effects which can inflate sample phenotypic variance. The model-wise approach with the hierarchical decomposition prior can address these situations.

There exists previous work on penalized estimation of genetic covariances (e.g., Meyer et al., 2011; Meyer, 2016, 2019) that also uses Bayesian principles and scale-free penalty functions to reduce variation of the estimates from small datasets and for large numbers of traits. Our proposed priors and expert knowledge reduced variation of estimates in the simulated case study. However, our estimates were biased, which is expected given the small sample size and that the Bayesian approach induced bias towards a lower variance (e.g, Sorensen and Gianola, 2007). It is worth noting that the maximum likelihood estimates of genetic variance also were largely underestimated, which we believe is due to the small sample size and a large number of parameters to estimate. We see in Note S2 in the Supplemental materials (File S1) that the data informs about phenotypic variance and broad-sense heritability, but only weakly about the division of the additive and nonadditive, and dominance and epistasis. Further, for some datasets we could not obtain the maximum likelihood estimates, while priors robustified the modeling by penalizing the genetic effects. The real case study also showed that using expert knowledge increases the inference robustness in datasets with a large amount of unobserved phenotypes, and reduces overfitting. We saw this improvement in both the Bayesian approach and the maximum likelihood approach where we used the results from the Bayesian models as initial values for the optimization algorithm. However, the latter approach requires specific expert knowledge on the size of the variances, which in the same way as the component-wise expert knowledge priors, is difficult to elicit from experts in the field. We note, however, that genomic models are inherently misspecified by trying to estimate the effect of causal loci through correlated marker loci (Gianola et al., 2009; de los Campos et al., 2015). Also, linkage and linkage disequilibrium are challenging the decomposition of genetic variance into its components (Gianola et al., 2013; Morota et al., 2014; Morota and Gianola, 2014). Indeed, our variance estimates were not very accurate in the simulated case study.

Future research could expand the hierarchical decomposition prior framework to other settings. For example, to multiple traits or modeling genotype-by-environment interactions, which are notoriously noisy, and we aim to find parsimonious models (e.g., Meyer, 2016, 2019; Tollhurst et al., 2019). Also, expand

to model macro- and micro-environmental effects (e.g., Selle et al., 2019) and to model multiple layers of sparse, yet high-dimensional, "omic" data from modern biological experiments using network-like models (Damianou and Lawrence, 2013).

4.2 Importance of priors for breeding

Robust genomic modeling of nonadditive variation is important for breeding programs. There is substantial literature indicating sizeable nonadditive genetic variation (e.g., Oakey et al., 2006; Muñoz et al., 2014; Bouvet et al., 2016; Varona et al., 2018; de Almeida Filho et al., 2019; Santantonio et al., 2019; Tolhurst et al., 2019), but robust modeling of this variation is often challenging. We have shown how to achieve this robust modeling with the proposed priors and expert knowledge. We evaluated this approach with a simulated wheat breeding program where we assessed the ability to select the genetically best individuals on their genetic value (variety selection) and additive value (parent selection). The results showed that the proposed priors and the expert knowledge improved variety and parent selection. We observed more improvement in the variety selection, which is expected because there is more variation in genetic values than its first-order approximation additive values. However, this additional nonadditive variation is hard to model due to a small signal from the data relative to environmental variation and confounding with the environmental variation. This confounding is expected. As pointed by one of the reviewers, we obtain the epistasis covariance matrix using the Hadamard product of the additive covariance matrix with itself, and such repeated Hadamard multiplication converges to an identity matrix, i.e., to the covariance matrix of the environmental effect. Both the simulated and real case studies showed that including nonadditive effects in the model requires some sort of penalization to avoid overfitting environmental noise as nonadditive genetic effects. The proposed priors and the expert knowledge helped us to achieve this.

Importantly, all models improved upon sole phenotypic selection in the simulated case study, which shows the overall importance of genomic modeling. While the differences between the different models and priors were small, the improved genomic modeling can contribute to the much needed improvements in plant breeding (Ray et al., 2013; Asseng et al., 2015). Also, even a small improvement in the variety selection has a huge impact on production, because varieties are used extensively (Acquaah, 2009). In the terms of model complexity, the answer to whether to use the additive model, the additive and dominance model or the nonadditive model depended on the aim of the analysis. The latter models were the best in selecting the genetically best individuals on genetic value, whereas

the additive model performed best in selecting the genetically best individuals on additive value. The reason for this is likely the small sample size and large number of parameters to estimate with the nonadditive model (Varona et al., 2018). In the real case study adding nonadditive effects to the model improved the phenotypic prediction accuracy beyond the additive model, and the expert knowledge helped us to avoid overfitting, which shows the advantage of the expert knowledge.

Of note is the observation that the proposed priors and the expert knowledge improved the selection of the genetically best individuals, but not the estimation of the different genetic values. We did not expect this difference. In principle both of these metrics are important, but for breeding the ability to select the genetically best individuals is more important (de los Campos et al., 2013). A possible explanation for the difference between the two metrics is that the top individuals deviated more from the overall distribution and the overall metrics do not capture well the tail-behaviour.

The importance of the proposed priors and the expert knowledge will likely vary with the stage and size of a breeding program, and as the simulation study with increasing amount of observations and the real case study shows, the importance of priors increases with the decreasing amount of observations. Prior importance is known to decrease as the amount of data increases (Sorensen and Gianola, 2007), but the required amount of data for accurate estimation of nonadditive effects is huge compared to the size of most breeding programs. Therefore the proposed penalized complexity and hierarchical decomposition priors could be helpful also in large breeding programs as they enforce shrinkage according to the expert knowledge unless the data indicates otherwise, reducing the risk of estimating spurious effects.

4.3 Limitations of our work

The aim of this paper was to describe the use of the expert knowledge to improve genomic modeling, which we achieved through two recently introduced prior frameworks (Simpson et al., 2017; Fuglstad et al., 2020), and demonstrated their use in a simulated and a real case study of wheat breeding. There are multiple other possible uses of the proposed priors in genomic modeling and prediction. The simulated case study is small with only 100 individuals at the advanced yield trials of a wheat breeding program, and up to 700 individuals at the preliminary yield trials. A small number of individuals and a limited genetic variation at this stage made a good case study to test the importance of priors, and we show that using our approach can be beneficial beyond the standard genomic model. We have also chosen this stage for computationally simplicity and speed because we

evaluate the robustness of estimation over many replicates. Studies with more individuals are a natural next step, but is beyond the scope of this paper due to computational reasons. Finally, we could have tested more prior options, in particular the shrinkage of the nonadditive values towards the additive values, i.e., the $PC_0(\cdot)$ versus the $PC_M(\cdot)$ prior. More research is needed in the future to see how the expert knowledge can improve genetic modeling further.

Interesting areas for future research are also in other breeding domains with the recent rise in volumes of individual genotype and phenotype data, which provide power for estimating dominance and epistasis values (e.g., Alves et al., 2020; Joshi et al., 2020). The ability to estimate the nonadditive values would be very beneficial in breeding programs that aim to exploit biotechnology (e.g., Gottardo et al., 2019). Finally, an exciting area for estimating nonadditive individual values is in the area of personalized human medicine (de los Campos et al., 2010; Mackay and Moore, 2014; Sackton and Hartl, 2016; de los Campos et al., 2018; Begum, 2019).

The proposed priors are novel and require further computational work to facilitate widespread use. The penalized complexity priors (Simpson et al., 2017) are increasingly used in the R-INLA software (Rue et al., 2009, 2017), while hierarchical decomposition priors (Fuglstad et al., 2020) have been implemented with the general purpose Bayesian software Stan (Carpenter et al., 2017; Stan Development Team, 2019). This implementation is technical and Stan is slow for genomic models, although there is active development to increase its computational performance (Margossian et al., 2020).

We are in the process of developing an R package that will offer an intuitive user interface to specify hierarchical decomposition priors. The clear graphical representation of the priors along the model defined tree encourages increased transparency within the scientific community. It facilitates communication and discussion between statisticians and non-statisticians in the process of the model design, prior specification but also model assessment. Existing expert knowledge is intuitively incorporated into PC prior distributions for the parameters where it applies to. The resulting model-wise prior can be fed directly into Stan or INLA, or can be pre-computed for use in other Bayesian software. Thus, the new priors will be straightforward to apply for statisticians and non-statisticians, robustify the analysis, and the use of INLA will speed up computations. Further work is needed to enable Bayesian treatment of large genomic models fitted to datasets with hundreds of thousands of individuals.

4.4 Conclusion

In conclusion, we have presented how to use the expert knowledge on relative magnitude of genetic variation and its additive and nonadditive components in the context of a Bayesian approach with two novel prior frameworks. We believe that when modeling a phenomenon for which there exists a lot of knowledge, we should employ methods that are able to take advantage of this resource. We have demonstrated with a simulated and a real case study that such methods are important and helpful in the breeding context, and they might have potential also in other areas that use genomic modeling.

References

- Acquaah, G. (2009). *Principles of Plant Genetics and Breeding*. John Wiley & Sons.
- Alves, K., Brito, L. F., Baes, C. F., Sargolzaei, M., Robinson, J. A. B., and Schenkel, F. S. (2020). Estimation of additive and non-additive genetic effects for fertility and reproduction traits in North American Holstein cattle using genomic information. *Journal of Animal Breeding and Genetics*, 00(n/a):1–15.
- Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., Kimball, B., Ottman, M. J., Wall, G., White, J. W., et al. (2015). Rising temperatures reduce global wheat production. *Nature Climate Change*, 5(2):143–147.
- Begum, R. (2019). A decade of Genome Medicine: toward precision medicine. *Genome Medicine*, 11(13).
- Belamkar, V., Guttieri, M. J., Hussain, W., Jarquín, D., El-basyoni, I., Poland, J., Lorenz, A. J., and Baenziger, P. S. (2018). Genomic selection in preliminary yield trials in a winter wheat breeding program. *G3: Genes, Genomes, Genetics*, 8(8):2735–2747.
- Betancourt, M. (2016). Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1604.00695 [stat.ME]*.
- Bouvet, J.-M., Makouanzi, G., Cros, D., and Vigneron, P. (2016). Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: prediction accuracy implications. *Heredity*, 116(2):146–157.

- 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).
- Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215.
- de Almeida Filho, J. E., Guimarães, J. F. R., Fonseca e Silva, F., Vilela de Resende, M. D., Muñoz, P., Kirst, M., and de Resende Júnior, M. F. R. (2019). Genomic prediction of additive and non-additive effects using genetic markers and pedigrees. *G3: Genes, Genomes, Genetics*, 9(8):2739–2748.
- de los Campos, G., Gianola, D., and Allison, D. B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11(12):880–886.
- de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345.
- de los Campos, G., Sorensen, D., and Gianola, D. (2015). Genomic heritability: what is it? *PLoS Genetics*, 11(5):e1005048.
- de los Campos, G., Sorensen, D. A., and Toro, M. A. (2019). Imperfect linkage disequilibrium generates phantom epistasis (& perils of big data). *G3: Genes, Genomes, Genetics*, 9(5):1429–1436.
- de los Campos, G., Vazquez, A. I., Hsu, S., and Lello, L. (2018). Complex-trait prediction in the era of big data. *Trends in Genetics*, 34(10):746–754.
- Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Longman Group, Harlow, 4 edition.
- FAO, IFAD, UNICEF, WFP, and WHO (2019). *The State of Food Security and Nutrition in the World 2019*. FAO, Rome.
- Farrow, M. (2013). Prior elicitation. In Dubitzky, W., Wolkenhauer, O., Cho, K.-H., and Yokota, H., editors, *Encyclopedia of Systems Biology*, pages 1743–1743. Springer New York, New York, NY.
- Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., Hearne, S. J., Gonen, S., and Hickey, J. M. (2016). Alphasim: Software for breeding program simulation. *The Plant Genome*, 9(3):1–14.

- Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433.
- Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*. Advance publication.
- Gaynor, C. (2019). Alphasimr: Breeding program simulations. <https://CRAN.R-project.org/package=AlphaSimR>. R package version 0.10.0.
- Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., Mackay, I. J., and Hickey, J. M. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, 57(5):2372–2386.
- General Assembly of the United Nations (2015). Resolution adopted by the General Assembly on 25 September 2015. A/RES/70/1.
- Gianola, D. and de los Campos, G. (2008). Inferring genetic values for quantitative traits non-parametrically. *Genetics Research*, 90(6):525–540.
- Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, 183(1):347–363.
- Gianola, D. and Fernando, R. L. (1986). Bayesian methods in animal breeding theory. *Journal of Animal Science*, 63(1):217–244.
- Gianola, D., Hospital, F., and Verrier, E. (2013). Contribution of an additive locus to genetic variance when inheritance is multi-factorial with implications on interpretation of GWAS. *Theoretical and Applied Genetics*, 126(6):1457–1472.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- Gottardo, P., Gorjanc, G., Battagin, M., Gaynor, R. C., Jenko, J., Ros-Freixedes, R., Bruce A. Whitelaw, C., Mileham, A. J., Herring, W. O., and Hickey, J. M. (2019). A strategy to exploit surrogate sire technology in livestock breeding programs. *G3: Genes, Genomes, Genetics*, 9(1):203–215.
- Gowda, M., Zhao, Y., Würschum, T., Longin, C. F., Miedaner, T., Ebmeyer, E., Schachschneider, R., Kazman, E., Schacht, J., Martinant, J., et al. (2014). Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. *Heredity*, 112(5):552–561.

- 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.
- Henderson, C. R. (1985). Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. *Journal of Animal Science*, 60(1):111–117.
- Hill, W. and Mäki-Tanila, A. (2015). Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. *Journal of Animal Breeding and Genetics*, 132(2):176–186.
- Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4(2):e1000008.
- Horn, R. A. (1990). The Hadamard Product. In *Proceedings of Symposia in Applied Mathematics*, volume 40, pages 87–169.
- Houle, D. (1992). Comparing evolvability and variability of quantitative traits. *Genetics*, 130(1):195–204.
- Huang, W. and Mackay, T. F. C. (2016). The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS Genetics*, 12(11):e1006421.
- Johnson, S. G. (2020). The NLOpt nonlinear-optimization package. Accessed 2020-03-01.
- Joshi, R., Meuwissen, T. H., Woolliams, J. A., and GjØen, H. M. (2020). Genomic dissection of maternal, additive and non-additive genetic effects for growth and carcass traits in Nile tilapia. *Genetics Selection Evolution*, 52(1).
- Legarra, A. (2016). Comparing estimates of genetic variance across different relationship models. *Theoretical Population Biology*, 107:26–30.
- Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528.
- Lynch, M., Walsh, B., et al. (1998). *Genetics and Analysis of Quantitative Traits*, volume 1. Sinauer Sunderland, MA.
- Mackay, I., Horwell, A., Garner, J., White, J., McKee, J., and Philpott, H. (2011). Reanalyses of the historical series of UK variety trials to quantify the contributions of genetic and environmental factors to trends and variability in yield over time. *Theoretical and Applied Genetics*, 122(1):225–238.

- Mackay, T. F. and Moore, J. H. (2014). Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*, 6(6):42.
- Mäki-Tanila, A. and Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355–367.
- Margossian, C. C., Vehtari, A., Simpson, D., and Agrawal, R. (2020). Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation. *arXiv preprint arXiv:2004.12550 [stat.CO]*.
- Martini, J. W., Gao, N., Cardoso, D. F., Wimmer, V., Erbe, M., Cantet, R. J., and Simianer, H. (2017). Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinformatics*, 18(1):3.
- Martini, J. W., Toledo, F. H., and Crossa, J. (2020). On the approximation of interaction effect models by Hadamard powers of the additive genomic relationship. *Theoretical Population Biology*, 132:16–23.
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.
- Meyer, K. (2016). Simple penalties on maximum-likelihood estimates of genetic parameters to reduce sampling variation. *Genetics*, 203(4):1885–1900.
- Meyer, K. (2019). "Bending" and beyond: Better estimates of quantitative genetic parameters? *Journal of Animal Breeding and Genetics*, 136(4):243–251.
- Meyer, K., Kirkpatrick, M., Gianola, D., et al. (2011). Penalized maximum likelihood estimates of genetic covariance matrices with shrinkage towards phenotypic dispersion. In *Proc Ass Advan Anim Breed Genet*, volume 19, pages 87–90.
- Misztal, I. (1997). Estimation of variance components with large-scale dominance models. *Journal of Dairy Science*, 80(5):965–974.
- Morita, S., Thall, P. F., and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602.
- Morota, G., Boddhireddy, P., Vukasinovic, N., Gianola, D., and DeNise, S. (2014). Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. *Frontiers in Genetics*, 5:56.

- Morota, G. and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in Genetics*, 5:363.
- Muñoz, P. R., Resende, M. F. R., Gezan, S. A., Resende, M. D. V., de los Campos, G., Kirst, M., Huber, D., and Peter, G. F. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics*, 198(4):1759–1768.
- Nocedal, J. (1980). Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782.
- Oakey, H., Verbyla, A., Pitchford, W., Cullis, B., and Kuchel, H. (2006). Joint modeling of additive and non-additive genetic line effects in single field trials. *Theoretical and Applied Genetics*, 113:809–819.
- O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts’ Probabilities*. John Wiley & Sons.
- Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PLoS ONE*, 8(6):e66428.
- Reif, J. C., Maurer, H. P., Korzun, V., Ebmeyer, E., Miedaner, T., and Würschum, T. (2011). Mapping QTLs with main and epistatic effects underlying grain yield and heading time in soft winter wheat. *Theoretical and Applied Genetics*, 123(2):283.
- 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.
- Rife, T. W., Graybosch, R. A., and Poland, J. A. (2019). A field-based analysis of genetic improvement for grain yield in winter wheat cultivars developed in the US Central Plains from 1992 to 2014. *Crop Science*, 59(3):905–910.
- 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.
- 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.
- Sackton, T. B. and Hartl, D. L. (2016). Genotypic context and epistasis in individuals and populations. *Cell*, 166(2):279–287.

- Santantonio, N., Jannink, J.-L., and Sorrells, M. (2019). Prediction of subgenome additive and interaction effects in allohexaploid wheat. *G3: Genes, Genomes, Genetics*, 9(3):685–698.
- Selle, M. L., Steinsland, I., Hickey, J. M., and Gorjanc, G. (2019). Flexible modelling of spatial variation in agricultural field trials with the R package INLA. *Theoretical and Applied Genetics*, 132(12):3277–3293.
- Shewry, P. R. and Hey, S. J. (2015). The contribution of wheat to human diet and health. *Food and Energy Security*, 4(3):178–202.
- 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.
- Sørbye, S. H. and Rue, H. (2018). Fractional Gaussian noise: prior specification and model comparison. *Environmetrics*, 29(5-6):e2457.
- Sorensen, D. and Gianola, D. (2007). *Likelihood, Bayesian, and MCMC methods in Quantitative Genetics*. Springer Science & Business Media.
- Stan Development Team (2019). RStan: the R interface to Stan. <http://mc-stan.org/>. R package version 2.19.2.
- Sweeney, D. W., Sun, J., Taagen, E., and Sorrells, M. E. (2019). Genomic selection in wheat. In Miedaner, T. and Korzun, V., editors, *Applications of Genetic and Genomic Research in Cereals*, Woodhead Publishing Series in Food Science, Technology and Nutrition, pages 273–302. Woodhead Publishing.
- Tolhurst, D. J., Mathews, K. L., Smith, A. B., and Cullis, B. R. (2019). Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. *Journal of Animal Breeding and Genetics*, 136(4):279–300.
- Varona, L., Legarra, A., Toro, M. A., and Vitezica, Z. G. (2018). Non-additive effects in genomic selection. *Frontiers in Genetics*, 9:78.
- Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.
- Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics*, 206(3):1297–1307.

Wittenburg, D., Melzer, N., and Reinsch, N. (2011). Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. *BMC Genetics*, 12(1):74.

Young, A. I. (2019). Solving the missing heritability problem. *PLoS Genetics*, 15(6):e1008222.

Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H. P., Würschum, T., Mock, H.-P., Matros, A., Ebmeyer, E., Schachschneider, R., et al. (2015). Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proceedings of the National Academy of Sciences*, 112(51):15624–15629.

Zhu, Z., Bakshi, A., Vinkhuyzen, A. A., Hemani, G., Lee, S. H., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., Esko, T., Milani, L., et al. (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics*, 96(3):377–385.

Supplementary materials:

Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge

Ingeborg Gullikstad Hem¹, Maria Lie Selle¹, Gregor Gorjanc², Geir-Arne Fuglstad¹, and Andrea Riebler²

¹Department of Mathematical Sciences, NTNU, Norway

²The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK

S1 Note S1: Detailed method description

In this note we describe the two approaches A-comp* and A-tree* in detail. We focus on the additive model (Model A) since this allows us to maximize readability and is sufficient to illustrate the main ideas, but the final section provides a description on how to extend to the additive and dominance model (Model AD) with the two approaches AD-comp* and AD-tree*. We also provide examples of the resulting prior and posterior distributions for two specific datasets with 100 and 500 observations. The aim is that this note contains all details necessary to reproduce the results.

S1.1 Model description

The additive genetic model is given by

$$y_i = \mu + a_i + e_i, \quad i = 1, 2, \dots, n,$$

where n is the number of individuals, μ is the intercept, $\mathbf{a} = (a_1, a_2, \dots, a_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2 \mathbf{A})$ are the additive values, and $\mathbf{e} = (e_1, e_2, \dots, e_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ is the

environmental noise. The covariance matrix \mathbf{A} is singular with rank less than n due to the construction from the SNP matrix. If we collect the phenotypes in a vector $\mathbf{y} = (y_1, y_2, \dots, y_n)$, the model can be formulated as the Gaussian likelihood

$$\begin{aligned} \pi(\mathbf{y}|\mu, \sigma_a^2, \sigma_e^2) &= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{|\Sigma(\sigma_a^2, \sigma_e^2)|^{n/2}} \\ &\quad \times \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{1}\mu)^T \Sigma(\sigma_a^2, \sigma_e^2)^{-1}(\mathbf{y} - \mathbf{1}\mu)\right), \quad \mathbf{y} \in \mathbb{R}^n, \end{aligned} \quad (\text{S1.1})$$

where $\Sigma(\sigma_a^2, \sigma_e^2) = \sigma_a^2 \mathbf{A} + \sigma_e^2 \mathbf{I}_n$, and $\mathbf{1} = (1, 1, \dots, 1)$ is the n -dimensional vector of ones.

In Bayesian statistics, the likelihood must be combined with a prior distribution for the parameters, $\pi(\mu, \sigma_a^2, \sigma_e^2)$. This prior should encapsulate our prior beliefs about the parameters based on expert knowledge, and stabilize inference in low-data settings. Bayes' theorem gives the posterior distribution for the parameters as

$$\pi(\mu, \sigma_a^2, \sigma_e^2|\mathbf{y}) = \frac{\pi(\mu, \sigma_a^2, \sigma_e^2)\pi(\mathbf{y}|\mu, \sigma_a^2, \sigma_e^2)}{\pi(\mathbf{y})} \propto \pi(\mu, \sigma_a^2, \sigma_e^2)\pi(\mathbf{y}|\mu, \sigma_a^2, \sigma_e^2), \quad (\text{S1.2})$$

where $\pi(\mathbf{y})$ is the marginal distribution of the phenotypes, and the proportionality is with respect to everything that does not vary as functions of the parameters. The constant $\pi(\mathbf{y})$ in Equation (S1.2) is not needed for Markov chain Monte Carlo methods.

S1.2 Component-wise priors

The common approach for selecting priors is to select independent component-wise prior distributions for σ_a^2 , σ_e^2 and μ so that $\pi(\sigma_a^2, \sigma_e^2, \mu) = \pi(\sigma_a^2)\pi(\sigma_e^2)\pi(\mu)$. A recent development are the penalized complexity (PC) priors that are derived in a principled way (Simpson et al., 2017). They are constructed based on a *base model* and "distance" to a more complex model that extends the base model. In the case of a model (S1.1) the base model does not have a genetic effect or equivalently genetic variance is zero. The proposed prior penalizes increased model complexity through Kullback-Leibler divergence between the base and complex model, but the details are not essential to our presentation, and we encourage interested readers to see Simpson et al. (2017) for details. The PC prior for the variance of a Gaussian distribution is an exponential distribution on the standard deviation,

which leads to the following priors transformed to variances for the model (S1.1):

$$\begin{aligned}\pi(\sigma_a^2) &= \frac{\lambda_a}{2\sqrt{\sigma_a^2}} \exp\left(-\lambda_a \sqrt{\sigma_a^2}\right), \quad \sigma_a^2 > 0, \\ \pi(\sigma_e^2) &= \frac{\lambda_e}{2\sqrt{\sigma_e^2}} \exp\left(-\lambda_e \sqrt{\sigma_e^2}\right), \quad \sigma_e^2 > 0.\end{aligned}$$

These priors are combined with a weakly informative Gaussian prior for the intercept, $\mu \sim \mathcal{N}_1(0, \sigma_{\text{Int}}^2)$. This is a $\text{PC}_0(\cdot)$ prior on variance.

Simpson et al. (2017) proposes to select λ_a and λ_e by eliciting a statement about a quantile for each variance from experts. In this paper we consider the specification of the two hyperparameters by specifying the upper quartiles of the priors through V_a and V_e for σ_a^2 and σ_e^2 , respectively. In this case, one must select

$$\lambda_a = -\frac{\ln(0.25)}{\sqrt{V_a}}, \quad \text{and} \quad \lambda_e = -\frac{\ln(0.25)}{\sqrt{V_e}}.$$

Applying Bayes' theorem from Equation (S1.2) results in

$$\begin{aligned}\pi(\sigma_a^2, \sigma_e^2, \mu | \mathbf{y}) &\propto \pi(\mu) \pi(\sigma_a^2) \pi(\sigma_e^2) \pi(\mathbf{y} | \mu, \sigma_a^2, \sigma_e^2) \\ &\propto \frac{\lambda_a \lambda_e}{4\sqrt{\sigma_a^2 \sigma_e^2} |\Sigma(\sigma_a^2, \sigma_e^2)|^{n/2}} \\ &\quad \times \exp\left(-\frac{1}{2}(\mathbf{y} - \mu \mathbf{1})^T \Sigma(\sigma_a^2, \sigma_e^2)^{-1} (\mathbf{y} - \mu \mathbf{1})\right. \\ &\quad \left. - \frac{\mu^2}{2\sigma_{\text{Int}}^2} - \lambda_a \sqrt{\sigma_a^2} - \lambda_e \sqrt{\sigma_e^2}\right), \quad (\text{S1.3})\end{aligned}$$

for $\mu \in \mathbb{R}$, $\sigma_a^2 > 0$ and $\sigma_e^2 > 0$.

Even though the normalizing constant is not analytically tractable in Equation (S1.3), Markov chain Monte Carlo (MCMC) methods can be devised for performing Bayesian inference through sampling. In this paper, we used Hamiltonian Monte Carlo (HMC) through Stan (Carpenter et al., 2017), a probabilistic programming language for statistical inference. Stan takes advantage of the No U-Turn Sampler (NUTS Hoffman and Gelman, 2014), which replaces random walks with a more efficient exploration strategy based on numerical solution of differential equations. NUTS also has a reduced need for tuning compared to other MCMC algorithms and, in some cases, the algorithm can be run completely without manually setting tuning parameters. Sampling from the posterior in Equation (S1.3) through Stan requires writing the expression for $\ln(\pi(\sigma_a^2, \sigma_e^2, \mu | \mathbf{y}))$ in a Stan description file. Clever parametrizations such as using the logarithms of

the variances instead of the variances themselves will improve the efficiency, but the complexity involved in implementing MCMC is greatly reduced from writing one's own MCMC algorithms. The coding language in the description file is similar as C++.

We have used Stan through the R-package `rstan` (Stan Development Team, 2019) to implement the inference in the paper. This required us only to provide the expression for the joint posterior, and we did not need to calculate the different full conditionals such as in a standard Gibbs sampling algorithm. We rephrased the additive model as a hierarchical model, where the additive values are be sampled together with the parameters,

$$\begin{aligned} \mathbf{y}|\mu, \mathbf{a}, \sigma_e^2 &\sim \mathcal{N}_n(\mathbf{1}\mu + \mathbf{a}, \sigma_e^2 \mathbf{I}_n), \\ \mathbf{a}|\sigma_a^2 &\sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2 \mathbf{A}), \\ (\mu, \sigma_a^2, \sigma_e^2) &\sim \pi(\mu, \sigma_a^2, \sigma_e^2), \end{aligned}$$

where $\mathbf{1} = (1, 1, \dots, 1)$ is the n -dimensional vector of ones. Again Bayes' theorem provides

$$\pi(\mathbf{a}, \mu, \sigma_a^2, \sigma_e^2 | \mathbf{y}) \propto \pi(\mu) \pi(\sigma_a^2) \pi(\sigma_e^2) \pi(\mathbf{a} | \sigma_a^2) \pi(\mathbf{y} | \mu, \mathbf{a}, \sigma_e^2),$$

where all terms in the product are known distributions. This joint posterior is implemented in Stan through a calculation of $\ln(\pi(\mathbf{a}, \mu, \sigma_a^2, \sigma_e^2 | \mathbf{y}))$. The details of the sampling is handled by the software. This approach is termed A-comp* in the paper, and the prior $\pi(\sigma_a)$ is denoted $\text{PC}_0(\sqrt{V_a^2}, 0.25)$ and the prior $\pi(\sigma_e)$ is denoted $\text{PC}_0(\sqrt{V_a^2}, 0.25)$.

S1.3 Model-wise prior

A shortcoming of using independent component-wise priors is that it does not provide a direct way to include expert knowledge about the relative sizes of variance parameters. In the context of the additive model, this refers to the situation where expert knowledge could be available about the phenotypic variance $\sigma_P^2 = \sigma_a^2 + \sigma_e^2$ and the broad-sense heritability $h_g^2 = p_{\frac{g}{g+e}} = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$, i.e., $\sigma_a^2 = p_{\frac{g}{g+e}} \sigma_P^2$ and $\sigma_e^2 = (1 - p_{\frac{g}{g+e}}) \sigma_P^2$. If component-wise priors are chosen for σ_a^2 and σ_e^2 , it can be extremely challenging to understand the *a priori* assumptions being imposed on σ_P^2 and $p_{\frac{g}{g+e}}$. Furthermore, we are not ensuring that we end up with reasonable families of priors for σ_P^2 and $p_{\frac{g}{g+e}}$, which have desirable properties.

A complementary framework to PC priors are the recent hierarchical decomposition (HD) priors by Fuglstad et al. (2020). Using this framework, we can

directly incorporate expert knowledge about the quantities σ_P^2 and $p_{\frac{g}{g+e}}$. Assume that the expected variability in phenotypes is not known *a priori* and that we aim for a scale-invariant prior. We can address this situation with a Jeffreys' prior

$$\pi(\sigma_P^2) \propto 1/\sigma_P^2, \quad \sigma_P^2 > 0.$$

Next, we apply the approach detailed in Fuglstad et al. (2020) by assuming that the model described by Equation (S1.1) is a flexible extension of the *base model* $\mathbf{y}|\mu, \sigma_P^2 \sim \mathcal{N}(\mathbf{1}\mu, \sigma_P^2 \mathbf{I}_n)$, where the heritability $p_{\frac{g}{g+e}} = 0$. This means that we use a $\text{PC}_0(\cdot)$ prior on $p_{\frac{g}{g+e}}$. Based on Fuglstad et al. (2020), we can then calculate the prior

$$\pi(p_{\frac{g}{g+e}}) = \lambda_h |d'(p_{\frac{g}{g+e}})| \exp\left(-\lambda_h d(p_{\frac{g}{g+e}})\right), \quad 0 < p_{\frac{g}{g+e}} < 1,$$

where

$$d(p_{\frac{g}{g+e}}) = \sqrt{p_{\frac{g}{g+e}} (\text{tr}(\mathbf{A}) - n) - \ln(\det(p_{\frac{g}{g+e}} \mathbf{A} + (1 - p_{\frac{g}{g+e}}) \mathbf{I}_n))},$$

and

$$d'(p_{\frac{g}{g+e}}) = \frac{\text{tr}(\mathbf{A}) - n - \text{tr}\left\{(p_{\frac{g}{g+e}} \mathbf{A} + (1 - p_{\frac{g}{g+e}}) \mathbf{I}_n)^{-1} (\mathbf{A} - \mathbf{I}_n)\right\}}{2\sqrt{p_{\frac{g}{g+e}} (\text{tr}(\mathbf{A}) - n) - \ln(\det(p_{\frac{g}{g+e}} \mathbf{A} + (1 - p_{\frac{g}{g+e}}) \mathbf{I}_n))}}$$

denotes the derivative of the function $d(\cdot)$. Here λ_h is a hyperparameter, $\text{tr}(\cdot)$ denotes the matrix trace, and $\det(\cdot)$ denotes the matrix determinant. The function $d(\cdot)$ is the Kullback-Leibler distance and expresses the added complexity of having a broad-sense heritability $p_{\frac{g}{g+e}} > 0$ compared to having the broad-sense heritability $p_{\frac{g}{g+e}} = 0$. We set the hyperparameter λ_h by specifying the median $R_{\frac{g}{g+e}}$ of the prior for $p_{\frac{g}{g+e}}$. This is achieved by setting

$$\lambda_h = -\frac{\ln(0.5)}{d(p_{\frac{g}{g+e}} = R_{\frac{g}{g+e}})}.$$

We choose to use independent priors for μ , σ_P^2 and $p_{\frac{g}{g+e}}$ so that $\pi(\mu, \sigma_P^2, p_{\frac{g}{g+e}}) = \pi(\mu)\pi(\sigma_P^2)\pi(p_{\frac{g}{g+e}})$.

Since the prior is formulated in terms of phenotypic variance and heritability, it is useful to reparametrize the hierarchical model as

$$\begin{aligned} \mathbf{y}|\mu, \sigma_P^2, p_{\frac{g}{g+e}} &\sim \mathcal{N}_n(\mathbf{1}\mu + \mathbf{a}, \sigma_P^2(1 - p_{\frac{g}{g+e}}) \mathbf{I}_n) \\ \mathbf{a}|\sigma_P^2, p_{\frac{g}{g+e}} &\sim \mathcal{N}_n(\mathbf{0}, \sigma_P^2 p_{\frac{g}{g+e}} \mathbf{A}) \\ \mu, \sigma_P^2, p_{\frac{g}{g+e}} &\sim \pi(\mu, \sigma_P^2, p_{\frac{g}{g+e}}). \end{aligned}$$

We calculate the posterior up to proportionality through Bayes' theorem,

$$\pi(\mathbf{a}, \mu, \sigma_{\mathbf{P}}^2, p_{\frac{g}{g+c}} | \mathbf{y}) \propto \pi(\mu) \pi(\sigma_{\mathbf{P}}^2) \pi(p_{\frac{g}{g+c}}) \pi(\mathbf{a} | \sigma_{\mathbf{P}}^2, p_{\frac{g}{g+c}}) \pi(\mathbf{y} | \mu, \mathbf{a}, \sigma_{\mathbf{P}}^2, p_{\frac{g}{g+c}}),$$

where all terms in the product are known distributions. The sampling in Stan is automatically handled based on code that calculates the value of the joint posterior up to proportionality. We precompute $\ln(\pi(p_{\frac{g}{g+c}}))$ for a range of values and approximate the function by a spline. This greatly reduces the computational burden and provides a large speed-up under MCMC sampling. This approach is termed A-tree* in the paper, and the prior $\pi(p_{\frac{g}{g+c}})$ is denoted $\text{PC}_0(R_{\frac{g}{g+c}})$.

S1.4 Data example

We simulated two datasets using the breeding program described in the main paper by reducing the 10,000 individuals at the first trial stage (headrow) to $n = 700, 600, \dots, 100$ individuals, and used one dataset of size 100 and one of size 500. The choices of hyperparameters for the priors follow the main paper: $V_a = 0.25 \cdot 1.86$ and $V_e = 0.75 \cdot 1.86$. Figures S1.1 and S1.2 clearly demonstrates that the dataset of size $n = 500$ provides more information than the data set of size $n = 100$. Figure S1.2a indicates that the inference about the phenotypic variance is not sensitive to the choice between the two priors for both $n = 100$ and $n = 500$, whereas Figure S1.2b shows that inference about heritability is influenced by the prior for $n = 100$, but not for $n = 500$. This suggests that it is important to select a plausible prior that encodes prior belief for heritability. This is an argument for a principled approach for prior construction directly targeting heritability and phenotypic variance instead of additive and environmental variances separately.

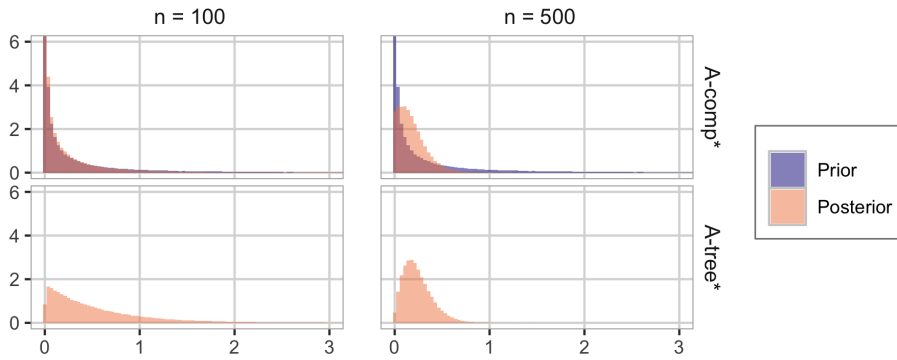
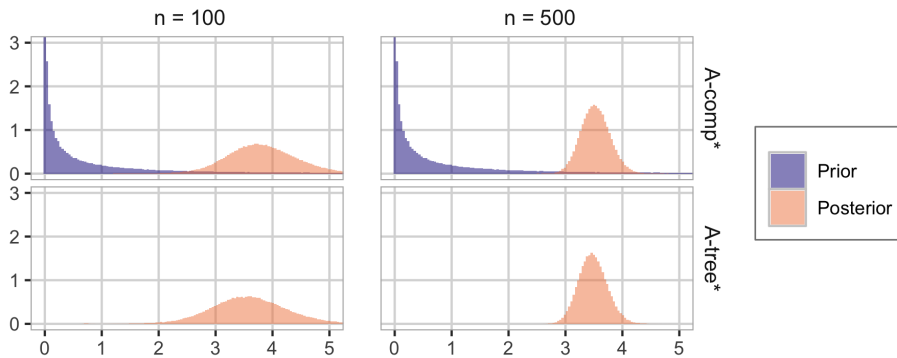
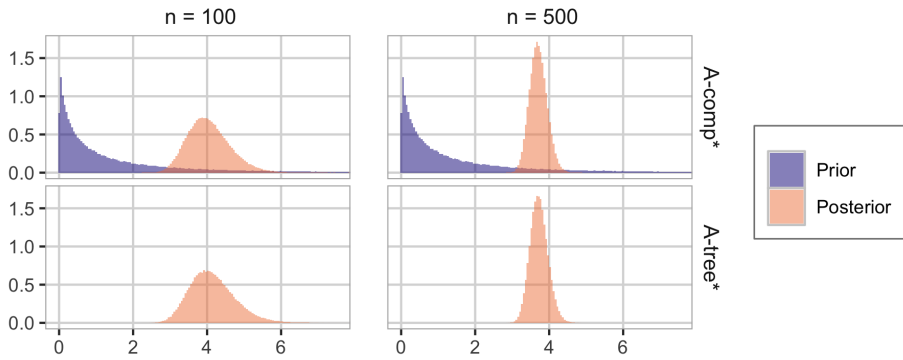
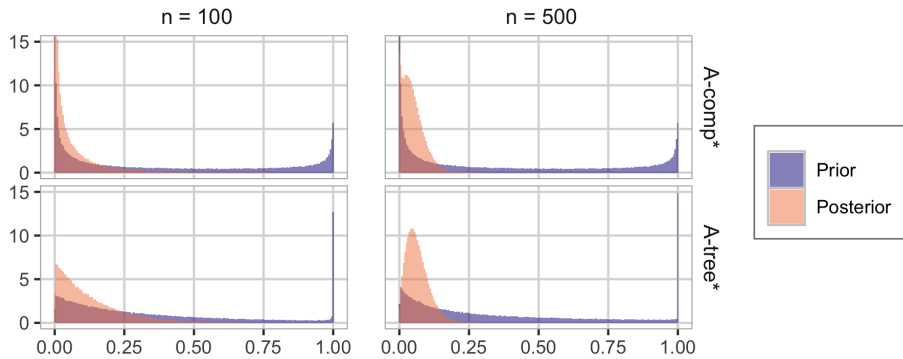
(a) Additive variance, σ_a^2 .(b) Environmental variance, σ_e^2 .

Figure S1.1: Columns indicate data sizes $n = 100$ and $n = 500$, and rows indicate priors A-comp* and A-tree*. The upper panel (a) shows priors and posteriors for additive variance σ_a^2 and the lower panel (b) shows priors and posteriors for environmental variance σ_e^2 . Priors are not plotted for A-tree* because the prior on the phenotypic variance σ_P^2 and thus also on $\sigma_a^2 = p_{\frac{g}{g+e}} \sigma_P^2$ and $\sigma_e^2 = (1 - p_{\frac{g}{g+e}}) \sigma_P^2$ are scale-invariant, and therefore improper.



(a) Phenotypic (total) variance, σ_P^2 .



(b) Heritability, $p_{\frac{g}{g+e}}$.

Figure S1.2: Columns indicate data sizes $n = 100$ and $n = 500$, and rows indicate priors A-comp* and A-tree*. The upper panel (a) shows priors and posteriors for phenotypic variance σ_P^2 and the lower panel (b) shows priors and posteriors for heritability $p_{\frac{g}{g+e}}$. Priors are not plotted for the combination A-tree* and σ_P^2 because the prior on σ_P^2 is scale-invariant, and therefore improper.

S1.5 Extension to the additive and dominance model

For Model AD, dominance values are added to the model,

$$y_i = \mu + a_i + d_i + e_i, \quad i = 1, 2, \dots, n,$$

where n is the number of individuals, μ is the intercept, $\mathbf{a} = (a_1, a_2, \dots, a_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2 \mathbf{A})$ are the additive values, $\mathbf{d} = (d_1, d_2, \dots, d_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_d^2 \mathbf{D})$ are dominance values, and $\mathbf{e} = (e_1, e_2, \dots, e_n) \sim \mathcal{N}_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ is the environmental noise. Let $\sigma_p^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2$ be the phenotypic variance, $p_{\frac{g}{g+e}} = (\sigma_a^2 + \sigma_d^2) / (\sigma_a^2 + \sigma_d^2 + \sigma_e^2)$ be the broad-sense heritability.

Following the main paper, we plan to describe the prior on the variances through a joint prior on σ_p^2 , $p_{\frac{g}{g+e}}$, and the proportion of additive to genetic variance, $p_{\frac{a}{g}} = \sigma_a^2 / \sigma_g^2$. The details are technical, and we therefore present the rationale behind each prior before we describe the technical details. We will apply independent priors to σ_p^2 , $p_{\frac{g}{g+e}}$, and $p_{\frac{a}{g}}$. These three priors will be described in reverse order.

We believe that $p_{\frac{a}{g}}$ should be around $R_{\frac{a}{g}}$, and desire a prior that favors this value and penalizes deviations from $R_{\frac{a}{g}}$. Therefore, we apply Fuglstad et al. (2020, Theorem 1) with the *base model* $p_{\frac{a}{g}} = R_{\frac{a}{g}}$, which yields

$$\pi(p_{\frac{a}{g}}) = \begin{cases} \frac{\lambda_1 |d'_1(p_{\frac{a}{g}})|}{2(1 - \exp(-\lambda_1 d_1(0)))} \exp(-\lambda_1 d_1(p_{\frac{a}{g}})), & 0 < p_{\frac{a}{g}} < R_{\frac{a}{g}}, \\ \frac{\lambda_1 |d'_1(p_{\frac{a}{g}})|}{2(1 - \exp(-\lambda_1 d_1(1)))} \exp(-\lambda_1 d_1(p_{\frac{a}{g}})), & R_{\frac{a}{g}} < p_{\frac{a}{g}} < 1. \end{cases}$$

This formulation guarantees that the median is $R_{\frac{a}{g}}$. This is a $\text{PC}_M(\cdot)$ prior. The distance is calculated as

$$d_1(p_{\frac{a}{g}}) = \sqrt{\text{tr}(\Sigma_0^{-1} \Sigma(p_{\frac{a}{g}})) - n - \ln(\det(\Sigma_0^{-1} \Sigma(p_{\frac{a}{g}}))},$$

where $\Sigma_0 = R_{\frac{a}{g}} \mathbf{A} + (1 - R_{\frac{a}{g}}) \mathbf{D}$ and $\Sigma(p_{\frac{a}{g}}) = p_{\frac{a}{g}} \mathbf{A} + (1 - p_{\frac{a}{g}}) \mathbf{D}$. In practice, we have found

$$\pi(p_{\frac{a}{g}}) = \frac{\lambda_1 |d'_1(p_{\frac{a}{g}})|}{2} \exp(-\lambda_1 d_1(p_{\frac{a}{g}})), \quad 0 < p_{\frac{a}{g}} < 1,$$

to be a reasonable approximation for the datasets in this paper as $d_1(0)$ and $d_1(1) \gg 1/\lambda_1$. The hyperparameter λ_1 controls the spread of the prior around the median $R_{\frac{a}{g}}$ and is selected by numerical optimization so that *a priori*

$$\text{P}(\text{logit}(R_{\frac{a}{g}}) - 1 < \text{logit}(p_{\frac{a}{g}}) < \text{logit}(R_{\frac{a}{g}}) + 1) = 0.75,$$

where $\text{logit}(p) = \ln(p/(1-p))$.

In the next step we construct a prior for $p_{\frac{g}{g+e}}$ with the *base model* $p_{\frac{g}{g+e}} = 0$. In this construction we assume the total genetic effect $\mathbf{g}|\sigma_P^2, p_{\frac{g}{g+e}} = (\mathbf{a} + \mathbf{d})|\sigma_P^2, p_{\frac{g}{g+e}} \sim \mathcal{N}_n(\mathbf{0}, \sigma_P^2 p_{\frac{g}{g+e}} (R_{\frac{a}{g}} \mathbf{A} + (1-R_{\frac{a}{g}}) \mathbf{D}))$. This means that we fix $p_{\frac{a}{g}} = R_{\frac{a}{g}}$ under the construction of the prior for $p_{\frac{g}{g+e}}$. Following Fuglstad et al. (2020, Theorem 1), a prior is constructed based on the distance measure

$$d_2(p_{\frac{g}{g+e}}) = \sqrt{\text{tr} \left(p_{\frac{g}{g+e}} \Sigma_2 + (1-p_{\frac{g}{g+e}}) \mathbf{I}_n \right) - n - \ln \left(\det(p_{\frac{g}{g+e}} \Sigma_2 + (1-p_{\frac{g}{g+e}}) \mathbf{I}_n) \right)},$$

where $\Sigma_2 = R_{\frac{a}{g}} \mathbf{A} + (1-R_{\frac{a}{g}}) \mathbf{D}$. The resulting prior is

$$\pi(p_{\frac{g}{g+e}}) = \lambda_2 |d'_2(p_{\frac{g}{g+e}})| \exp(-\lambda_2 d_2(p_{\frac{g}{g+e}})), \quad 0 < p_{\frac{g}{g+e}} < 1.$$

We set the hyperparameter λ_2 by specifying the median $R_{\frac{g}{g+e}}$ of the prior for $p_{\frac{g}{g+e}}$. This is achieved by setting

$$\lambda_2 = -\frac{\ln(0.5)}{d_2(p_{\frac{g}{g+e}} = R_{\frac{g}{g+e}})}.$$

This is a $\text{PC}_0(\cdot)$ prior.

We want a scale-invariant prior for the phenotypic variance, and choose a Jeffreys' prior:

$$\pi(\sigma_P^2) \propto \frac{1}{\sigma_P^2}, \quad \sigma_P^2 > 0.$$

These three priors are combined with the previous prior on μ to form

$$\pi(\mu, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}) = \pi(\mu) \pi(\sigma_P^2) \pi(p_{\frac{g}{g+e}}) \pi(p_{\frac{a}{g}}).$$

To allow sampling of both additive and dominance effects, we describe the model as a hierarchical model

$$\begin{aligned} \mathbf{y}|\mu, \mathbf{a}, \mathbf{d}, \sigma_P^2, p_{\frac{g}{g+e}} &\sim \mathcal{N}_n(\mathbf{1}\mu + \mathbf{a} + \mathbf{d}, (1-p_{\frac{g}{g+e}}) \sigma_P^2 \mathbf{I}_n), \\ \mathbf{a}|\sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}} &\sim \mathcal{N}_n(\mathbf{0}, p_{\frac{a}{g}} p_{\frac{g}{g+e}} \sigma_P^2 \mathbf{A}), \\ \mathbf{d}|\sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}} &\sim \mathcal{N}_n(\mathbf{0}, (1-p_{\frac{a}{g}}) p_{\frac{g}{g+e}} \sigma_P^2 \mathbf{D}), \\ \mu, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}} &\sim \pi(\mu, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}). \end{aligned}$$

Bayes' theorem results in

$$\begin{aligned} \pi(\mu, \mathbf{a}, \mathbf{d}, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}} | \mathbf{y}) &\propto \pi(\mu) \pi(\sigma_P^2) \pi(p_{\frac{g}{g+e}}) \pi(p_{\frac{a}{g}}) \\ &\quad \times \pi(\mathbf{a} | \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}) \pi(\mathbf{d} | \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}}) \\ &\quad \times \pi(\mathbf{y} | \mu, \mathbf{a}, \mathbf{d}, \sigma_P^2, p_{\frac{g}{g+e}}). \end{aligned}$$

The inference is implemented through `rstan`, where we write code to calculate $\ln(\pi(\mathbf{a}, \mathbf{d}, \sigma_P^2, p_{\frac{g}{g+e}}, p_{\frac{a}{g}} | \mathbf{y}))$, and parameterization through the logarithm of the variances and the logit transformation of the proportions are needed. This is the approach termed AD-tree* in the paper, and the prior $\pi(p_{\frac{g}{g+e}})$ is denoted $PC_0(R_{\frac{g}{g+e}})$ and the prior $\pi(p_{\frac{a}{g}})$ is denoted $PC_M(R_{\frac{a}{g}})$. Plots of priors and posteriors offer little added value over those shown for Model A and are therefore omitted.

The approach given in this section extends further to the model also including epistasis (Model ADX). One additional step is required, but the same approach is taken for all steps as described above for Model AD.

Full details of the implementation in Stan for the approaches used in the paper are found in Files S3 and S4. File S3 contains the full Stan-code used for model fitting in the simulated case study, with necessary R functions and scripts for constructing the prior distributions and running inference, and File S4 contains the full Stan-code used in the real case study, with necessary R functions and scripts.

References

- 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).
- Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*. Advance publication.
- Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.
- 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.
- Stan Development Team (2019). RStan: the R interface to Stan. <http://mc-stan.org/>. R package version 2.19.2.

S2 Note S2: Results

Here we give a detailed description of the results for the the additive model with model-wise default prior (A-tree) and the maximum likelihood approach (A-ML), the additive and dominance model and the nonadditive model with model-wise expert knowledge prior (AD-tree* and ADX-tree*), in addition to phenotype selection. We show the results of the remaining settings in the Figures S11-S16.

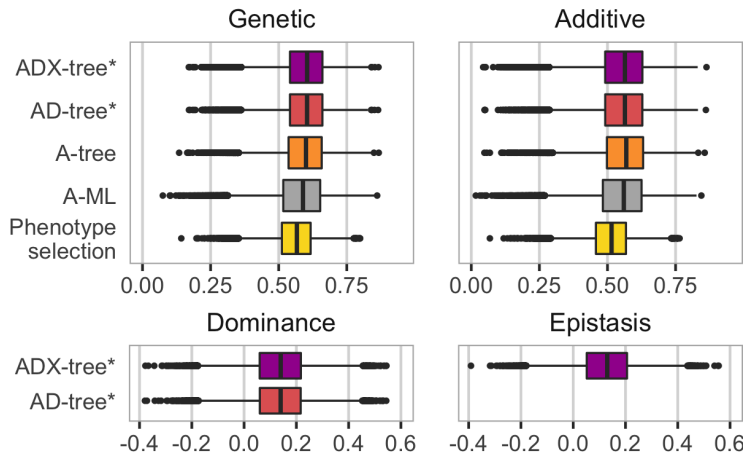


Figure S2.1: Accuracy of estimating the different genetic values for all individuals by model and prior setting - correlation (high value is desired, boxplots show variation over replicates). Genetic (upper left) means the estimated additive values for Model A, and the sum of estimated additive, dominance and epistasis values for Model ADX.

S2.1 Estimating genetic values

While using the model-wise priors and expert knowledge significantly improved the selection of the genetically best individuals compared to the maximum-likelihood approach (see the main paper), it did not significantly improve the accuracy of estimating different genetic values across all individuals. We show this in Figure S2.1 with the correlation between the true (simulated) values and corresponding posterior means and in Figure S2.2 with the continuous rank probability score (CRPS) between the true values and corresponding posterior distributions. We show this for the genetic, additive, dominance and epistasis values

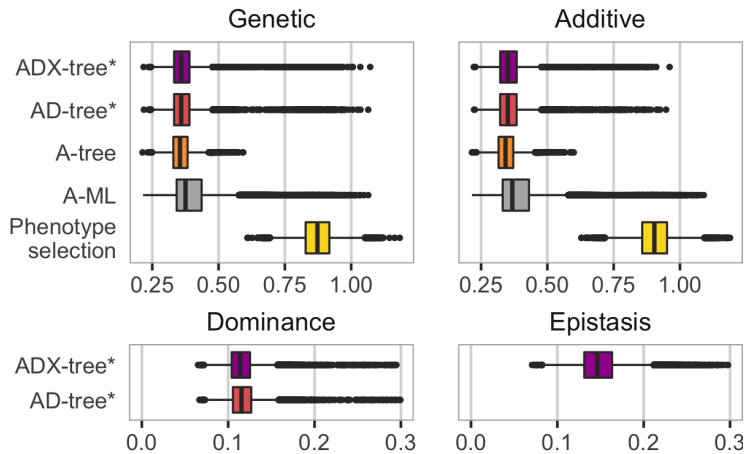


Figure S2.2: Accuracy of estimating the different genetic values for all individuals by model and prior setting - continuous rank probability score (CRPS; low value is desired, boxplots show variation over replicates). Genetic (upper left) means the estimated additive values for Model A, and the sum of estimated additive, dominance and epistasis values for Model ADX.

separately. While there was a tendency of more favourable correlation and CRPS for certain model and prior settings, the variation between replicates was much larger than variation between the model and prior settings. The model-wise prior tended to perform better than the component-wise prior, expert knowledge tended to perform better than the default non-informative prior knowledge and use of prior knowledge via the Bayesian approach tended to perform better than the maximum likelihood. All models performed better in estimating the genetic and additive values, especially in the terms of CRPS, than the phenotype selection where we treat the phenotype as a point estimate of the genetic value.

S2.2 Estimating variances

Variance component estimates varied considerably around the true values for all models and prior settings, but the estimates from the Bayesian inference showed slightly larger biases and smaller variance estimates than the maximum likelihood approach. We show this in Figure S2.3 with the ratio of estimated to true variances (value close to 1 is desired and values below/above 1 denote underestimation/overestimation). Of the model and prior settings in Figure S2.3,

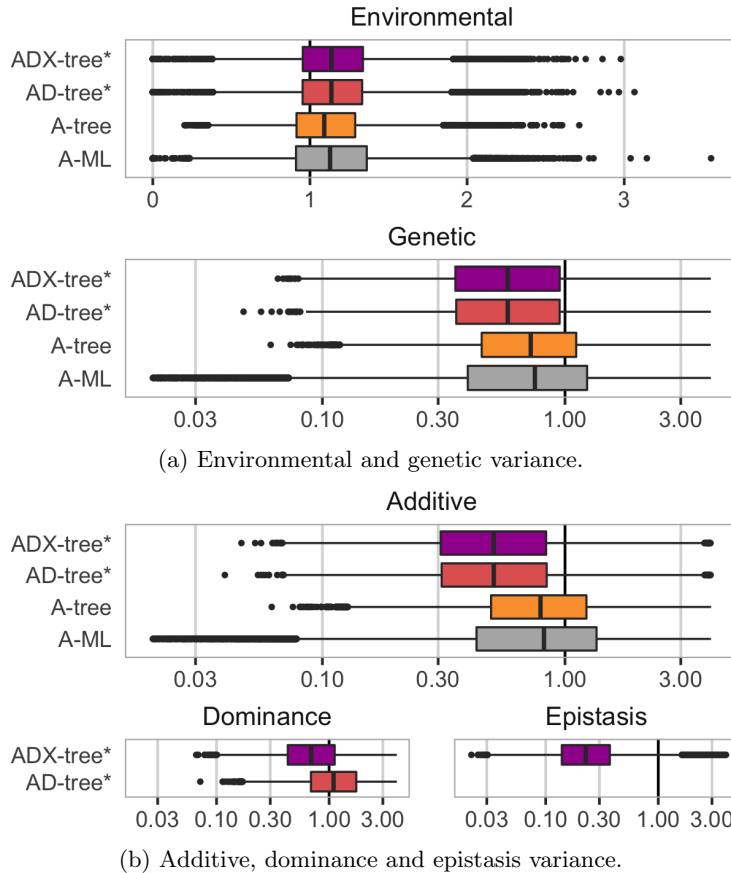


Figure S2.3: Accuracy of estimating (a) environmental and genetic variance and (b) additive, dominance and epistasis variance by model and prior setting - expressed as the estimated posterior median divided by the true value (a value close to 1 is desired, boxplots show variation over replicates, x -axes have a log-scale (except for environmental variance) and is focused on area around 1 with some outliers excluded).

A-ML was the closest to the true value on average in estimating the genetic variance, but also had the largest variation between replicates. Bayesian analysis with A-tree reduced variance between replicates, but did not improve bias. AD-tree* and ADX-tree* further increased the bias (underestimation) compared to A-ML and A-tree. When estimating dominance variance, AD-tree* performed better than ADX-tree*, but does not give estimates of the epistasis variance. Estimates for epistasis variance were considerably more underestimated than for the dominance variance.

In Figure S2.4 we show the posterior distributions of the environmental, additive, dominance and epistasis variances from one year in one simulated breeding program for the ADX-tree* setting (model-wise expert knowledge priors for the additive and nonadditive model). We see what we would expect: The environmental variance is larger than the variances of the genetic components, the additive effect stands for most of the genetic variation, and the dominance and epistasis variances are small.

We show the prior and posterior distributions of the phenotypic variance, the proportion of genetic to phenotypic variance, proportion of additive to genetic variance, proportion of dominance to nonadditive variance, also for the ADX-tree* setting in Figure S2.5. We see that the data informs about the phenotypic variance and the broad-sense heritability, but only weakly informs about the two other splits.

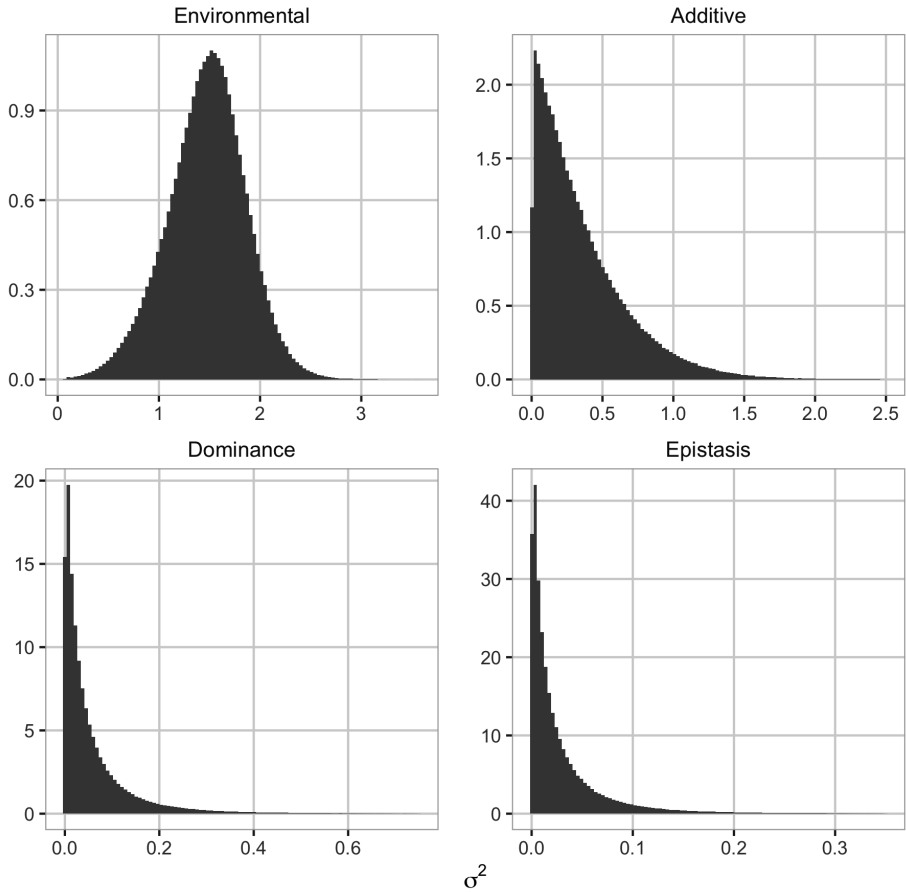


Figure S2.4: Posterior distribution of the environmental, additive, dominance and epistasis variance from the ADX-tree* setting, from one year in one simulated breeding program. Priors are not plotted because the prior on the phenotypic variance σ_P^2 and thus also on the variance parameters σ_e^2 , σ_a^2 , σ_d^2 and σ_x^2 , are scale-invariant, and therefore improper.

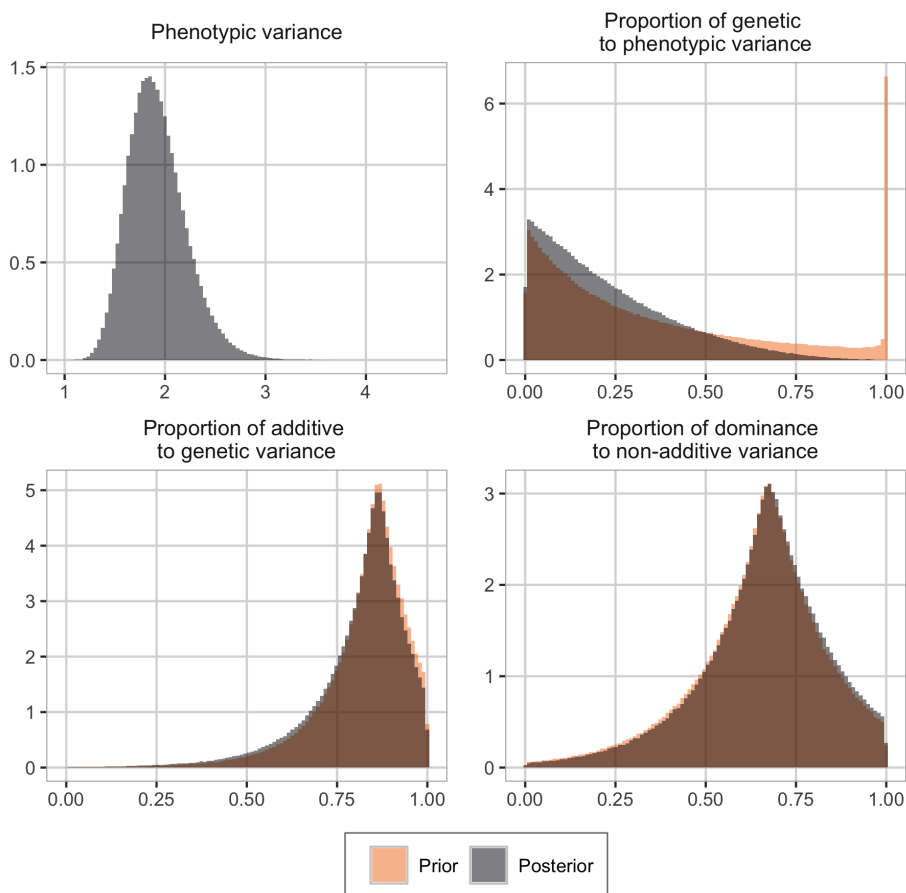


Figure S2.5: Prior and distribution of the phenotypic variance and variance proportions from the ADX-tree* setting, from one year in one simulated breeding program. The prior is not plotted for phenotypic variance σ_P^2 because it is scale-invariant, and therefore improper.

S2.3 Increasing number of observations

Figure S2.6 shows the ability of estimating the model variances for increasing number of individuals for the additive (A) model and the additive and nonadditive (ADX) model. The plot shows the posterior median from each model fit divided by the true variance from the simulated breeding program. Figure S2.6 shows the variance estimates from Models A and ADX, and datasets of size 100, 300, 500 and 700 here, and include the full results with variance estimates, correlation and continuous rank probability score (CRPS) for all models and number of individuals in the Figures S14-S16. From the environmental variance estimates we see that the variation between replicates decreases for all models for increasing number of observations. The maximum likelihood approach underestimated the additive, dominance and epistasis variances to a larger extent than the Bayesian approach did, and this underestimation decreased when the number of individuals increased. However, 700 observations is not enough for the maximum likelihood approach to obtain a bias in dominance and epistasis variance estimates as low as the Bayesian approach, indicating that the need for good priors decrease with increasing number of observations, but suitable priors are still necessary also for 700 observations. The inference stability did not increase with increasing number of observations for any of the models fitted with the maximum likelihood approach. The Bayesian models had the same high inference stability as in Table 2 in the main paper.

The correlation did not differ significantly between models and approaches, and increased with increasing number of observations for all settings (Figure S15). The CRPS of the genetic and additive effects was significantly lower (better) for the models fitted with the Bayesian approach for a low number of individuals, but the maximum likelihood approach improved quickly when the dataset size increased (Figure S14). The Bayesian models had a significantly lower CRPS of the dominance and especially epistasis effects than the maximum likelihood approach for all dataset sizes. The results from the additive and dominance (AD) model did, with an exception of slightly more overestimation of the dominance variation for the maximum likelihood approach, not differ from the results from the nonadditive model (ADX).

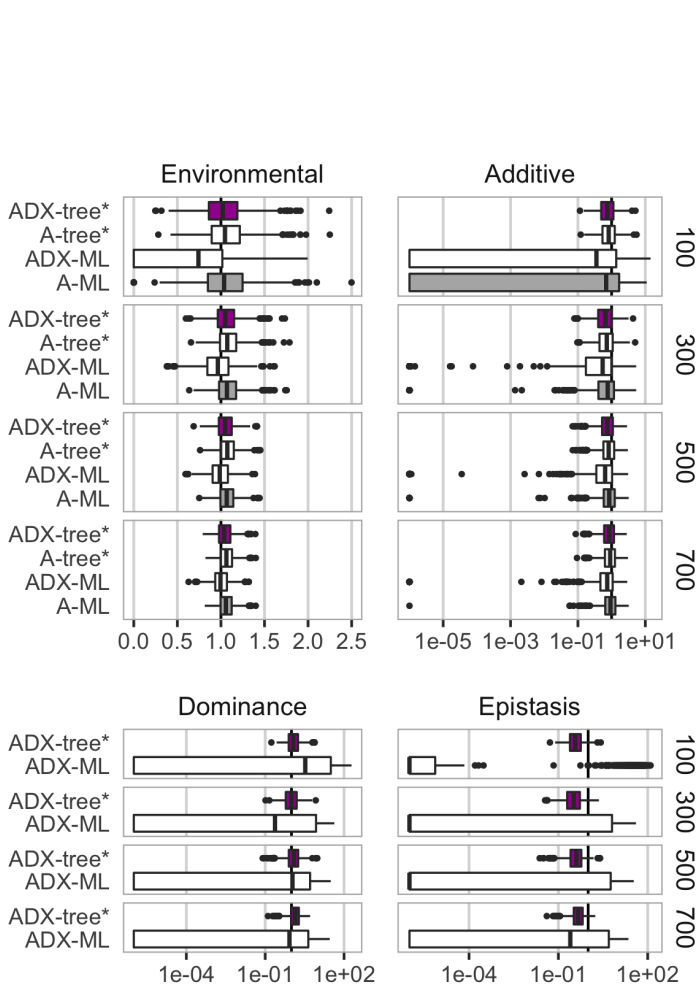


Figure S2.6: Accuracy of estimating environmental, additive, dominance and epistasis variance expressed as the estimated posterior median divided by the true value (a value close to 1 is desired). The dataset size is indicated for each box, and the y-axis shows the model and prior settings. x -axes have a log-scale (except for environmental variance), and all values smaller than 10^{-6} are set to 10^{-6} as those values are essentially zero.

S3 Supplemental tables and figures

Trial	No. of obs.
Adenstedt (Ade13)	1,729
Böhnshausen (Boh12)	1,101
Böhnshausen (Boh13)	1,692
Hadmersleben (Had12)	1,738
Hadmersleben (Had13)	1,669
Harzhof (Hhof12)	1,736
Harzhof (Hhof13)	1,738
Hohenheim (Hoh12)	1,720
Hohenheim (Hoh13)	1,703
Seligenstadt (Sel12)	834
Seligenstadt (Sel13)	1,739

Table S1: The number of observed phenotypes for each of the 11 trials in the 6 locations in Germany for the Central European wheat dataset. The total number of individuals in the dataset is 1,739, where 15 are male parents, 120 are female parents and 1,604 are hybrids. Names in parentheses are the abbreviations used.

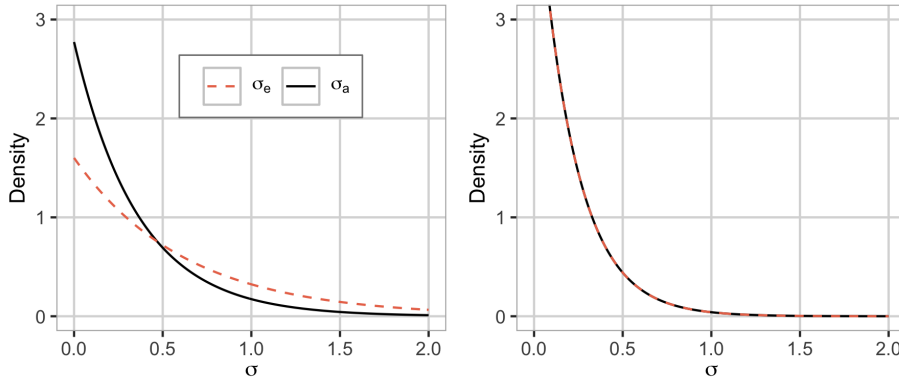


Figure S1: The prior used for the A-comp* (additive model with component-wise expert knowledge prior) (left) and A-comp (additive model with component-wise default prior) (right) settings. For A-comp*, we use $h_g^2 = 0.25$. We have plotted the priors for $V_p = 1$. For A-comp, additive and environmental variances have the same prior.

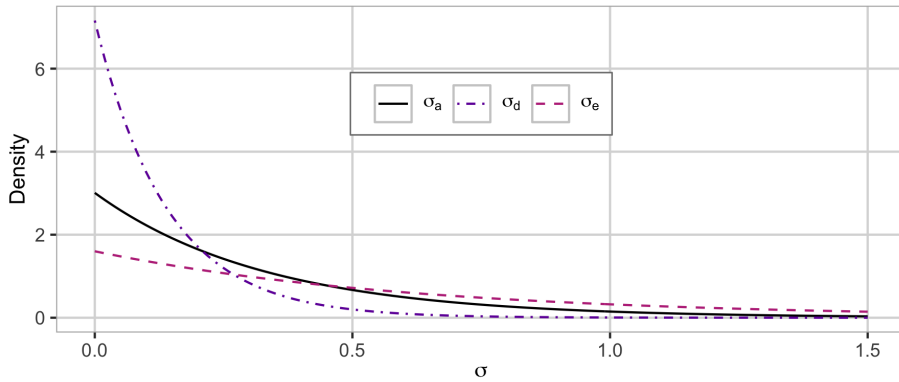


Figure S2: The prior used for the AD-comp* (additive and dominance model with component-wise expert knowledge prior) setting. We use $R_{\frac{g}{g+e}} = 0.25$ and $R_{\frac{p}{g}} = 0.85$. We have plotted the priors for $V_p = 1$.

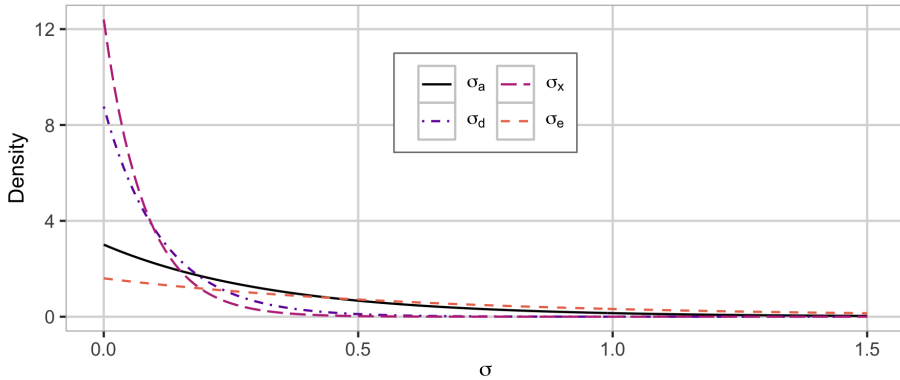


Figure S3: The prior used for the ADX-comp* (additive and nonadditive model with component-wise expert knowledge prior) setting. We use $R_{\frac{g}{g+e}} = 0.25$, $R_{\frac{a}{g}} = 0.85$ and $R_{\frac{d}{d+x}} \approx 0.67$. We have plotted the priors for $V_p = 1$.

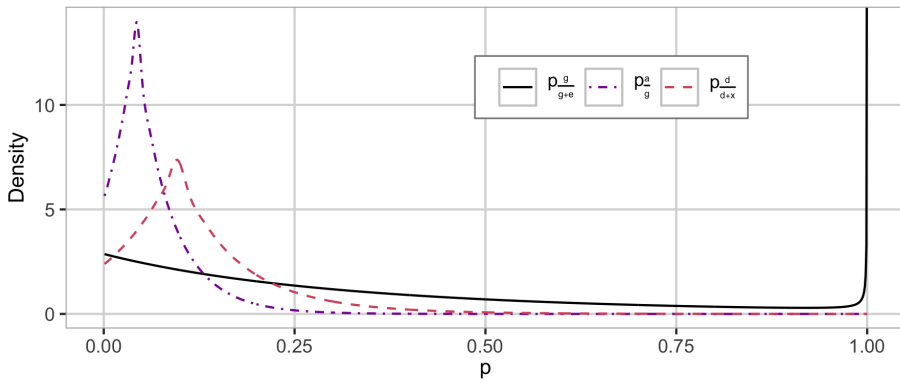
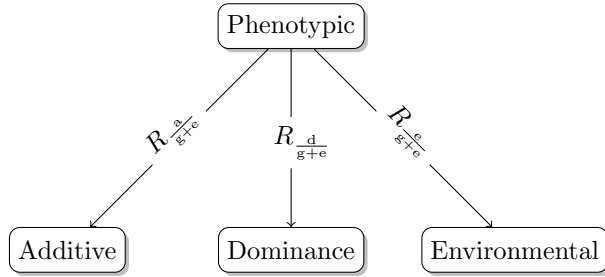
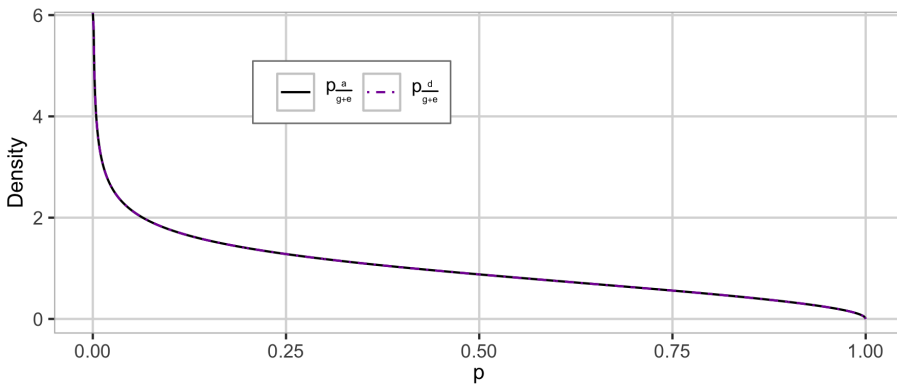


Figure S4: The HD prior used for the ADX-tree-opp* (additive and nonadditive model with model-wise opposite expert knowledge prior) setting with the proportion of genetic to phenotypic variance $p_{\frac{g}{g+e}}$, additive to genetic variance $p_{\frac{a}{g}}$, and dominance to nonadditive variance $p_{\frac{d}{d+x}}$. We use $R_{\frac{g}{g+e}} = 0.25$, $R_{\frac{a}{g}} = 0.05$ and $R_{\frac{d}{d+x}} \approx 0.11$. This is a dataset specific prior.

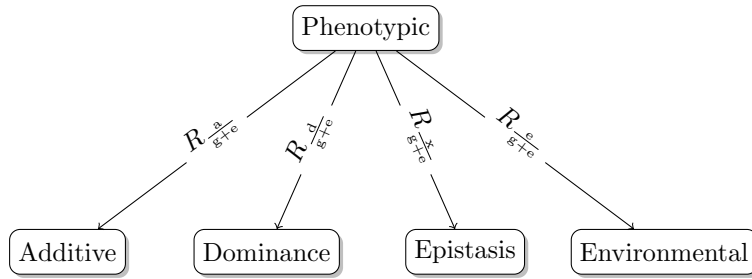


(a) Tree structure.

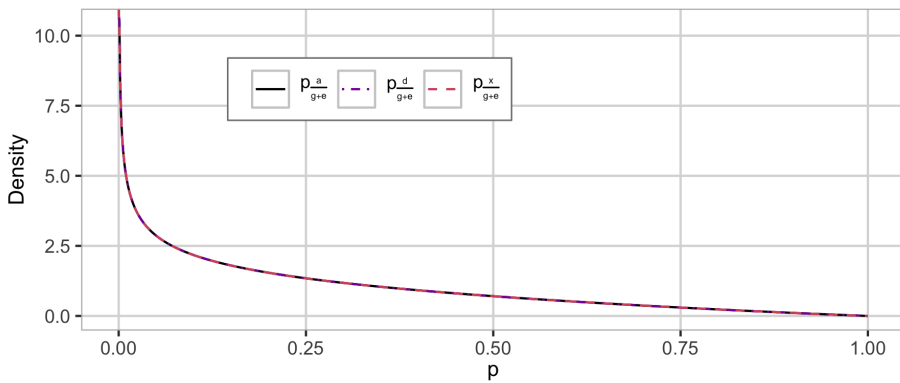


(b) Prior.

Figure S5: The (a) tree structure and (b) HD prior for the AD-tree (additive and dominance model with model-wise default prior) setting with equal magnitude for the four sources of variation without using expert knowledge - the proportion of additive to phenotypic variance $p_{\frac{a}{g+e}}$, and dominance to phenotypic variance $p_{\frac{d}{g+e}}$. This corresponds to a Dirichlet (3) prior on the variance proportions.



(a) Tree structure.



(b) Prior.

Figure S6: The (a) tree structure and (b) HD prior for the ADX-tree (additive and nonadditive model with model-wise default prior) setting with equal magnitude for the four sources of variation without using expert knowledge - the proportion of additive to phenotypic variance $p_{\frac{a}{g+e}}$, dominance to phenotypic variance $p_{\frac{d}{g+e}}$, and epistasis to phenotypic variance $p_{\frac{x}{g+e}}$. This corresponds to a Dirichlet(4) prior on the variance proportions.

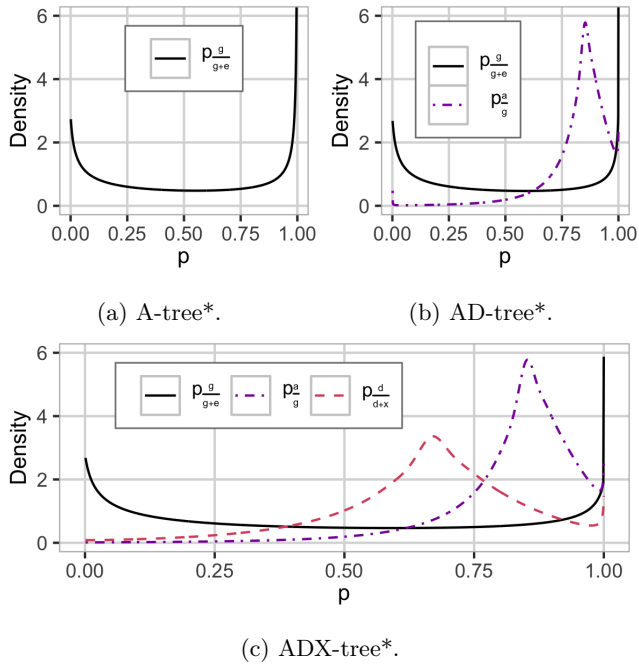


Figure S7: The model-wise expert knowledge HD prior used in (a) A-tree*, (b) AD-tree* and (c) ADX-tree* settings in the analysis of the Central European winter wheat data. $R_{\frac{g}{g+e}} = 0.75$, $R_{\frac{a}{g}} = 0.85$ and $R_{\frac{d}{d+x}} \approx 0.67$.

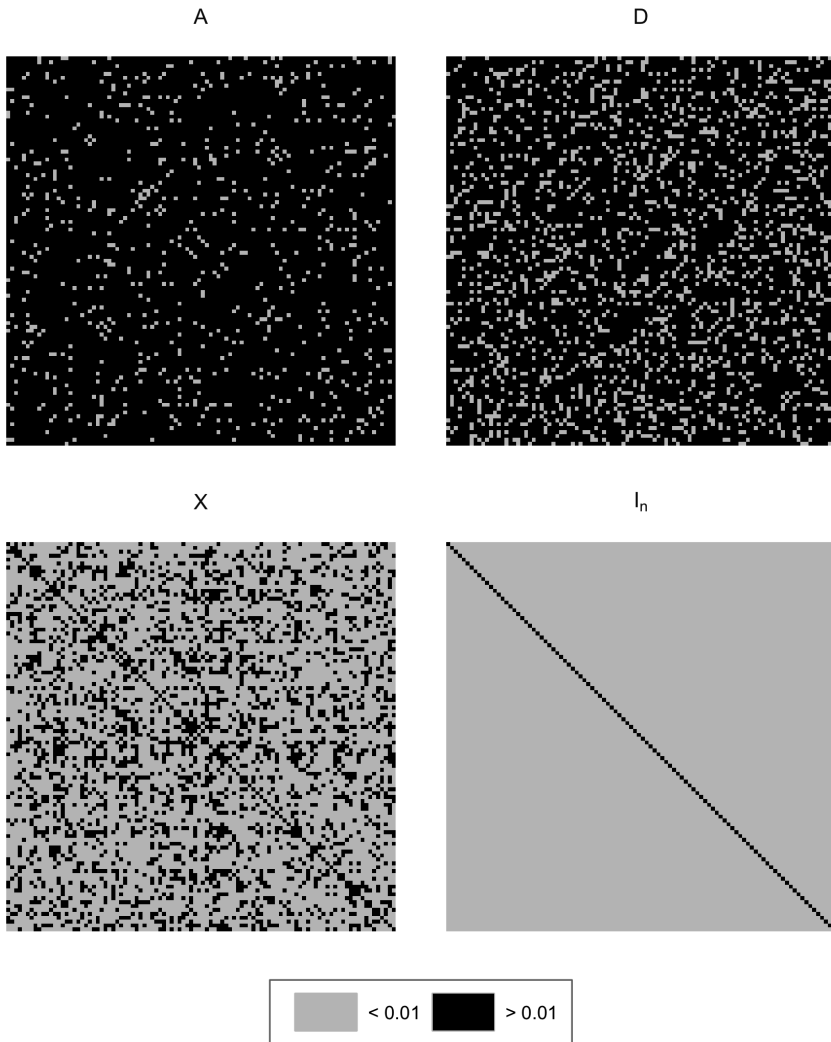


Figure S8: Covariance matrices for the additive (**A**), dominant (**D**), epistasis (**X**) and environmental (I_n) sources of variation for one year in one simulated breeding program.

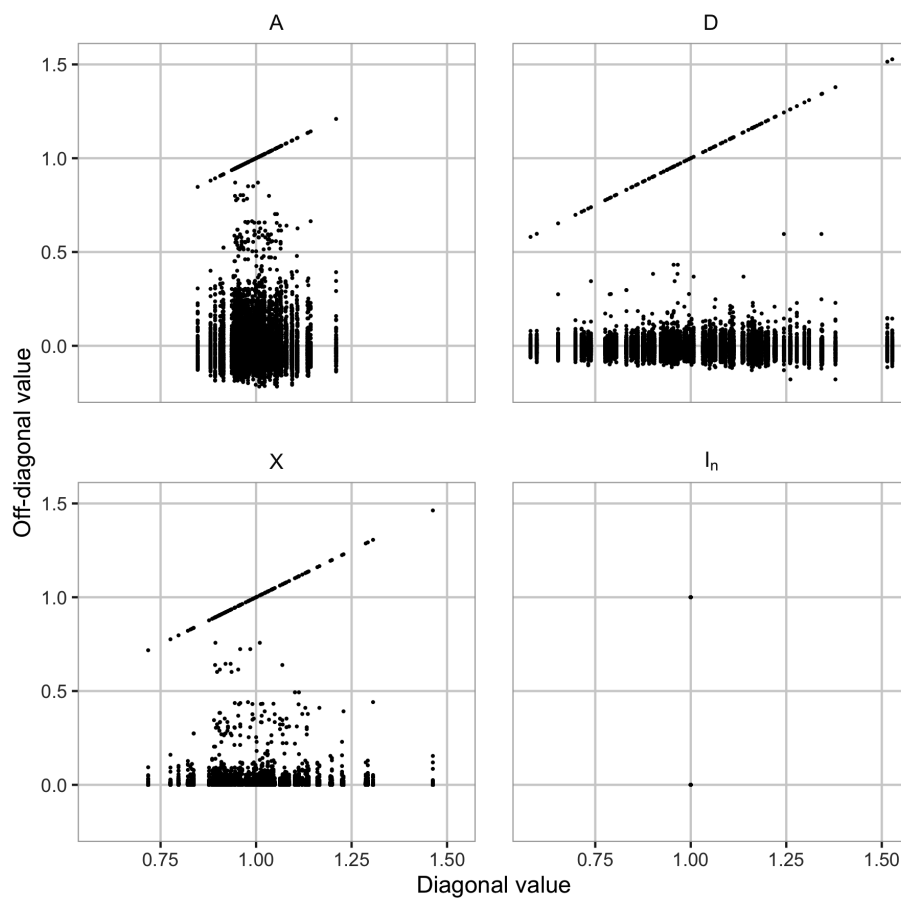


Figure S9: Scatterplot of entries of the covariance matrices for the additive (**A**), dominant (**D**), epistasis (**X**) and environmental (I_n) sources of variation for one year in one simulated breeding program. The off-diagonal values for each row is plotted pairwise against the diagonal value on the same row.

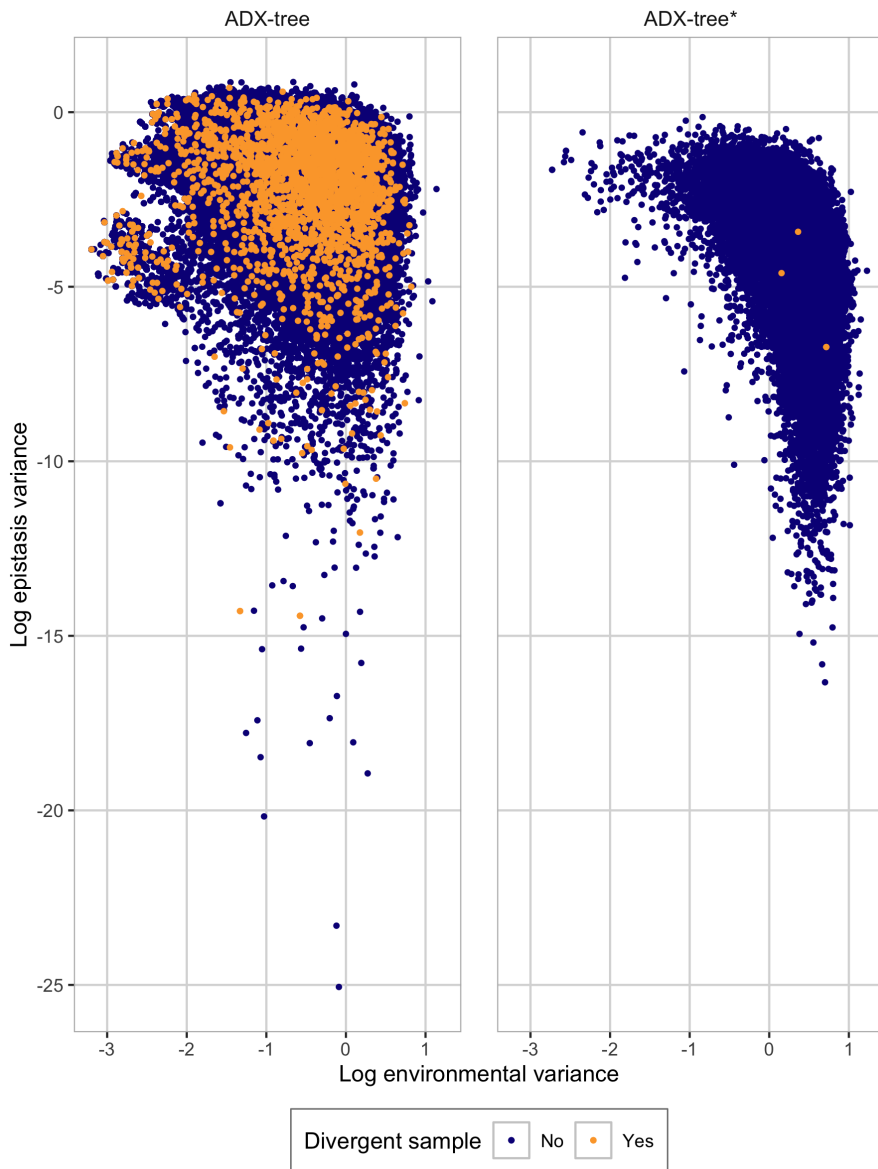


Figure S10: The joint posterior of the environmental and epistasis variances (log-scale) for one year in one simulated breeding program with the ADX-tree (left) and ADX-tree* (right) settings. By a divergent sample we mean a sample where the MCMC sampler had a divergent transition.

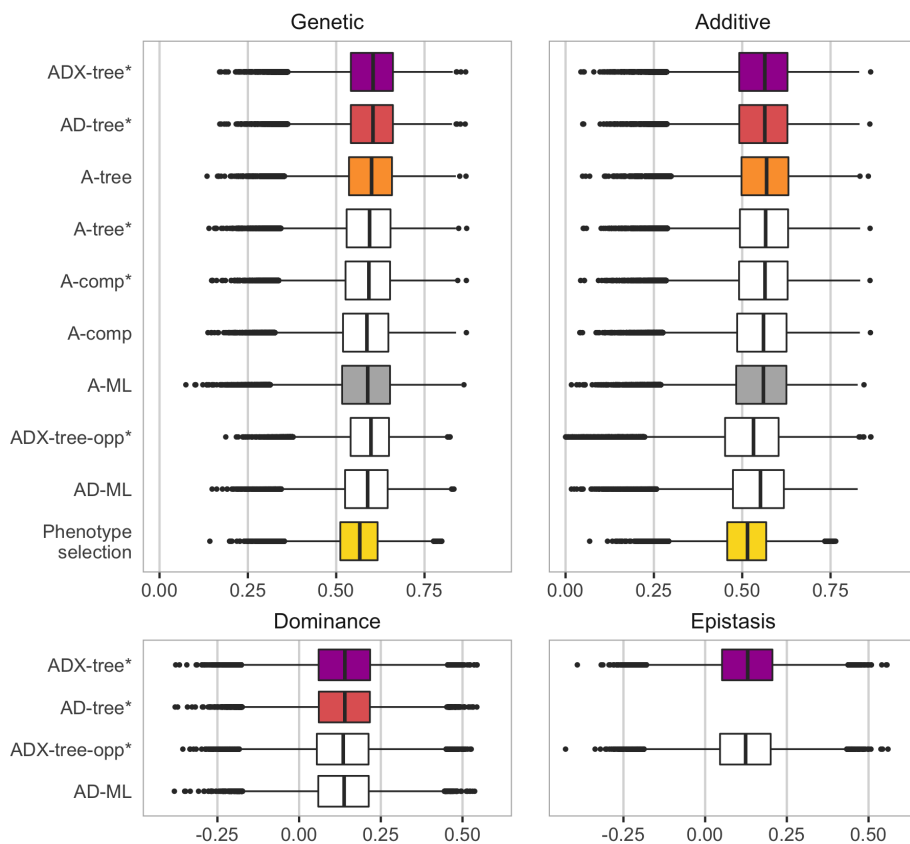


Figure S11: The ability to estimate the different genetic values for all individuals by the model and prior setting - correlation (high value is desired, boxplots show variation over replicates). Genetic (upper left) means the estimated additive values for Model A, the sum of the estimated additive and dominance values for Model AD, and the sum of estimated additive, dominance and epistasis values for Model ADX.

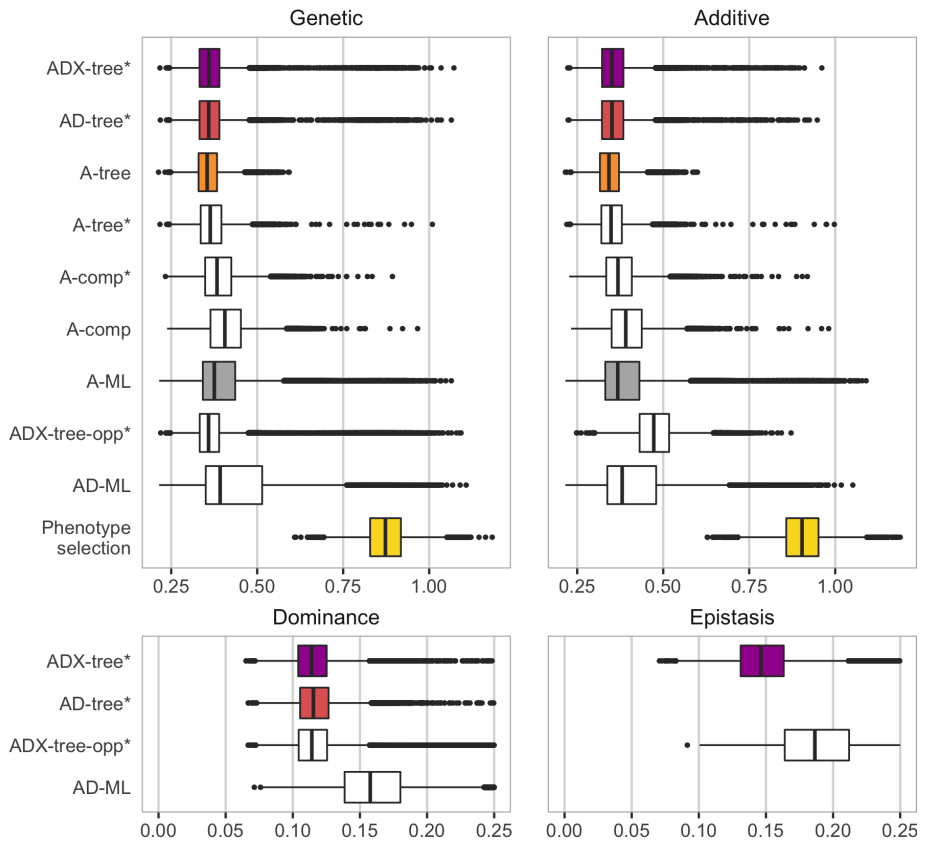
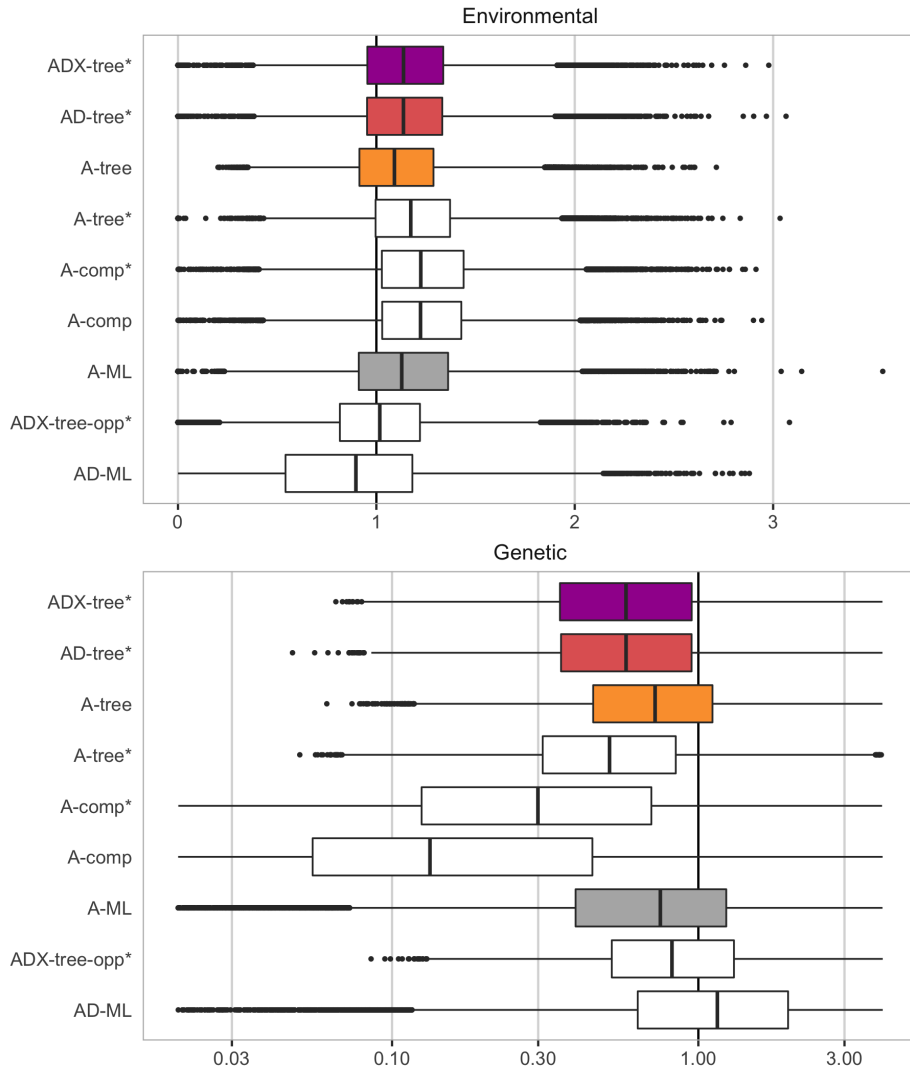
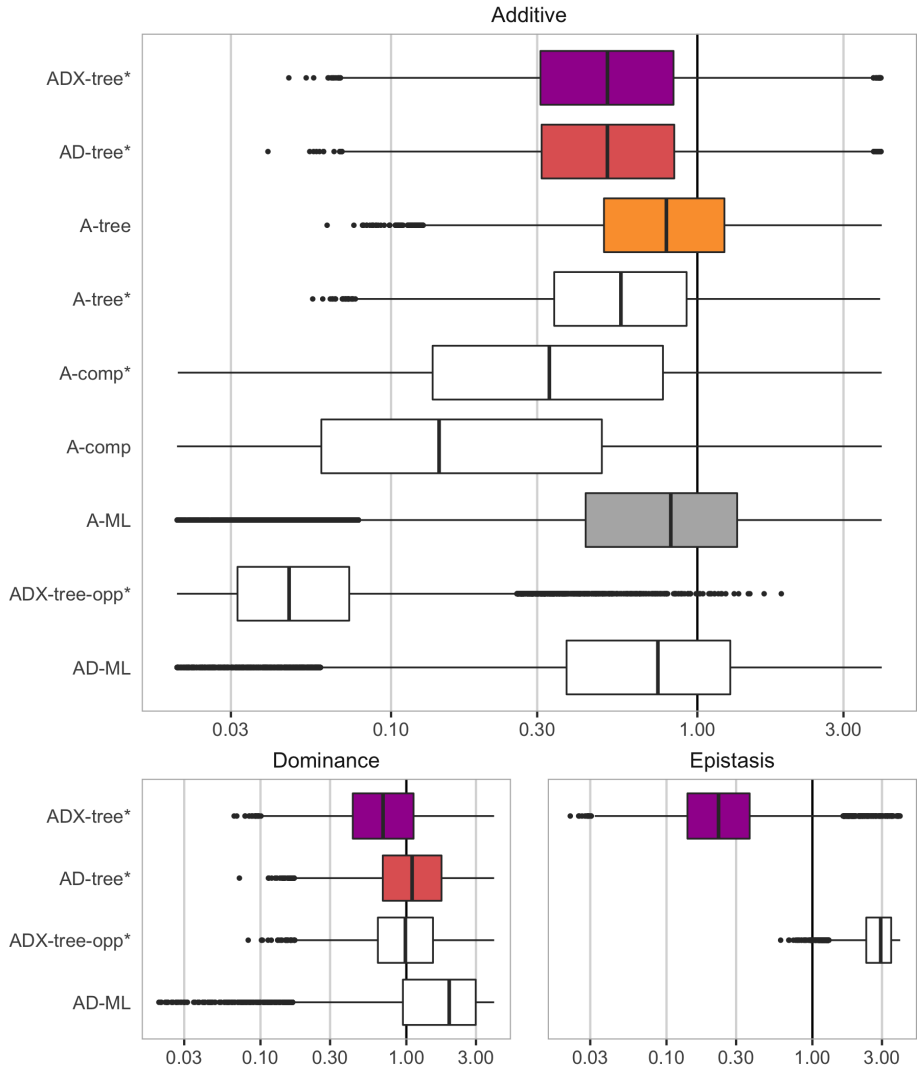


Figure S12: The ability to estimate the different genetic values for all individuals by the model and prior setting - continuous rank probability score (CRPS; low value is desired, boxplots show variation over replicates). Genetic (upper left) means the estimated additive values for Model A, the sum of the estimated additive and dominance values for Model AD, and the sum of estimated additive, dominance and epistasis values for Model ADX.

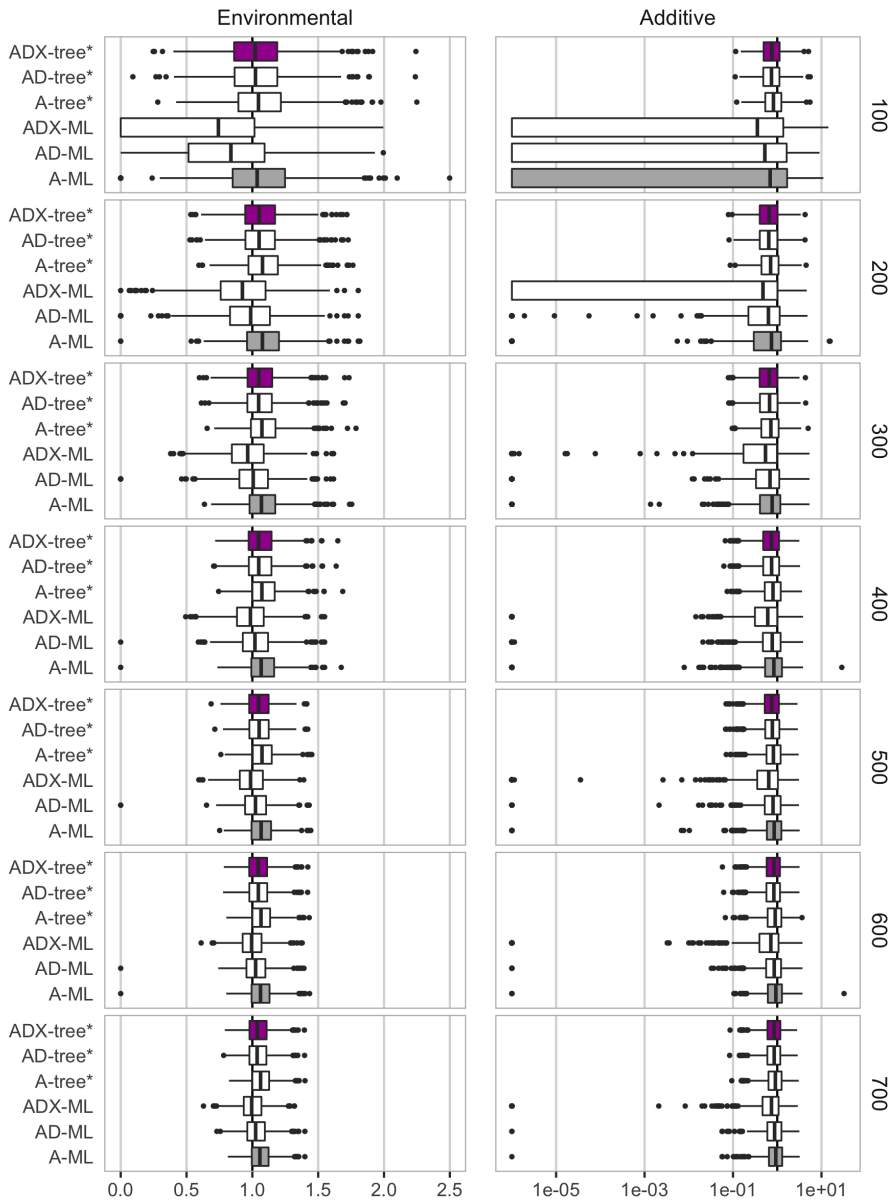


(a) The environmental and genetic variance.

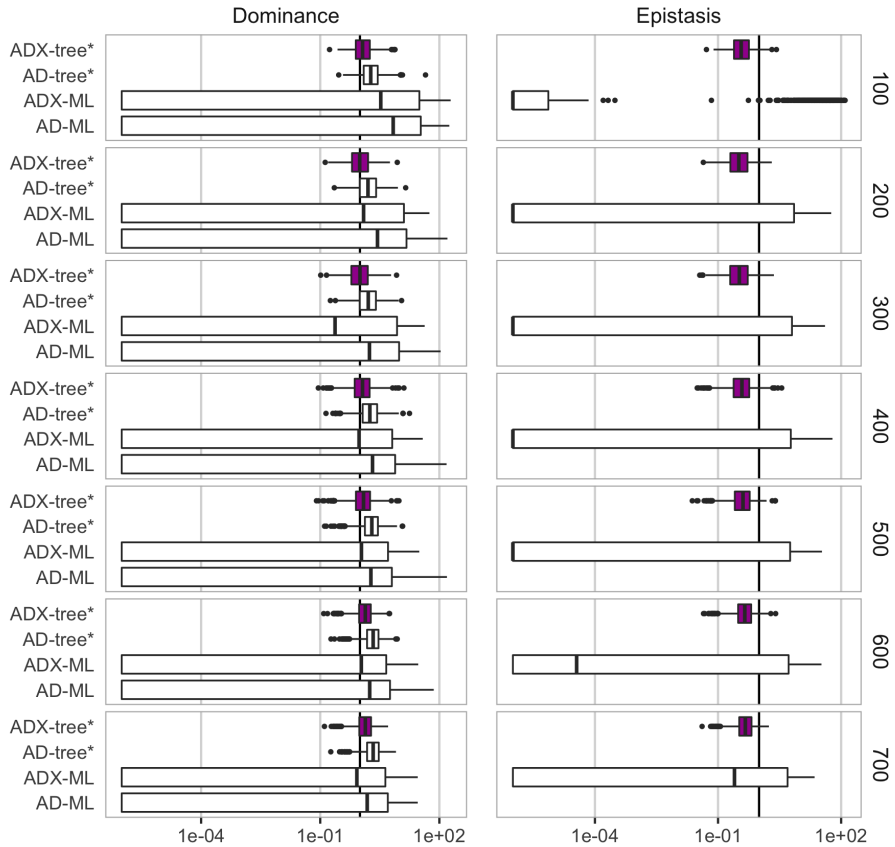


(b) The additive, dominance and epistasis variance.

Figure S13: The ability to estimate (a) environmental and genetic variance and (b) additive, dominance and epistasis variance by the model and prior setting - expressed as the estimated posterior median divided by the true value (a value close to 1 is desired, boxplots show variation over replicates, x -axes have a log-scale (except for environmental variance) and is focused on area around 1 with some outliers excluded).

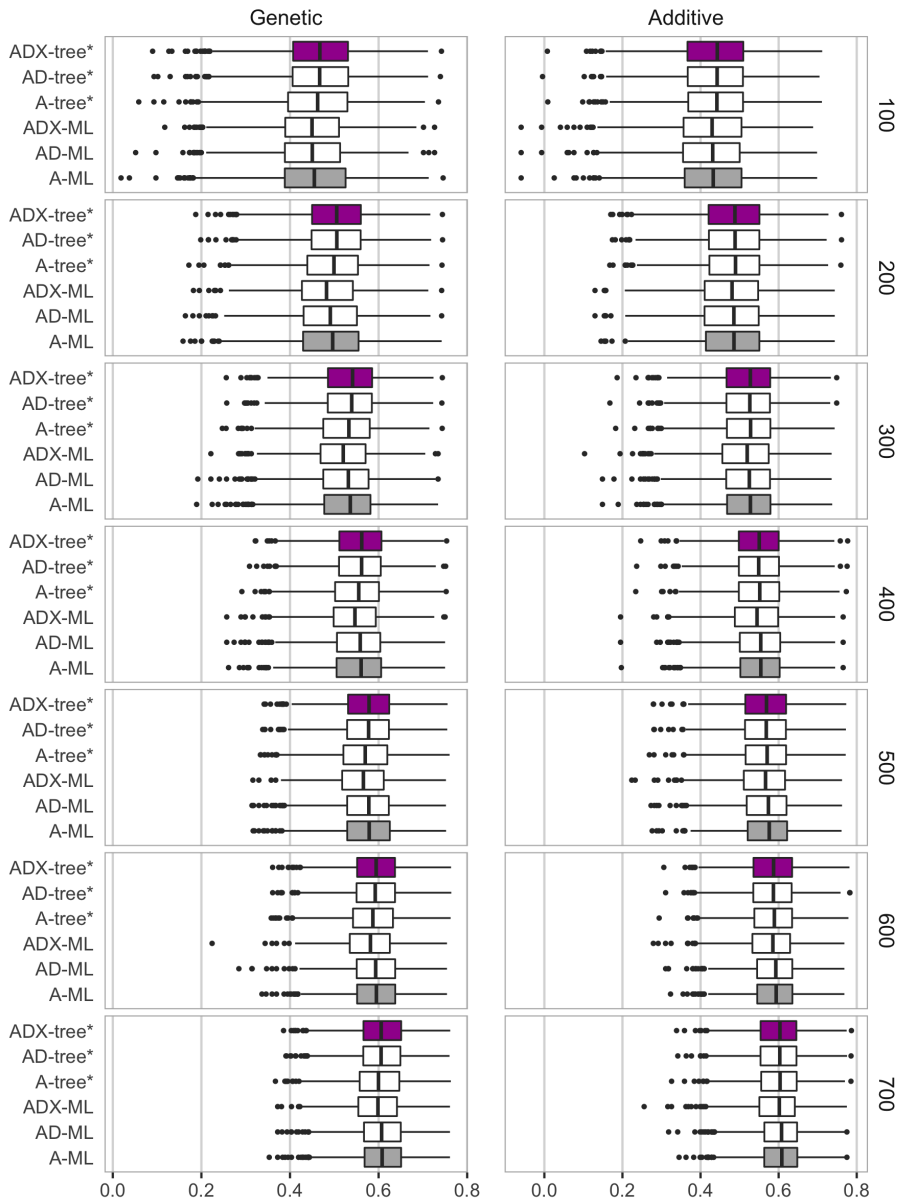


(a) The environmental and additive variance.

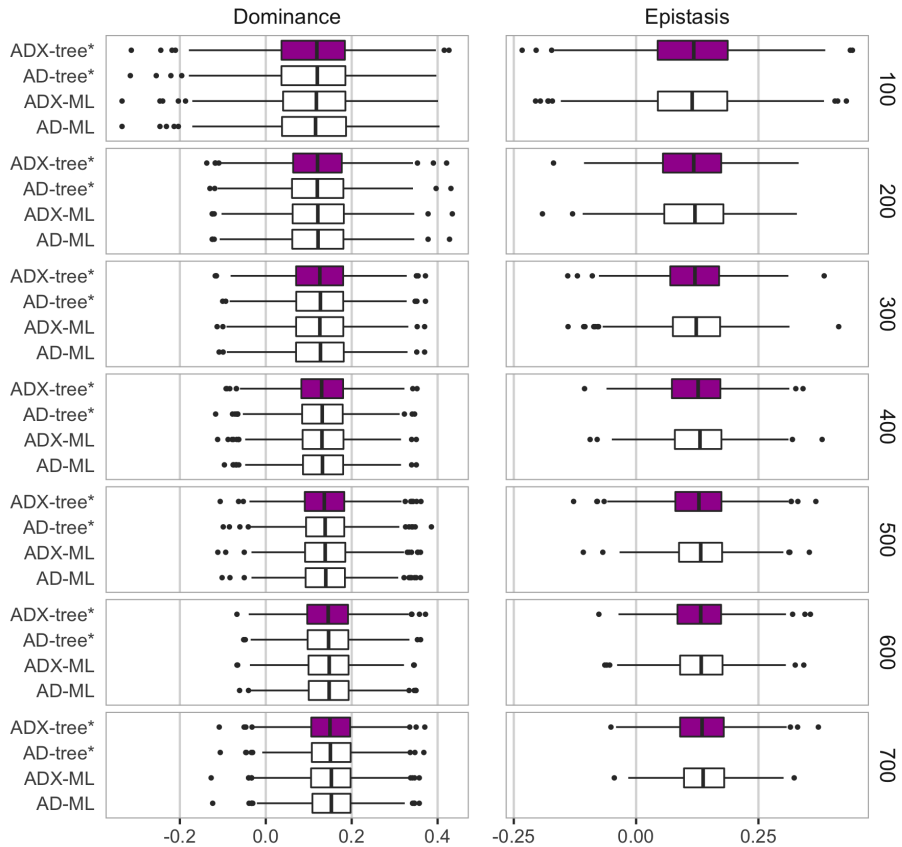


(b) The dominance and epistasis variance.

Figure S14: The ability to estimate (a) environmental and additive variance and (b) dominance and epistasis variance by model, prior setting and size - expressed as the estimated posterior median divided by the true value (a value close to 1 is desired).

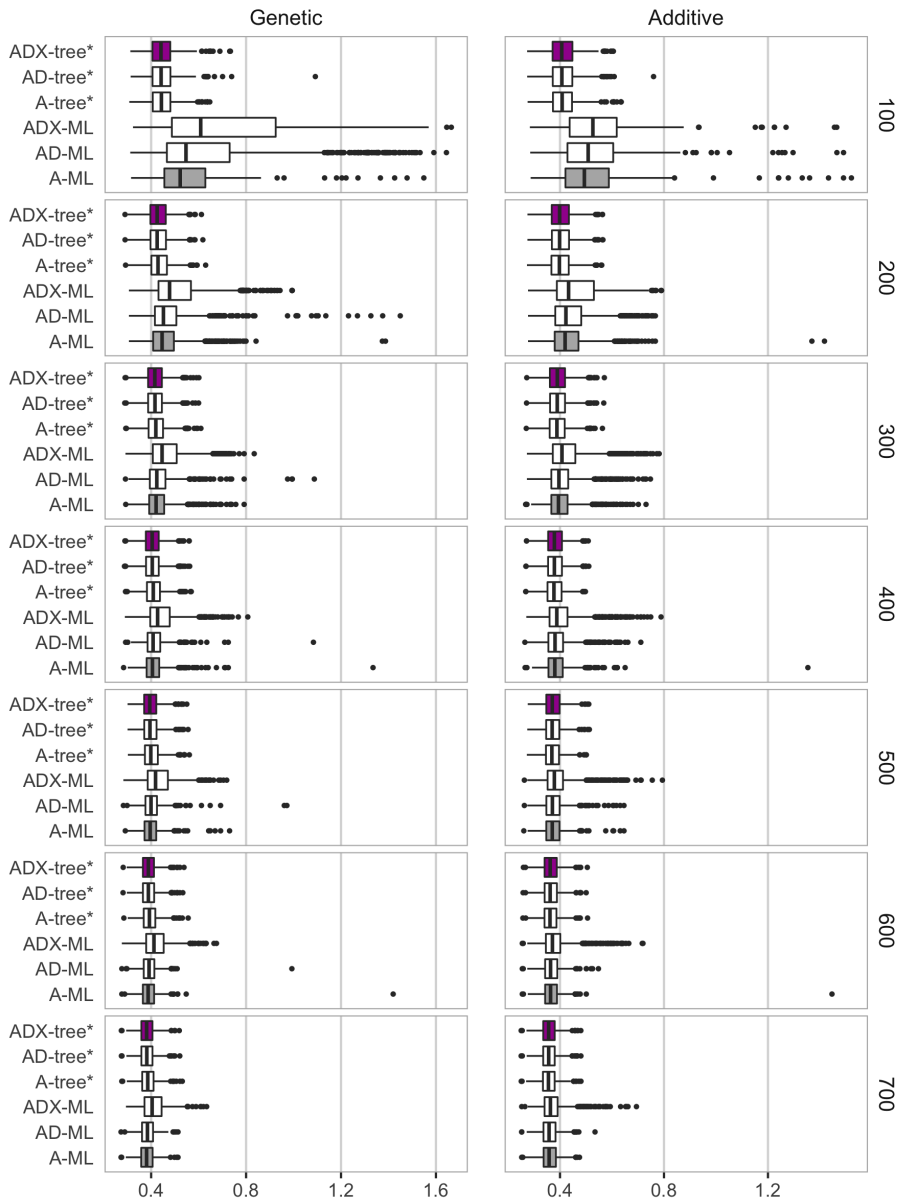


(a) The environmental and additive effect.

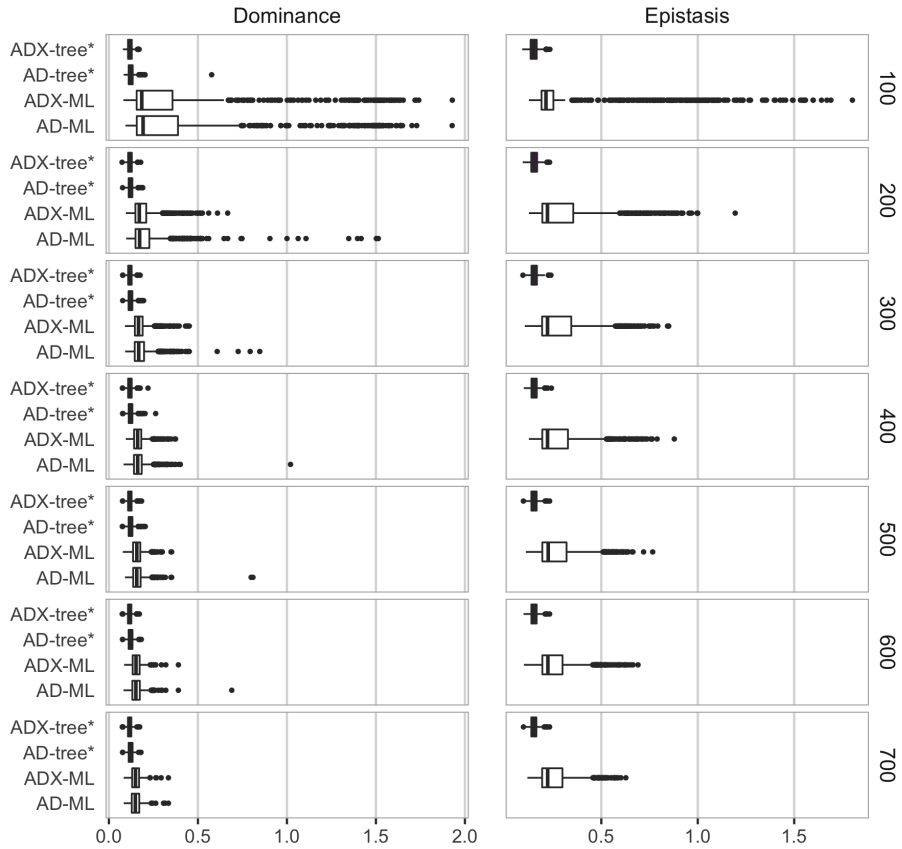


(b) The dominance and epistasis effect.

Figure S15: The ability to estimate (a) environmental and additive effect and (b) dominance and epistasis effect by the model, prior setting and size - correlation (high value is desired, boxplots show variation over replicates).



(a) The environmental and additive variance.



(b) The dominance and epistasis variance.

Figure S16: The ability to estimate (a) environmental and additive effect and (b) dominance and epistasis effect by the model, prior setting and size - continuous rank probability score (CRPS; low value is desired, boxplots show variation over replicates).

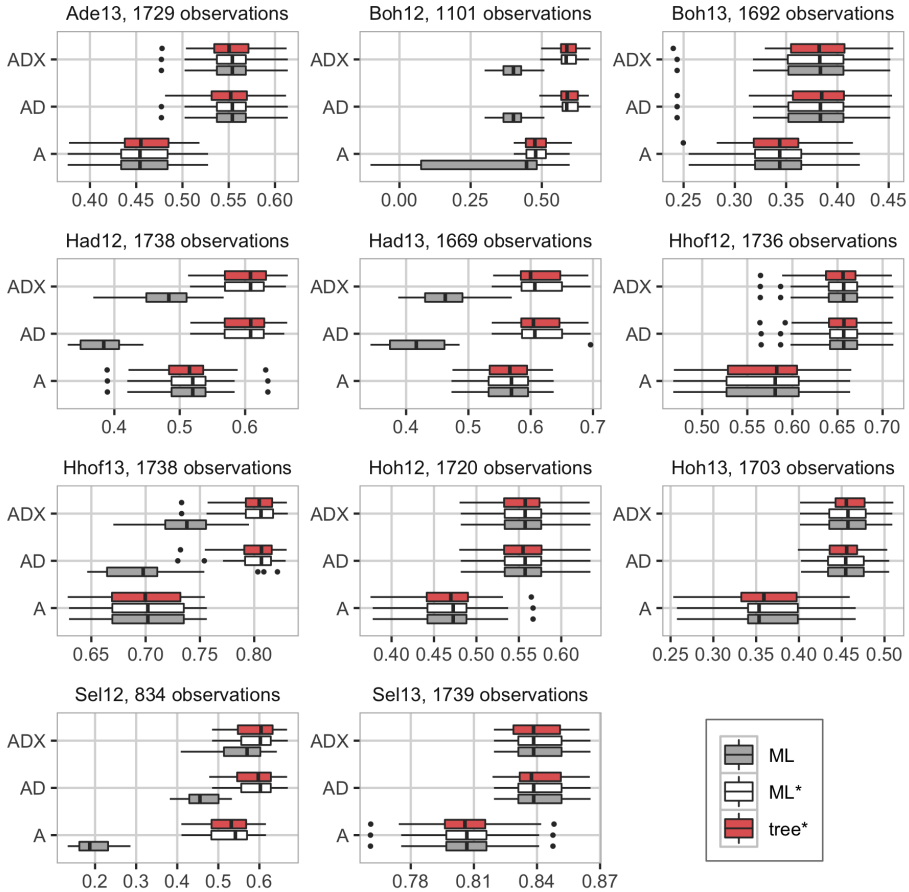


Figure S17: The ability of phenotype prediction in the real case study for all 11 trials, measured using correlation (high value is desired, boxplots show variation over cross-validations and folds). The number of observations in each dataset is indicated for each trial. The total number of parents and hybrids is 1,739.

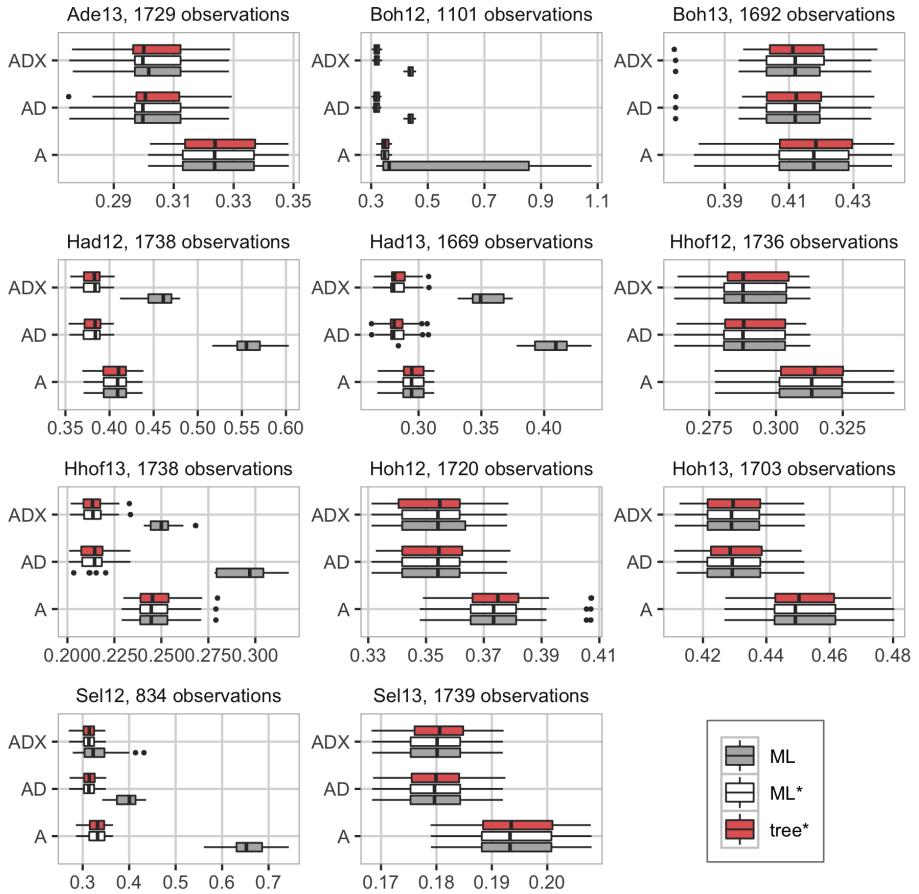
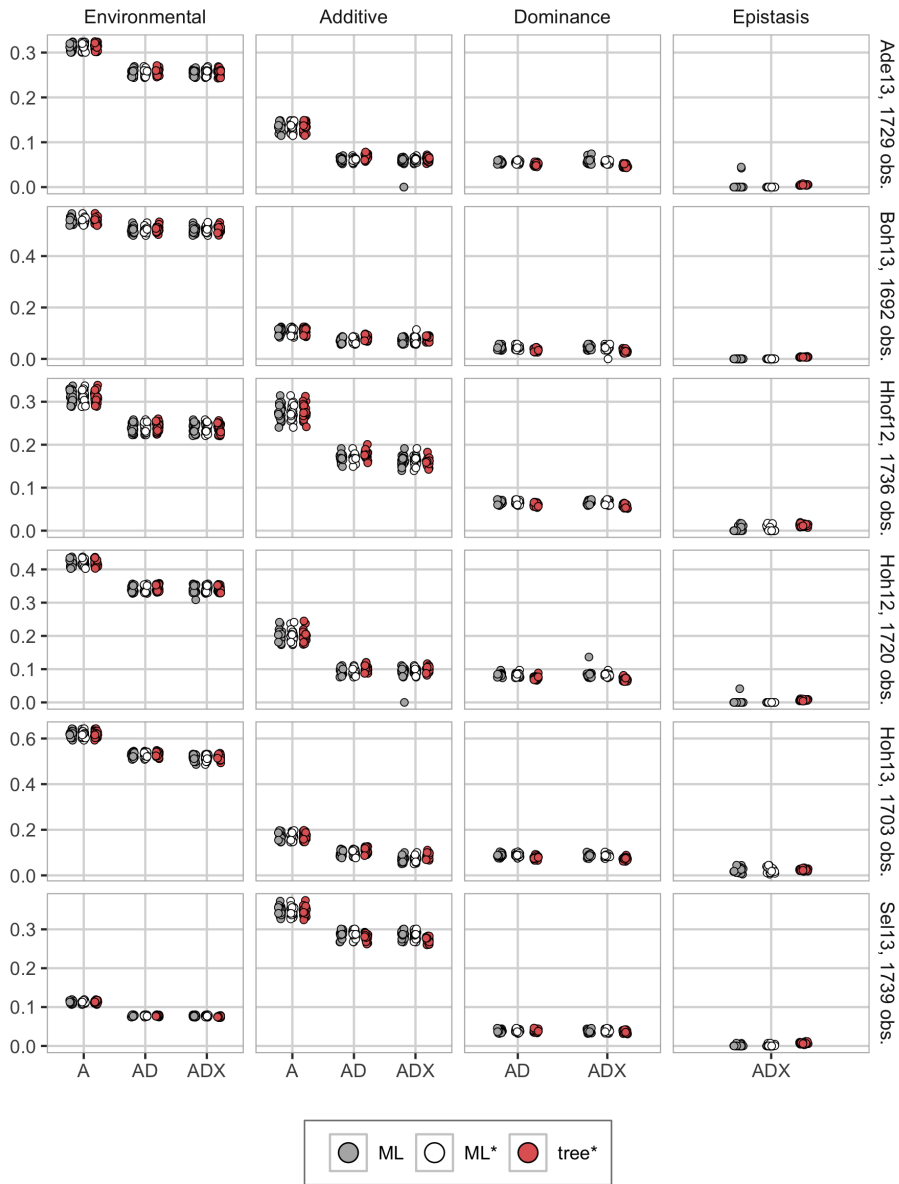


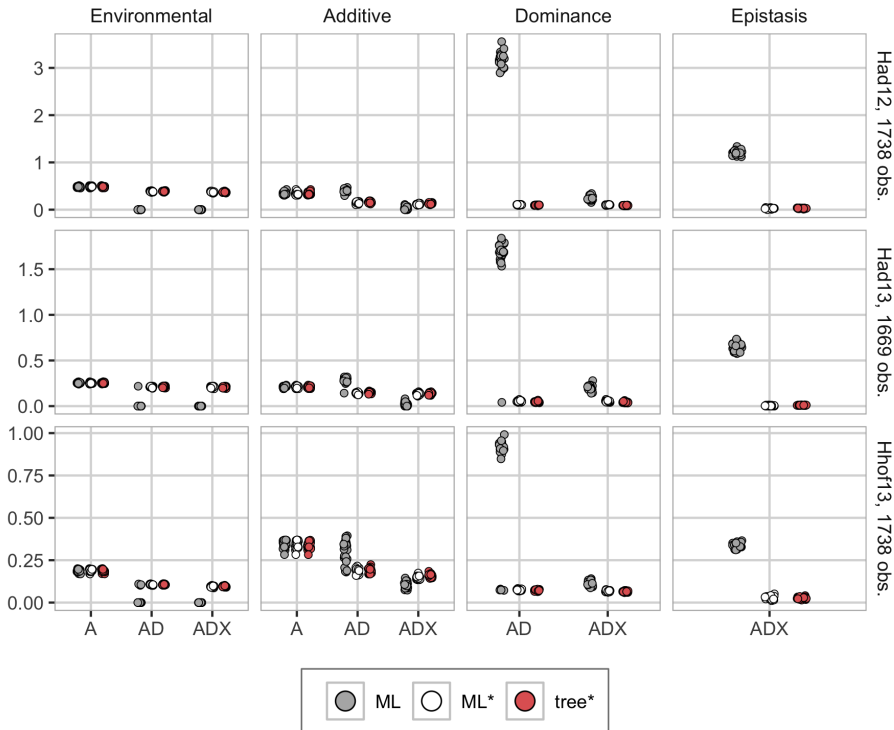
Figure S18: The ability of phenotype prediction in the real case study for all 11 trials, measured using continuous rank probability score (CRPS; low value is desired, boxplots show variation over cross-validation and folds). The number of observations in each dataset is indicated for each trial. The total number of parents and hybrids is 1,739.



(a) Trials where the approaches perform equally good.



(b) Trials where we have a lot of unobserved phenotypes, and ML is diverging. For Boh12 and Sel12, A-ML (additive model fitted with the maximum likelihood approach) is overestimating the additive variance to be so large (estimates over 900 and 400, respectively) that we have truncated the y-axes at 2.5 and 1.5, respectively, to highlight the other results.



(c) Trials where the maximum likelihood approach leads to overfitting.

Figure S19: Posterior median variances from (a) Ade13, Boh13, Hhof12, Hoh12, Hoh13 and Sel13, where both approaches perform equally good, (b) Boh12 and Sel12, where large amounts of the phenotypes are unobserved and the maximum likelihood approach is diverging, and (c) Had12, Had13 and Hhof13, where ML is overfitting. We have included the variances from the five 5-fold cross-validations, giving 25 estimates for each trial and model. We have removed results where the maximum likelihood optimizer did not converge. The y-axes of (b) are truncated to highlight the other results. The total number of parents and hybrids is 1,739.

Paper III

makemyprior: Intuitive construction of joint priors for variance parameters in R

Hem, I. G., Fuglstad, G.-A., and Riebler, A.

2021. In preparation.

makemyprior: Intuitive Construction of Joint Priors for Variance Parameters in R

Ingeborg Gullikstad Hem¹, Geir-Arne Fuglstad¹, and Andrea Riebler¹

¹Department of Mathematical Sciences, NTNU, Norway

Abstract

`makemyprior` is an R package that formulates prior distributions for variance parameters taking the entire model structure into account. Existing prior or expert knowledge can be intuitively incorporated at the level it applies to. Independent of the existence of any prior knowledge the package enhances the consciousness of prior selection and makes it an active component in the Bayesian workflow performed by the applied researcher. The prior distribution is constructed based on a hierarchical decomposition of the total variance in the model along a tree. It is proper and leads to robust inference. The user controls the prior construction by specifying their prior beliefs or ignorance at each level of the prior tree and obtains the resulting prior distributions without any additional involvement. A graphical user interface facilitates prior construction through visualization of the tree and returns the priors graphically for each variance proportion parameter.

Keywords: Bayesian hierarchical models, robust inference, variance parameters, prior distributions, hierarchical variance decomposition, graphical user interface, R.

1 Introduction

Bayesian modelling is more available than ever through fast and easy-to-use softwares for Bayesian inference such as Integrated Nested Laplace Approximations (INLA, Rue et al., 2009), Stan (Carpenter et al., 2017) with the R interface `rstan` (Stan Development Team, 2020) and dependencies such as `rstanarm` (Goodrich et al., 2020), `shinystan` (Gabry, 2018) and `loo` (Vehtari et al., 2020), Template Model Builder (TMB, Kristensen et al., 2016), JAGS (Plummer, 2017),

and Bayesian Analysis Toolkit (BAT, Caldwell et al., 2009). These softwares offer many ways to construct complex models suitable for a wide range of applications, and often give default settings and as such not requiring any action by the user. It is convenient to use the default priors, but then the Bayesian framework is not fully utilized. How to implement the priors is explained in the softwares, but they lack thorough guides on how the priors should be chosen, even though there is an increasing focus on that prior distributions should be chosen consciously (Zondervan-Zwijnenburg et al., 2017; Gelman et al., 2020; Smid and Winter, 2020). Our goal is to empower users to actively select priors that are suitable for their model structure and application at hand, and to increase the awareness of prior choices. The user is confronted with which priors are chosen and what they express, also when the default settings are chosen.

We focus on Bayesian hierarchical models that model the variation in the observations through a combination of an observation model, a linear latent model, and priors for the parameters. The observation model defines a distribution for the observed data conditional on the latent model. The link between the latent model and each observation acts through a transformation of a linear predictor, which is a linear combination of fixed and random effects. The goal of the linear predictor is to explain the variation in the true signal. We concentrate on latent Gaussian models, where the linear predictor is composed of random model components that follow multivariate Gaussian distributions conditional on the parameters. These parameters control the random effects and need prior distributions. We set focus to the most central type of model parameters: the variance parameters. Other parameters such as correlations are outside the scope of the paper.

Parameters controlling mean and medians, such as the coefficients of fixed effects, are close to the data and tolerate vague priors (Goel and Degroot, 1981; Gelman et al., 2020), and are commonly given Gaussian priors with zero mean and a fixed high variance. We follow Fuglstad et al. (2020) and do not give attention to the fixed effect coefficients. The specification of a prior distribution for a variance parameter can be regarded as a challenge (Lambert et al., 2005; Gelman et al., 2017), but is at the same time a strong feature of Bayesian inference, as here prior knowledge, obtained from previous experiments or comparable investigations, and expert knowledge can be included to make the model more robust and more complete. So far, there is no R package that intuitively allows the proper inclusion of such knowledge and simple visualization of chosen priors in a straightforward manner. Here, we present the R package `makemyprior` that tries to fill this gap, and is a tool for increasing the awareness around prior choices.

The `makemyprior` package applies the hierarchical decomposition (HD) prior framework proposed by Fuglstad et al. (2020) and distributes the total variation

in the observed data to the random effects following a *prior tree structure*. In this tree, the leaf nodes represent the random effects specified in the linear predictor, and the root node represents the sum of the random effect variance, denoted the *total variance*. We do not include the variance of the fixed effects in neither the tree nor in the total variance, only the latent random effects and their variances. How much variation is distributed from a parent node to its child nodes is determined by a split in the tree, and this procedure continues until the leaf nodes. This gives us a parameterization with *proportions of variation*, instead of the more common variance parameter parameterization. The priors for those variance proportion parameters can be specified intuitively and transparently as they often coincide with the scale on which prior or expert knowledge exists, such as in genomic modelling (Holand et al., 2013; Hem et al., 2021) and disease mapping (Wakefield, 2006). The main goal with `makemyprior` is to raise the awareness of the need of prior distributions and to help the user to formulate, compute and visualize sensible and proper prior distributions.

`makemyprior` allows the user to specify the prior distribution either directly within R or through a graphical user interface (GUI). In the GUI the user can inspect the prior tree and adapt it as needed. Consequently, one can click through the splits independently, or be guided, to specify the beliefs for each split. The user be ignorant and distribute the variance equally to the child nodes through a Dirichlet prior, or exploit expert knowledge implemented via penalized complexity (PC, Simpson et al., 2017) priors. The PC prior is a principle-based, weakly-informative proper prior, and is used by Fuglstad et al. (2020) in the HD prior framework to formulate a joint prior where the whole model is taken into account. After completing the prior specification the package computes the joint prior, which has desired properties such as being robust and interpretable, the package allows to feed the prior directly into the R packages `rstan` and `INLA` for inference.

We begin with explaining the concepts of total variance and hierarchical variance decomposition through two motivating examples in Section 2. We then introduce the necessary background in Section 3 before we present the `makemyprior` package with general explanations on how to use it in Section 4. Section 5 gives more detailed examples showing how to use the package in various situations. A summary and discussion is given in Section 6. The package is available at https://github.com/ingebogh/makemyprior_0.1.0 with instructions on how to install it.

2 Motivating examples

We demonstrate the core ideas of total variance and hierarchical decomposition of the total variance through two illustrative examples. These examples are simplified versions of examples we use later to demonstrate the usage of the `makemyprior` package.

First, in quantitative genetics, one of the key quantities, heritability, concerns the distribution of observed variation to genetic and environmental sources. In this setting, good intuition exists on the ratio of genetic to phenotypic variation, also known as the heritability, whereas it is more difficult to define suitable priors separately on the two variance parameters.

Example 2.1 (Genomic models). *Consider a group of n individuals, where each individual i has an observed phenotype y_i . A simple genomic model is*

$$y_i = \mu + a_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where μ is an intercept, $\mathbf{a} = (a_1, \dots, a_n)^\top \sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2 \mathbf{A})$ is an additive genetic effect, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top \sim \mathcal{N}_n(0, \sigma_\varepsilon^2 \mathbf{I}_n)$ is environmental noise. The covariance matrix \mathbf{A} is calculated based on genetic sequencing of the n individuals and is scaled so that σ_a^2 is representative of the variance arising from the genetic effect; see Selle et al. (2019); Hem et al. (2021) for details.

In this simple model, two key quantities are the phenotypic variance $\sigma_{a+\varepsilon}^2 = \sigma_a^2 + \sigma_\varepsilon^2$, which is the total variance, and the heritability $\omega_{\frac{a}{a+\varepsilon}} = \sigma_a^2 / (\sigma_a^2 + \sigma_\varepsilon^2)$, which is the proportion of the phenotypic variance explained by the genetic effect. When expert knowledge is available about these two quantities, this information can be directly exploited through a joint prior assigned to phenotypic variance and heritability. A simple and intuitive visualization of this parameterization is given by the tree in Figure 1a where the phenotypic variance $\sigma_{a+\varepsilon}^2 = \sigma_a^2 + \sigma_\varepsilon^2$ in the top (root) node is distributed to the genetic variance σ_a^2 and the environmental variance σ_ε^2 in the two leaf nodes. Note that we do not include the intercept in the tree, and treat it independently with a wide Gaussian prior. See Hem et al. (2021) for a detailed description.

The idea of expressing a parameterization that is given in terms of total variance and proportions of variances through a tree, extends to more complex models with more random effects. For example, when analysing data arising from designed experiments.

Example 2.2 (Design of experiments). *Based on the ideas in Fuglstad et al. (2020), we assume that a field is split into rows and columns resulting in a 9×9*

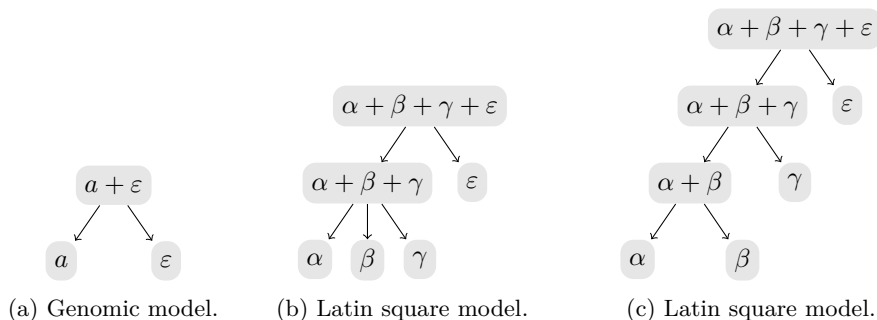


Figure 1: Prior trees for (a) the genomic model in Example 2.1, and (b, c) the latin square model in Example 2.2.

grid, and that one out of 9 different strengths of fertilizer is applied in each grid cell. Outcomes $y_{i,j}$ are observed in row i and column j under treatment $k[i, j]$, and modelled through

$$y_{i,j} = \alpha_i + \beta_j + \gamma_{k[i,j]} + \varepsilon_{i,j}, \quad i, j = 1, \dots, 9, \quad (2)$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_9)^\top \sim \mathcal{N}_9(\mathbf{0}, \sigma_\alpha^2 \mathbf{I}_9)$ is a row effect, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_9)^\top \sim \mathcal{N}_9(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_9)$ is a column effect, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_9)^\top \sim \mathcal{N}_9(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_9)$ is a treatment effect, and the residual noise is $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}, \varepsilon_{1,2}, \dots, \varepsilon_{9,9})^\top \sim \mathcal{N}_{81}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{81})$. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have sum-to-zero constraints. The individual variances, σ_α^2 , σ_β^2 , σ_γ^2 and σ_ε^2 , are nuisance parameters, and it may be difficult to have prior knowledge about them.

Figures 1b and 1c visualize two ways in which the total variance $\sigma_{\alpha+\beta+\gamma+\varepsilon}^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$ can be distributed to the individual variances σ_α^2 , σ_β^2 , σ_γ^2 and σ_ε^2 in the leaf nodes. Using Figure 1b, we could envision that, in the top split, shrinkage is applied to the latent variance $\sigma_{\alpha+\beta+\gamma}^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2$ relative to the residual variance σ_ε^2 with the goal of reducing the risk of overfitting. Then in the second split, we could envision that we want to express ignorance about how the latent variance is distributed to the individual variances σ_α^2 , σ_β^2 and σ_γ^2 . Alternatively, using Figure 1c, we may want to express ignorance about how $\sigma_{\alpha+\beta}^2 = \sigma_\alpha^2 + \sigma_\beta^2$ is distributed to σ_α^2 and σ_β^2 , but apply shrinkage to the variance of the treatment variance σ_γ^2 relative to $\sigma_{\alpha+\beta}^2$. The full details can be found in Fuglstad et al. (2020).

Examples 2.1 and 2.2 make it clear that in some cases it is natural to think in terms of proportions of variances. However, explicitly writing out how the

proportions are defined may obfuscate the key ideas that one want to express. Therefore, trees such as shown in Figure 1 are critical to define such priors. We want to emphasize that a prior can be chosen in many ways, and there are no wrong choices. Our message is that it is a choice to use the default prior, it is a choice to use literature-based priors, and it is a choice to use prior and expert knowledge in the prior. We believe it is important to communicate this and to make prior selection an active part of the Bayesian workflow.

3 Background

In this section, we present the necessary theoretical and methodical background behind the `makemyprior` package, in addition to terminology used throughout the paper.

3.1 Hierarchical variance decomposition along a prior tree

3.1.1 General definition of a prior tree

A prior tree is a directed acyclic graph consisting of a *top node*, *split nodes* and *leaf nodes*, connected by directed edges. Each random model components is represented by a leaf node. The top node in the tree represents total variance, i.e., the sum of the variance of the leaf nodes (random effects). Note that even though we may have fixed effects in the model that contribute to the total variance in the data, we follow Fuglstad et al. (2020) and omit intercept and fixed effects, in the hierarchical decomposition (HD) prior and thus also in the tree, and the total variance refers to the sum of the variance of the random effects. We combine two or more leaf nodes in a split node in a way that reflects the hierarchical model and our prior beliefs about how the total variance is distributed. The split nodes represent a *variance proportion*. Note that a top node is also a split node.

In some cases we might want to include only a subset of the model components in the same prior tree, and in that way have several trees. An example is the animal model (e.g., Holand et al., 2013), a mixed model that is usually used to decompose environmental and genetic variances in animal populations. We may have a good intuition on the absolute magnitude of the variation that is explained by the environment and by all the genetic contributions separately, but the latter might be split into several sub-components, like additive, dominance and mutational variance where we only have intuition on the relative magnitude of the variation. The genetic contributions are confounded, thus it is useful to split the variance of the genetic contribution using a prior tree, and in that way have

both variance parameters, total variance parameters, and variance proportions. If a variance component is assigned an individual prior, its corresponding random effect is represented by a *singleton*: a node not connected to any other nodes. The singletons can be considered as trees with only one node, and no variance to distribute. Several prior trees gives a *prior forest*. We refer to the forest of trees as the *prior tree structure* of the model. Each tree in a tree structure is associated with their own joint prior distribution, independent of the priors belonging to the other trees. The tree structure is made based on prior knowledge about the model, data and problem at hand.

3.1.2 Defining priors for the split nodes: Shrinkage versus ignorance

Given a tree structure describing the distribution of the variance in the model, we can use the rest of our pre-existing knowledge to steer the variance to the different model components by choosing suitable priors for the different parameters belonging to the prior tree.

A good prior distribution can improve the robustness of the inference, by helping to avoid estimating spurious effects. We apply penalized complexity (PC) priors as they shrink towards a so-called base model and are thus robust by design. PC priors are based on the *distance* $d(\cdot)$ between the base model and a flexible extension of the base model, measured using the Kullback-Leibler divergence (Simpson et al., 2017). In the base model the parameter of interest θ is fixed to θ_0 . An example of a base model for a random effect with mean zero and one variance parameter is to fix this variance parameter $\theta = \sigma^2$ to zero, which corresponds to removing the effect from the model. In the flexible model on the other hand, θ is allowed to vary. The PC prior induces shrinkage towards the base model, which gives a robust prior that aids to avoid overfitting. As in Simpson et al. (2017), we use an exponential prior for the distance $d(\cdot)$ between the two models, and transform this to a prior on the desired parameter. This means that the PC prior always is an exponential distribution on the distance, but for the parameter in question θ the distribution varies with parameterization, choice of base model and covariance matrix structures, and does in general not have an analytical expression. The parameter θ can be a variance or standard deviation (Simpson et al., 2017), a variance proportion (Fuglstad et al., 2020; Hem et al., 2021), or, for example, a correlation parameter (Guo et al., 2017). In *makemyprior* we consider standard deviation σ (and variance σ^2) parameters together with variance proportions ω . For a standard deviation the distance is simply $d(\sigma) = \sigma$ (Simpson et al., 2017). For a variance proportion parameter the distance will be a function of the covariance matrices of the random effects involved in the split, see Fuglstad et al. (2020) for details.

When constructing the joint prior, a bottom-up approach following the prior tree is used, and the prior will thus be dependent on prior tree structure. The distance measure $d(\cdot)$ for a variance proportion depends on the covariance matrices of the effects of the child nodes in a split (Fuglstad et al., 2020), and the covariance matrix of a split node will be a function of the variance proportions(s) involved in this split. This means we must condition on the variance proportions associated with splits lower in the tree (if any), and that each prior depends on choices and covariance matrices at that and lower levels. We omit the dependence of tree structure, covariance matrices and prior choices for other splits in the notation of the PC prior for readability. Consider a random intercept model $y_{i,j} = a_i + \varepsilon_{i,j}$ for $i, j = 1, \dots, 10$, where $a_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_a^2)$ is a group effect and $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ is a residual effect. We define the variance proportion $\omega_{\frac{a}{a+\varepsilon}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\varepsilon^2}$. Then we denote the different PC prior distributions as:

- $\sigma_* \sim \text{PC}_0(U, \alpha)$, with $\text{Prob}(\sigma_* > U) = \alpha$, and shrinkage towards $\sigma_* = 0$.
- $\omega_{\frac{a}{a+\varepsilon}} \sim \text{PC}_0(m)$ with $\text{Prob}(\omega_{\frac{a}{a+\varepsilon}} > m) = 0.5$ so that m defines the median, and shrinkage towards $\omega_{\frac{a}{a+\varepsilon}} = 0$, i.e., the base model is a model with only ε .
- $\omega_{\frac{a}{a+\varepsilon}} \sim \text{PC}_1(m)$ with $\text{Prob}(\omega_{\frac{a}{a+\varepsilon}} > m) = 0.5$ so that m defines the median, and shrinkage towards $\omega_{\frac{a}{a+\varepsilon}} = 1$, i.e., the base model is a model with only \mathbf{a} .
- $\omega_{\frac{a}{a+\varepsilon}} \sim \text{PC}_M(m, c)$ with $\text{Prob}(\omega_{\frac{a}{a+\varepsilon}} > m) = 0.5$ and $\text{Prob}(\text{logit}(1/4) < \text{logit}(\omega_{\frac{a}{a+\varepsilon}}) - \text{logit}(m) < \text{logit}(3/4)) = c$ so that m defines the median, and c says something about how concentrated the distribution is around the median. The shrinkage is towards $\omega_{\frac{a}{a+\varepsilon}} = m$, i.e., the base model is a combination of the effects \mathbf{a} and ε .

Note that $\text{PC}_1(m)$ on $\omega_{\frac{a}{a+\varepsilon}}$ is equivalent to $\text{PC}_0(1 - m)$ on $1 - \omega_{\frac{a}{a+\varepsilon}} = \omega_{\frac{\varepsilon}{a+\varepsilon}}$. Since the PC prior is a prior put on the distance between two models, and then transformed to the parameter of interest, we do not distinguish the notation between the PC prior on a standard deviation and variance parameter, as it will result in the same prior. In Figure 2 we show examples for the four different priors. Note that the shape of the PC prior on a standard deviation, shown in Figure 2a, is independent of the hyperparameters and other hyperparameters will simply give a rescaling of the axes.

If the covariance matrix of the base model is non-singular, we get a PC prior where the median is guaranteed to be where we choose it to be, as in Figure 2b. However, if the base model is chosen so that the covariance matrix is singular,

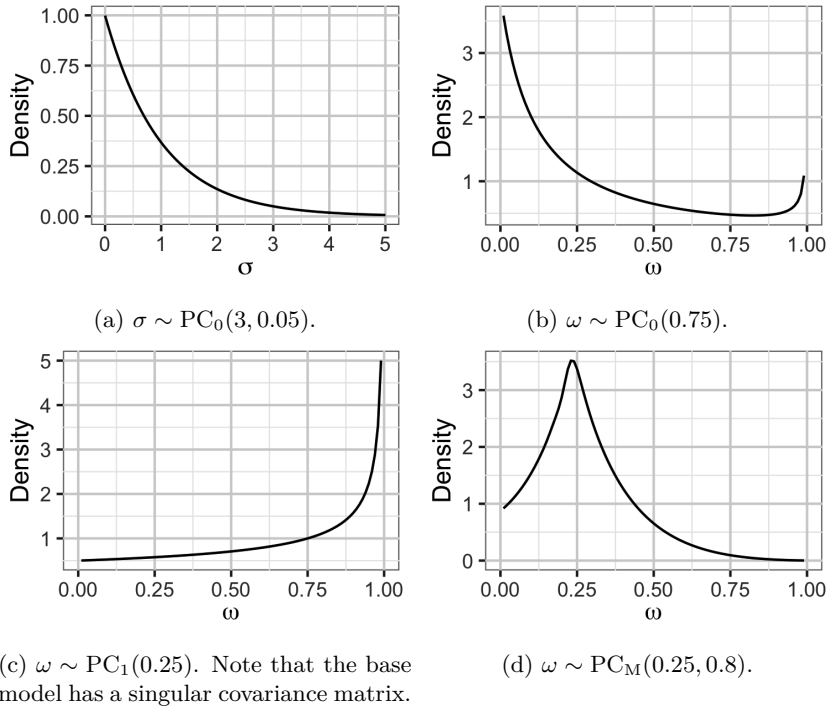


Figure 2: Examples of the different PC priors for a random intercept model $y_{i,j} = a_i + \varepsilon_{i,j}$ for $i, j = 1, \dots, 10$. $\omega = \sigma_a^2 / (\sigma_a^2 + \sigma_\varepsilon^2)$.

we get a distance measure that is infinite, and we cannot have a median that is further than 0.25 from the base model, which is the case for the prior in Figure 2c (see Fuglstad et al. (2020, Theorem 1) for details). The base model covariance matrix is always non-singular for a PC_M prior (shown in Figure 2d). The concentration parameter c in $\text{PC}_M(m, c)$ measures how certain we are about the prior median, and can be between 0.5 and 1. Fuglstad et al. (2020) suggest $c = 0.5$. A concentration of less than 0.5 will indicate that a prior with median at $\omega = 0.5$ is less concentrated around 0.5 than a uniform distribution on $[0, 1]$ would be, and thus we set the lower limit to 0.5. As this parameter says something about how much of the distribution mass is in an interval that is smaller than the parameter space ($0 \leq \omega \leq 1$), the distribution will not change much when c approaches 1.

The difference between the PC_0/PC_1 and PC_M priors is how certain the user

is in the prior knowledge. In the genomic model in Example 2.1, assume we have knowledge about the heritability from similar experiments, and that, for this phenotype and species, it is around 0.4. If we are certain that the additive genetic effect is contributing to the variation in the observed phenotype, we use a PC_M prior with median $m = 0.4$, and we choose the concentration parameter c based on how strongly we believe in this value. In this case we get a prior with shrinkage towards a heritability of 0.4. If the contribution of the additive effect to the total variation is not clear, for example due to a small data sample, we can use a PC_0 prior with median $m = 0.4$, which results in a prior with shrinkage towards the heritability being 0.

By using a *multi-split*, a split with more than two child nodes, the user expresses no strong opinions on how the variance is distributed among the components involved in the split, and we follow Fuglstad et al. (2020) and use the ignorant symmetric Dirichlet prior on such splits. For a split with p children this prior is given by:

$$\pi(\boldsymbol{\omega}, p) = \text{Dirichlet}(p) = \frac{\Gamma(p\alpha)}{\Gamma(\alpha)^p} \left(\prod_{i=1}^p \omega_i \right)^{\alpha-1},$$

where $\boldsymbol{\omega}$ is the vector of variance proportions involved in the split with $\omega_i > 0$ for $i = 1, \dots, p$ and $\sum_{i=1}^p \omega_i = 1$. Further, $\Gamma(\cdot)$ denoted the gamma function and $\alpha > 0$ is chosen so $\text{P}(\text{logit}(1/4) < \text{logit}(\omega_i) - \text{logit}(1/p) < \text{logit}(3/4)) = 1/2$ for $i = 1, \dots, p$ (if this is achieved for one i , it is by symmetry achieved for all i). This prior assigns equal amount of variance to each model component. For a split consisting of p nodes we denote this as $\boldsymbol{\omega} \sim \text{Dirichlet}(p)$ for each proportion in the split. Note that a $\text{Dirichlet}(2)$ prior can be used on a dual split where the user wants to be ignorant about the attribution of variance to the two child nodes. The Dirichlet distribution for a dual split reduces to a uniform distribution.

There may be situations where the user wants to assign unequal amounts of variance to the components in a multi-split. In that case, the user has opinions about the variance decomposition in the split, and should thus not use the ignorant Dirichlet prior. Instead we transform the multi-split to several dual splits and use a PC prior on each of the dual splits. In Example 2.2, assume that we want a (20, 30, 50) division of the latent total variance of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. This we achieve by first splitting the variance 50/50 between $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ with a $\text{PC}_M(0.5, c)$ prior, and then dividing the variance of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ 40/60 with a $\text{PC}_M(0.4, c)$ prior for some suitable value of c , for example $c = 0.5$. We show the corresponding tree structure in Figure 1c. We could also have chosen another effect to split off first, but Fuglstad et al. (2020) show that this order does not have much impact on the resulting prior.

To ease the computation of the joint PC prior, we follow Fuglstad et al. (2020) and condition on the base models of the splits below, instead of on the parameters themselves. In this way we can pre-compute the marginal priors for each split in the tree. The base model for the Dirichlet distribution is equal variance to each component, i.e., for a split with p components the base model is $1/p$ of the variance in the split node to each child node. A split where with a Dirichlet prior do not use information from lower levels in the tree.

3.1.3 Defining priors for top nodes and singletons

Appropriate prior distributions for variance parameters varies with the likelihood. We describe the train of thought when specifying prior distributions for variance parameters when the likelihood is Gaussian, binomial and Poisson.

In a model with a Gaussian likelihood, the total variance is usually easy to identify and does not need an informative prior. This is a parameter close to the data in the model, just as the intercept, and settles with a vague prior (Gelman et al., 2020). For the total variance in the top nodes, Fuglstad et al. (2020) recommend the scale-invariant, improper Jeffreys' prior when all model components are involved in one single tree. This prior does not require any hyperparameters and is straight-forward to use. In cases where the user has specific knowledge about the total model variance, a proper prior can be used to include this knowledge in the prior. In cases where the nodes are not all involved in the same tree, a scale-invariant prior is not meaningful, and an improper prior may lead to an improper posterior. Fuglstad et al. (2020) recommend the PC prior, and so do we due to its desirable shrinking properties, but any prior distribution suitable for variance parameters is applicable. The hyperparameters can be selected using a tail probability of the standard deviation, and we recommend a weak prior. Singletons must always be assigned a proper prior.

For the genomic model in Example 2.1, assume we have prior knowledge saying it is unlikely that the total standard deviation in the observed data is greater than 4. We want to use this knowledge, and choose a $PC_0(U, \alpha)$ prior with $U = \sqrt{4}$. The value we choose for α says something about how certain we are in the value of U . $\alpha = 0.05$ is in this case a suitable choice.

Other likelihoods require proper priors on all variances, also the total variance, and scale-invariance is not meaningful for data that are not Gaussian. Again we follow the recommendation of Fuglstad et al. (2020) and suggest PC priors. However, instead of choosing a prior using the upper tail probability of the standard deviation, we think on a different scale than for Gaussian data, and choose a credible interval for the variance parameter on a suitable scale that we

transform into an upper tail probability. For both binomial and Poisson data, an exponential scale with logit and log links, respectively, is appropriate. This will correspond to thinking on odds-ratio scale for binomial data, and on the scale of the data for Poisson data. This is a bit less intuitive than for the Gaussian likelihood, however, the interpretation of each single variance parameter is not straight-forward either, and with the HD prior framework the user only needs to consider one instead of multiple variance parameters.

Assume we have a model with linear predictor $\eta_i = \mu + a_i + b_i$ and binomial likelihood with logit link function. An intuitive way of choosing a prior for the total variance is to choose an equal-tailed credible interval for the effect of the random effects on the odds-ratio, $\exp(a_i + b_i)$, i.e., $\text{Prob}(l < \exp(a_i + b_i) < u) = p$ (Fong et al., 2010). For example, we can say we want $\text{Prob}(0.1 < \exp(a_i + b_i) < 10) = 0.9$. This corresponds to a 90% credible interval $[0.1, 10]$ for $\exp(a_i + b_i)$. The idea is the same for a Poisson likelihood with a log link: we can think on the effect of the random effects on the relative risk. Note that we do not include fixed effects when we choose parameters for the prior.

4 Software

Throughout this section, the use of the package is exemplified by the following model.

Model 1 (Example model). Consider the hierarchical model for the $n = m \cdot p$ observations $y_{i,j}$, $i = 1, \dots, p$ and $j = 1, \dots, m$, given by

$$y_{i,j} | \eta_{i,j}, \sigma_\varepsilon^2 \sim \mathcal{N}(\eta_{i,j}, \sigma_\varepsilon^2),$$

$$\eta_{i,j} = \mu + x_i \beta + a_i + b_j,$$

where μ is an intercept and x_i is a covariate with coefficient β . $a_1, a_2, \dots, a_p \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_a^2)$ and $b_1, b_2, \dots, b_m \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_b^2)$ are random effects, and $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ are residuals.

Several likelihoods, latent models and prior distributions are available in `makemyprior`, which can be listed with the function:

```
makemyprior_models(type = c("prior", "latent", "likelihood"),
  select = NULL)
```

4.1 Specifying the linear predictor and preparing the data object

First, the linear predictor is specified using a formula object with a syntax similar to e.g. `lm()` and `inla()`. Covariates are included directly by name in the formula, and `mc()` is used to include information about each random effect:

```
mc(label, model = "iid", ...)
```

The main arguments are `label` (name of the effect) and `model` (the type of latent model effect, i.i.d. is the default). The other arguments depend on the choice of latent model, see documentation for details.

For Model 1 where both \mathbf{a} and \mathbf{b} are i.i.d., the formula is:

```
> formula <- y ~ x + mc(a) + mc(b)
```

The intercept μ is included by default, but can be removed with `-1`. The residual effect for a Gaussian likelihood is not specified in the formula.

The next step is to gather data and create a data object as either a `data.frame` or `list` with names corresponding to the elements of the formula. For Model 1, we need the observed response y , the covariate x , and indexes for \mathbf{a} and \mathbf{b} (these must be specified as integers). A simple simulated dataset is:

```
> p <- 10
> m <- 10
> n <- m*p
>
> set.seed(1)
> data <- list(a = rep(1:p, each = m),
+           b = rep(1:m, times = p),
+           x = runif(n))
> data$y <- data$x + rnorm(p, 0, 0.5)[data$a] +
+   rnorm(m, 0, 0.3)[data$b] + rnorm(n, 0, 1)
```

We recommend the use of short names for the input data because these names need to be used to refer to model components in later steps. An example is "rain" instead of "rainfall_august_2020". Note that the observations $y_{i,j}$ are not used to make the prior.

	Tree structure	Text string
Prior 1		"(a); (b); (eps)"
Prior 2		"s1 = (a, b); (eps)"
Prior 3		"s1 = (a, b, eps)"
Prior 4		"s1 = (a, b); s2 = (s1, eps)"

Table 1: Four tree structures for Model 1 with text strings specifying them. s_1 and s_2 are names on the splits (variance proportions) chosen by the user in the initial specification of the prior, and are used in a nested formulation to specify the prior tree structure.

4.2 Exploring and selecting the prior graphically

We provide a graphical user interface (GUI) where the user can construct the prior in an interactive way. The graphical user interface is implemented as a **shiny** (Chang et al., 2020) app running locally and allows the user to first define the tree structure, and then be guided sequentially through the steps of selecting priors for each singleton, split and total variance. The interface supports the use of prior forests (multiple trees) such as Priors 1 and 2 in Table 1. **shiny** apps are useful to display and investigate results from analyses. For example, the user can customize graphs and tables in a simple way. **shiny** apps are used by Depaoli et al. (2020) to show why prior sensitivity analysis is important, and by Smid and Winter (2020) to let users explore the impact of prior distributions in inference. However, to the extent of our knowledge, as of today there are no packages or apps that allows the user to use a **shiny** app (or similar) to specify priors for

custom models and data and directly carry out inference.

The first step is to initialize a prior object. This is done with the function `make_prior()`:

```
make_prior(formula, data, family = "gaussian", prior = list(),
           intercept_prior = c(), covariate_prior = list())
```

`formula` and `data` are the objects created in Section 4.1, and `family` is the likelihood (Gaussian likelihood is the default; binomial and Poisson likelihoods are also available). The `prior` argument is not relevant when the GUI is used, and its description is deferred to Section 4.3. `intercept_prior` and `covariate_prior` specifies the parameters for the Gaussian priors on the intercept and covariate coefficients. The default in `makemyprior` is $\mathcal{N}(0, 1000^2)$ for both, and the coefficient priors are specified as a named list with names corresponding to the covariate names.

For Model 1, we can create a prior object with the following command:

```
> prior <- make_prior(formula, data, family = "gaussian",
+                    intercept_prior = c(0, 1000),
+                    covariate_prior = list(x = c(0, 100)))
```

Warning message:

```
Did not find a tree, using default tree structure instead.
```

This gives a prior with a single tree with one split as shown in Prior 3 in Table 1, which is the default setting. We get a warning, to make the user aware that the default prior is chosen. For Gaussian likelihoods, a Jeffreys' prior is set for the total variance. Let $\sigma_{a+b+\varepsilon}^2$ denote total variance, and $\omega = (\omega_{\frac{a}{a+b+\varepsilon}}, \omega_{\frac{b}{a+b+\varepsilon}}, 1 - \omega_{\frac{a}{a+b+\varepsilon}} - \omega_{\frac{b}{a+b+\varepsilon}})$ describe the attribution of variance to the three different sources, then the initial choice of priors is

$$\omega \sim \text{Dirichlet}(3) \quad \text{and} \quad \sigma_{a+b+\varepsilon}^2 \sim \text{Jeffreys}' \quad (3)$$

The intercept has a $\mathcal{N}(0, 1000^2)$ prior, and the covariate a $\mathcal{N}(0, 100^2)$ prior.

The function `makemyprior_gui()` allows the user to select the desired prior tree structure and choose prior distributions interactively:

```
makemyprior_gui(prior, guide = FALSE, no_pc = FALSE)
```

This function takes the arguments `prior`, which was created with `make_prior()` earlier, `guide`, which is a boolean that specifies whether or not the guide should

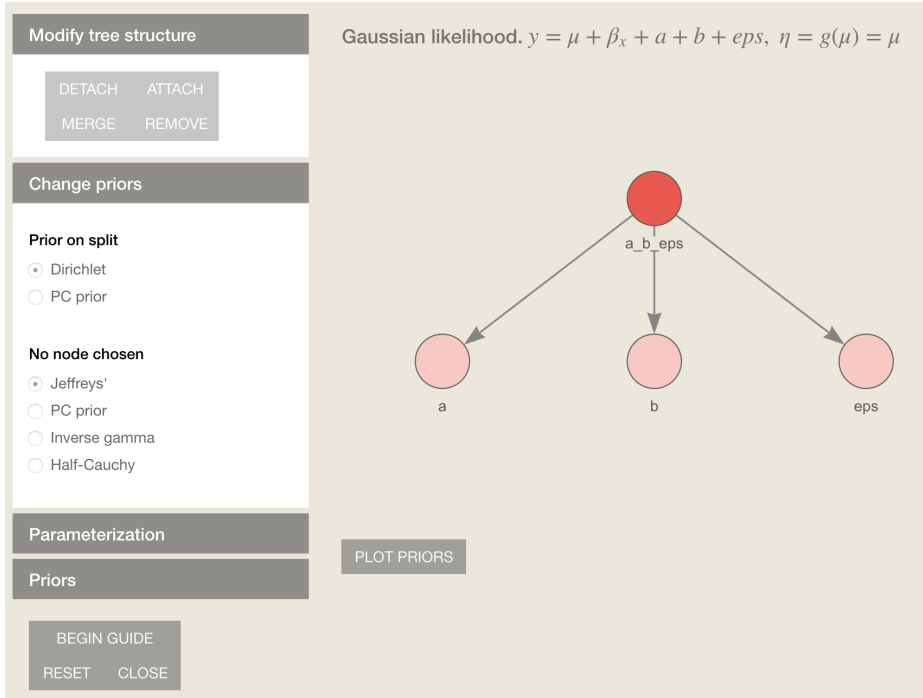


Figure 3: Screenshot of the GUI in `makemyprior` with the default prior (Equation 3).

automatically start (the guide can be started at any time), and `no_pc`. Since the PC prior is computed using the covariance matrix structure of the model components, it may be slow for large models. For a better user experience, the user can turn off the computation of the PC prior in the GUI using `no_pc = TRUE`. The prior will be computed upon closing, and this will only affect the plotting of the prior in the GUI.

For Model 1, we start the GUI by running:

```
> new_prior <- makemyprior_gui(prior)
```

This saves the changes made in the graphical interface to the variable `new_prior`. Figure 3 shows a screenshot of the GUI with Model 1 for $m = p = 10$ and the default prior. The user can create the desired tree structure, and then choose the desired priors for this tree. Note that every time the tree structure is modified, all splits are set to have the default Dirichlet prior. Figure 4 shows a screenshot

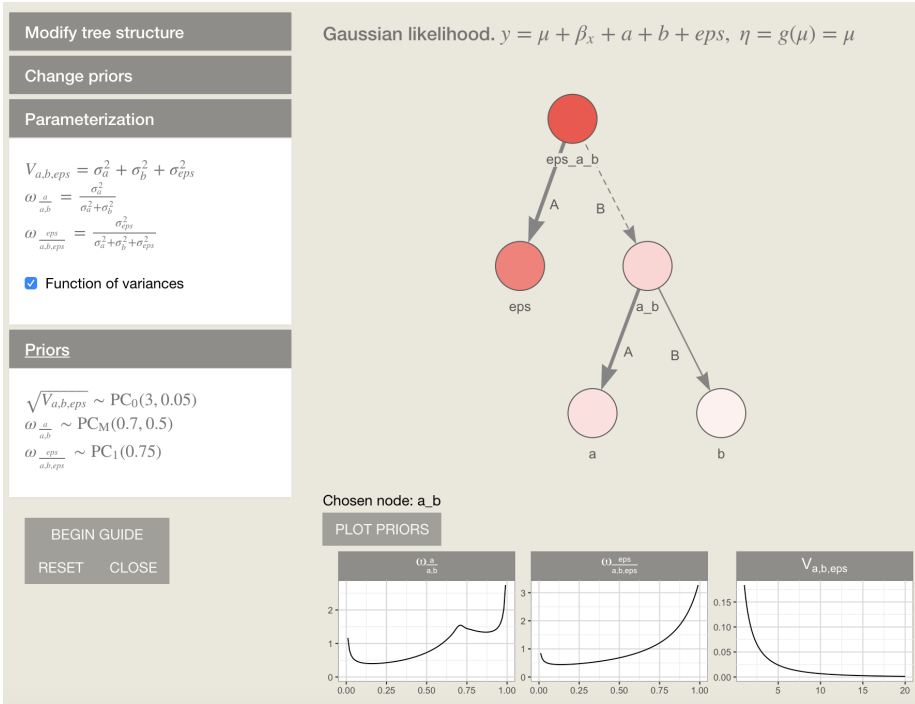


Figure 4: Screenshot of the GUI in **makemyprior** when the prior is chosen as in Equation 4.

of the GUI where we have chosen the tree in Prior 4 in Table 1 and the following distributions:

$$\omega_{\frac{a}{a+b}} \sim PC_M(0.7, 0.5), \omega_{\frac{eps}{a+b+\varepsilon}} \sim PC_0(0.25), \text{ and } \sigma_{a+b+\varepsilon} \sim PC_0(3, 0.05). \quad (4)$$

The chosen prior distributions and the connection between the parameterization and model variances are easily seen in the GUI. Figure 5 shows the initial page of the guide, where the user is asked about the tree structure, and then guides them through how to choose priors based on prior knowledge for each split, total variance and singleton through simple questions.

The GUI is intuitive and contains a thorough description of the options, and we do not explain the features in detail here. We instead recommend to use the guide in the GUI, either by running `makemyprior_gui()` with `guide = TRUE`, or clicking “Begin guide” inside the GUI itself. A summary of the prior object can be printed with:

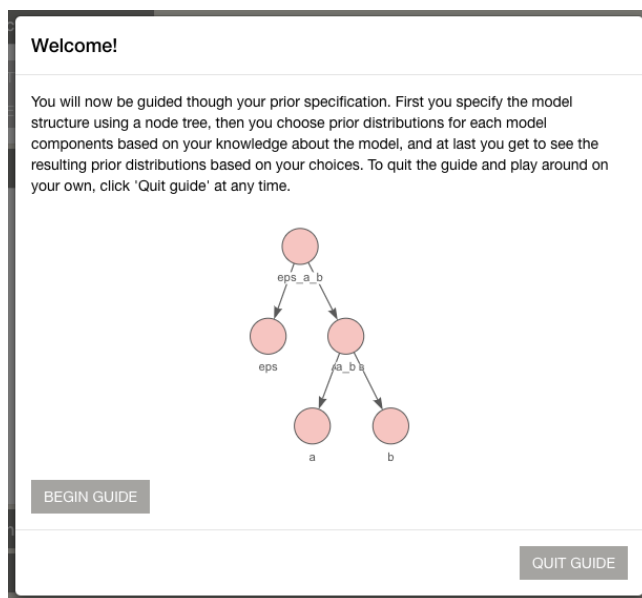


Figure 5: Screenshot of the guide in makemyprior.

```
> summary(new_prior)
```

```
Tree structure: a_b = (a,b); eps_a_b = (eps,a_b)
```

```
Weight priors:
```

```
  w[a/a_b] ~ PCM(0.7, 0.5)
```

```
  w[eps/eps_a_b] ~ PC1(0.75)
```

```
Total variance priors:
```

```
  sqrt(V)[eps_a_b] ~ PC0(3, 0.05)
```

```
Covariate priors: intercept ~ N(0, 1000^2), x ~ N(0, 100^2)
```

```
y ~ x + mc(a) + mc(b)
```

4.3 Selecting the prior non-graphically

If the prior has been constructed with the GUI, this section can be skipped. Here we describe an alternative way to specify the prior without the use of a GUI.

This is done with same function we use to make the default prior, `make_prior()`, through the `prior` argument. This is a named `list` with the following arguments:

- tree** The tree structure as a string. A split is specified as `s1 = (a, b)`, where `s1` represents a split node and can be any name except names of the input data in `data` and the reserved `eps`, which is used for residuals for a Gaussian likelihood. Short names are recommended. Note that these split names are just used in the initial specification. The child nodes for each split are included in parentheses separated by commas, and each split is separated by semicolons. Singletons are included as `(a)`. Examples of strings for different tree structures for Model 1 are shown in Table 1.
- V** A named list with information on the priors on each top node and singleton, i.e., all variances. Options are "pc", "jeffreys", "invgam", and "hc" (Half-Cauchy). The names in the list are the top node and singleton names from the `tree` argument.
- w** A named list with information on the priors on each split, i.e., all variance proportions. The names in the list are the split node names from the `tree` argument. Options are "pc0", "pc1", "pcM" and "dirichlet".

`V` and `w` must have the following structure for each element in the list: `list(prior = prior_name, param = parameter_vector)`, except for Jeffreys' and Dirichlet priors, where `param` should not be specified (as the distributions do not have hyperparameters). See `makemyprior_models()` for details, and see Section 5 for examples on how to specify priors in specific examples.

The prior specified in Equation 4 for Model 1 can be specified as

```
R> prior <- make_prior(
+   formula, data,
+   prior = list(
+     tree = "s1 = (a, b); s2 = (s1, eps)",
+     w = list(s1 = list(prior = "pcM", param = c(0.7, 0.5)),
+              s2 = list(prior = "pc0", param = 0.25)),
+     V = list(s2 = list(prior = "pc0", param = c(3, 0.05)))
+   ),
+   covariate_prior = list(x = c(0, 100)))
R> prior
```

Model: $y \sim x + mc(a) + mc(b)$

Tree structure: `a_b = (a,b); eps_a_b = (eps,a_b)`

Weight priors:

`w[a/a_b] ~ PCM(0.7, 0.5)`

`w[eps/eps_a_b] ~ PC1(0.75)`

Total variance priors:

`sqrt(V)[eps_a_b] ~ PC0(3, 0.05)`

Covariate priors: `intercept ~ N(0, 1000^2)`, `x ~ N(0, 100^2)`

The names `s1` and `s2` are chosen by the user in the initial specification of the prior, and can be any names that are not used for the data or `eps` which is reserved for residuals. `s1` and `s2` will be changed automatically by `make_prior()`, and are only used as a link between the splits and priors. The order we list the children for each split node in the `tree` argument decides which way we have shrinkage with the PC priors. For `s1 = (a, b)`, $PC_0(m)$ shrinks effect \mathbf{a} ($\omega_{\frac{a}{a+b}} = 0$ as base model), $PC_1(m)$ shrinks effect \mathbf{b} and $PC_M(m, c)$ gives shrinkage towards $m\mathbf{a} + (1 - m)\mathbf{b}$. All three has median at $\omega_{\frac{a}{a+b}} = m$. Note that $PC_0(m)$ on $\omega_{\frac{a}{a+b}}$ is equivalent to $PC_1(1 - m)$ on $\omega_{\frac{b}{a+b}}$. Note that `s1` and `s2` have been changed to `a_b` and `eps_a_b` by the function.

All top nodes, singletons and split nodes without a specified prior will get the default prior. The default settings in `makemyprior` are chosen based on the findings of Fuglstad et al. (2020) to ensure robust inference:

- If no prior is specified (neither tree structure nor priors), the prior will be a joint prior where all latent components (including a possible residual effect) get an equal amount of the total variance in the prior through the symmetric Dirichlet prior, and the default total variance prior.
- The default prior on the total variance (top nodes) varies with likelihood:
 - Jeffreys' prior for Gaussian likelihood for a tree structure with one tree, $PC_0(3, 0.05)$ otherwise.
 - $PC_0(1.6, 0.05)$ for binomial likelihood.
 - $PC_0(1.6, 0.05)$ for Poisson likelihood.
- The default prior on an individual variance (singletons) varies with likelihood:
 - $PC_0(3, 0.05)$ for Gaussian likelihood.
 - $PC_0(1.6, 0.05)$ for binomial likelihood.
 - $PC_0(1.6, 0.05)$ for Poisson likelihood.

- The default prior on a variance proportion (split node) is a Dirichlet prior assigning equal amount of variance to each of the model components involved in the split.

The reasoning behind these choices are as follows. For the variance proportions, it can do more harm than it will be helpful to use a PC prior, and the ignorant Dirichlet is chosen as the default. A standard Gaussian distribution with mean 0 and variance 1 will have close to all the density mass between -3 and 3 . To choose the default variance prior for other likelihoods, we have followed the idea of Fong et al. (2010), and use a credible interval on some suitable scale. The default prior for all variance parameters (both for top nodes and singletons) in `makemyprior` for binomial and Poisson likelihoods is a PC prior with a 95% credible interval between 0.2 and 5 for the multiplicative effect on the odds ratio and risk, respectively. This is obtained with a $PC_0(1.6, 0.05)$ prior. We want to emphasize that before selecting the default prior, both when using `makemyprior` and otherwise, you should stop and think about whether or not it is suitable for your model and data.

4.4 Performing inference

We include functions for inference that are compatible with the prior object obtained from `make_prior` (and `makemyprior_gui`). Both Stan (Carpenter et al., 2017) through `rstan` (Stan Development Team, 2020) and Integrated Nested Laplace Approximations (INLA, Rue et al., 2009) through INLA (see www.r-inla.org) can be used for the inference.

Stan is a probabilistic programming language, where Hamiltonian Monte Carlo (HMC) is used to sample from the posterior distribution (Carpenter et al., 2017). Stan implements HMC using the No U-Turn Sampler (NUTS, Hoffman and Gelman, 2014). NUTS reduces the need for tuning of the sampler, making it easy to use as no manual settings are needed for the algorithm to run, and the user only needs to provide the joint prior and likelihood model, implemented in a programming language similar to C++. We provide pre-compiled Stan-code for fitting latent Gaussian models with certain likelihoods and latent effects. The internal parameterization in the provided Stan-code is log-variance, and the prior is transformed from the parameterization given by the prior tree structure to log-variances.

INLA is a non-sampling based method for doing fast and efficient Bayesian inference on latent Gaussian models (Rue et al., 2009), utilizing Gaussian Markov random fields (GMRFs) with sparse precision matrices, which gives computational benefits due to the Markov property. The INLA method approximates the

posteriors by a mixture of Gaussian distributions and applies a skewness correction to the marginals (Rue et al., 2017). It is easy and straight-forward to use for inference, and can fit models with a broad range of latent effects. The internal parameterization of the model parameters in INLA is log-precision, and in the same way as for the provided Stan-code the parameterization following the prior tree is transformed to fit INLA.

Some common latent models are included in the code for the package: i.i.d. ("iid"), Besag ("besag"), random walk of first ("rw1") and second ("rw2") order, and effects with structured covariance matrices ("generic0").

In Section 5 we show how the inference can be performed. Here we describe the functions that can be used for inference.

4.4.1 Inference with Stan

Stan is a flexible tool for inference, however, it requires the user to write custom code for the model that is to be fitted. `makemyprior` contains pre-written Stan-code that can be used to do inference on latent Gaussian models with a selection of latent models. We recommend to compile the Stan-code before doing inference with Stan. This can be done with the following function:

```
compile_stan(save = FALSE, path = NULL)
```

where `save` indicates whether or not to save the compiled object (must be set this to `TRUE` to avoid recompiling the code every time inference is performed). `path` is only necessary if the default location for saving the compiled object is not possible to use (see the documentation for details). For inference with Stan we use the following function:

```
inference_stan(prior_obj, use_likelihood = TRUE,
               print_prior = TRUE, ...)
```

The first argument is the prior object from `make_prior` or `makemyprior_gui`. The user can specify whether to include the likelihood (`use_likelihood = TRUE`) or not (`use_likelihood = FALSE`); in the latter case we sample from the prior distribution. `print_prior` (`TRUE` by default) prints details about the chosen prior. Additional arguments that is sent directly to the `rstan` function `sampling()` can be specified for the inference. Useful arguments include `iter` (number of iterations for each chain), `warmup` (number of iterations for the warm-up), `chains` (number of chains), `seed` (for reproducibility), and `control` (for specifying algorithm tuning parameters).

The internal parameterization in the Stan-code included in the package is log-variance, however, since Stan works with samples we can look at any parameterization we want by transforming the log-variances. For using other latent models or more complex models than the ones provided in the included Stan-code (see above), the user must write customized Stan-code, see Section 4.6.

4.4.2 Inference with INLA

For inference with INLA we use the following function:

```
inference_inla(prior_obj, use_likelihoood = TRUE,
              print_prior = TRUE, ...)
```

The first three arguments are the same as in `inference_stan()`. Additional arguments can be fed to the INLA function `inla()`. Useful arguments include `Ntrials` for the binomial likelihood, used to specify the amount of trials, where the response is the number of successes.

4.5 Visualizing priors and posteriors

We offer several functions to visualize the prior and posterior distributions. The prior distributions for the random effects on the tree structure parameterization can be plotted with `plot_prior(obj)` which take an object from `make_prior()`, `makemyprior_gui()`, `inference_stan()` or `inference_inla()` as input.

The posterior distributions can be displayed with

```
plot_posterior_variance(obj)
plot_posterior_stdev(obj)
plot_posterior_precision(obj)
```

`obj` is an object from `inference_stan()` or `inference_inla()`.

The posterior distributions of random effects from inference with Stan can be plotted with:

```
plot_posterior_stan(
  obj, param = c("prior", "variance", "stdev", "precision"),
  prior = FALSE
)
```

`obj` is an object from `inference.stan()`, `param` specifies which parameterization the plots should have where `param = "prior"` gives the posterior on the same parameterization as the prior. `prior` indicates whether or not to plot the prior together with the posterior for `param = "prior"`. The total variance prior will only be plotted if it is not Jeffreys' prior. Fixed effect posteriors can be plotted with `plot.fixed.posterior(obj)`.

4.6 More complex models in Stan

Latent models may have parameters that are not variances, such as correlations. These non-variance parameters are handled independently, and are not included in the HD prior (Fuglstad et al., 2020). The Stan code included in `makemyprior` is only applicable for certain latent models and likelihoods (see Section 4.4). We provide a “skeleton” code and a description on how the user can write custom Stan-code and include the joint prior created with `make_prior()`. This can be accessed with:

```
R> create_stan_file(location = "")
```

`location` is a string to a path where a folder with necessary files will be stored. The user can edit the code and include custom latent components etc. We do not include details on this, as it will be highly model specific and is merely an offer to the users who want to apply the HD prior in more advanced models.

5 Using makemyprior: Examples

In this section we provide three examples where we use the `makemyprior` package to construct priors and run inference. Two examples are with Gaussian responses, and one with Binomial responses. We have used Stan for the inference (with `inference.stan()`), but the procedure is the same for inference with INLA (using `inference.inla()` instead).

5.1 Gaussian responses

5.1.1 Genomic selection in wheat breeding

This is an extended version of the model in Example 2.1. In addition to the additive genetic effect \mathbf{a} , we now also include two nonadditive effects: dominance

\mathbf{d} and additive-by-additive epistasis \mathbf{x} . This example is taken from Hem et al. (2021). The response y_i is grain yield. We only consider how to utilize the expert knowledge elicited from experts in the field, and create a prior distribution reflecting this knowledge. We model the response as

$$y_i = \mu + a_i + d_i + x_i + \varepsilon_i, \quad i = 1, \dots, 100 \quad (5)$$

where μ is an intercept with default $\mathcal{N}(0, 1000^2)$ prior and ε_i is the residual effect, representing environmental noise. a_i , d_i and x_i are additive, dominance and epistasis (additive-by-additive epistasis) effects, respectively. These three add up to the genetic effect $g_i = a_i + d_i + x_i$. We assume that $\mathbf{a} = (a_1, \dots, a_{100}) \sim \mathcal{N}_{100}(\mathbf{0}, \sigma_a^2 \mathbf{A})$, $\mathbf{d} = (d_1, \dots, d_{100}) \sim \mathcal{N}_{100}(\mathbf{0}, \sigma_d^2 \mathbf{D})$, and $\mathbf{x} = (x_1, \dots, x_{100}) \sim \mathcal{N}_{100}(\mathbf{0}, \sigma_x^2 \mathbf{X})$, and we use a sum-to-zero constraint on all genetic effects. The covariance matrices \mathbf{A} , \mathbf{D} and \mathbf{X} are computed from the single nucleotide polymorphism (SNP) matrix with thousands of genetic markers, see Hem et al. (2021) for details. This model has structured covariance matrices, and we use the "generic0" latent model. This requires the argument `Cmatrix`, which is the precision (inverse covariance) matrix \mathbf{Q}_* for the effect. With these data, we get the following formula:

```
R> formula <- ~
+ mc(a, model = "generic0", Cmatrix = Q_a, constr = T) +
+ mc(d, model = "generic0", Cmatrix = Q_d, constr = T) +
+ mc(x, model = "generic0", Cmatrix = Q_x, constr = T)
```

We go through the reasoning behind a prior where we utilize all available prior knowledge here, following the tree structure in Table 2.

The expert in genetics has information on the heritability, which is the amount of total variance attributed to the genetic effects and on the distribution of the genetic effect \mathbf{g} to the additive, dominance and epistasis effects \mathbf{a} , \mathbf{d} and \mathbf{x} . The expert says the heritability is around 0.25, but that we want to avoid overfitting. $\omega_{\frac{g}{g+\varepsilon}} \sim \text{PC}_0(0.25)$ fits this desire. The additive, dominance and epistasis effects have according to the expert a division of the genetic variance σ_g^2 that is around (85, 10, 5)%. To achieve this, we must use two dual-splits to decompose the genetic variation, and do this by splitting off the additive effect first, with a $\text{PC}_M(0.85, 0.8)$ prior on $\omega_{\frac{a}{g}}$. Then we attribute the remaining 15% of the genetic variance to \mathbf{d} and \mathbf{x} with 67% to \mathbf{d} with $\text{PC}_M(0.67, 0.8)$ on $\omega_{\frac{d}{d+x}}$. We choose a concentration parameter value of 0.8 because the expert is quite sure about the (85, 10, 5)% division. This corresponds to having 75% of the density mass in the interval $[\text{logit}(m) - 1, \text{logit}(m) + 1]$. The expert does not want to use expert knowledge for the total variance $\sigma_{a+d+x+\varepsilon}^2$, so we use Jeffreys' prior.

Tree structure	Parameters, priors
<pre> graph TD A["a + d + x + ε"] --> B["a + d + x"] A --> C["ε"] B --> D["a"] B --> E["d + x"] E --> F["d"] E --> G["x"] </pre>	$\sigma_{a+d+x+\varepsilon}^2 \sim \text{Jeffreys}'$ $\omega_{\frac{g}{g+\varepsilon}} \sim \text{PC}_0(0.25)$ $\omega_{\frac{a}{g}} \sim \text{PC}_M(0.85, 0.75)$ $\omega_{\frac{d}{d+x}} \sim \text{PC}_M(0.67, 0.75)$

Table 2: Tree structures and the corresponding parameters for the genomic example. $g_i = a_i + d_i + x_i$.

We have simulated a dataset following the description in Hem et al. (2021) (see also Gaynor et al., 2017; Selle et al., 2019), using the R package `AlphaSimR` (Faux et al., 2016; Gaynor, 2019). The source code for simulating the dataset is available in the Supplemental Materials in Hem et al. (2021) (Hem et al., 2020). The dataset is included as `wheat_data` in `makemyprior`. To incorporate the expert knowledge in a unified way, we first scale the covariance matrices have typical variance equal to 1 (for details, see Sørbye and Rue, 2017), using the function `scale_precmat` in `makemyprior`:

```
R> wheat_data_scaled <- wheat_data
R> wheat_data_scaled$Q_a <- scale_precmat(wheat_data$Q_a)
R> wheat_data_scaled$Q_d <- scale_precmat(wheat_data$Q_d)
R> wheat_data_scaled$Q_x <- scale_precmat(wheat_data$Q_x)
```

This model is implemented as follows:

```
R> prior <- make_prior(
+   formula, wheat_data_scaled, prior = list(
+     tree = "s1 = (d, x); s2 = (a, s1); s3 = (s2, eps)",
+     w = list(s1 = list(prior = "pcM", param = c(0.67, 0.8)),
+               s2 = list(prior = "pcM", param = c(0.85, 0.8)),
+               s3 = list(prior = "pc0", param = 0.25)))
```

We now do inference on this model and plot the results:

```
R> posterior <- inference_stan(prior, iter = 15000,
+                             warmup = 5000, seed = 1,
```

```

+                               init = "0", chains = 1)
R> plot_posterior_stan(posterior, param = "prior", prior = T)

Tree structure: d_x = (d,x); a_d_x = (a,d_x);
eps_a_d_x = (eps,a_d_x)

Weight priors:
  w[d/d_x] ~ PCM(0.67, 0.8)
  w[a/a_d_x] ~ PCM(0.85, 0.8)
  w[eps/eps_a_d_x] ~ PC1(0.75)
Total variance priors:
  V[eps_a_d_x] ~ Jeffreys'

SAMPLING FOR MODEL 'full_file' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0.000528 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition
would take 5.28 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:      1 / 15000 [  0%] (Warmup)
Chain 1: Iteration:    1500 / 15000 [ 10%] (Warmup)
Chain 1: Iteration:    3000 / 15000 [ 20%] (Warmup)
Chain 1: Iteration:    4500 / 15000 [ 30%] (Warmup)
Chain 1: Iteration:    5001 / 15000 [ 33%] (Sampling)
Chain 1: Iteration:    6500 / 15000 [ 43%] (Sampling)
Chain 1: Iteration:    8000 / 15000 [ 53%] (Sampling)
Chain 1: Iteration:    9500 / 15000 [ 63%] (Sampling)
Chain 1: Iteration:   11000 / 15000 [ 73%] (Sampling)
Chain 1: Iteration:   12500 / 15000 [ 83%] (Sampling)
Chain 1: Iteration:   14000 / 15000 [ 93%] (Sampling)
Chain 1: Iteration:   15000 / 15000 [100%] (Sampling)
Chain 1:
Chain 1: Elapsed Time: 31.2193 seconds (Warm-up)
Chain 1:                78.244 seconds (Sampling)
Chain 1:                109.463 seconds (Total)
Chain 1:

```

Figure 6 shows the prior and posterior together on the parameterization of the prior (see Table 2).

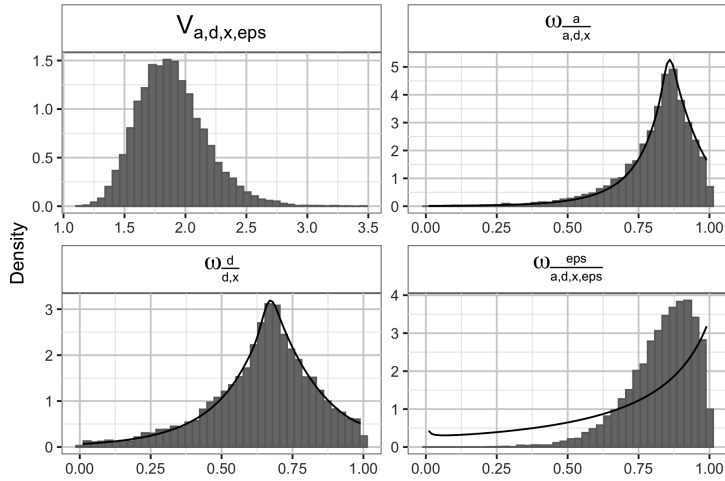


Figure 6: Prior and posterior distribution of the random effect parameters for the genomic selection example.

We see that we do not have enough data to estimate the variance proportion for the additive and nonadditive genetic effects; as the posterior distribution is almost identical to the prior distribution. Hem et al. (2021) have conducted an extensive simulation study on this and similar models. They saw a strong need for robust prior distributions, which we also see in Figure 6, because the nonadditive effects \mathbf{d} and \mathbf{x} are strongly confounded with the environmental effect $\boldsymbol{\varepsilon}$, and the number of observations is small compared to the number of generic markers that needs to be estimated (Sorensen and Gianola, 2007).

5.1.2 Latin square experiment

We consider the latin square experiment in Example 2.2. In line with Fuglstad et al. (2020, Section 5.2), we expand the model and assume the treatment effect now consists of a smooth signal $\boldsymbol{\gamma}^{(1)} = (\gamma_1^{(1)}, \dots, \gamma_9^{(1)}) \sim (\mathbf{0}, \sigma_{\text{RW2}}^2 \mathbf{Q}_{\text{RW2}}^{-1})$ where σ_{RW2}^2 is the variance and $\mathbf{Q}_{\text{RW2}}^{-1}$ is the covariance matrix describing the intrinsic second-order random walk (Rue and Held, 2005, Chapter 3), and random noise $\boldsymbol{\gamma}^{(2)} = (\gamma_1^{(2)}, \dots, \gamma_9^{(2)}) \sim \mathcal{N}_9(\mathbf{0}, \sigma_{\epsilon}^2 \mathbf{I}_9)$. Note that we focus on random effects and exclude intercept and fixed effects. We remove implicit intercept and linear effect by requiring $\sum_{i=1}^9 \gamma_i^{(1)} = 0$ and $\sum_{i=1}^9 i \gamma_i^{(1)} = 0$. To simplify the notation, we use $f_i = \alpha_i + \beta_i + \gamma_1^{(1)} + \gamma_1^{(2)}$.

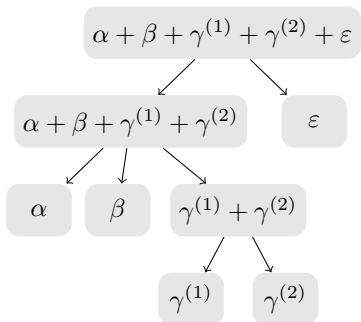
Tree structure	Parameters, priors
 <pre> graph TD A["α + β + γ⁽¹⁾ + γ⁽²⁾ + ε"] --> B["α + β + γ⁽¹⁾ + γ⁽²⁾"] A --> C["ε"] B --> D["α"] B --> E["β"] B --> F["γ⁽¹⁾ + γ⁽²⁾"] F --> G["γ⁽¹⁾"] F --> H["γ⁽²⁾"] </pre>	$\sigma_{\alpha+\beta+\gamma^{(1)}+\gamma^{(2)}+\varepsilon}^2 \sim \text{Jeffreys}'$ $\omega_{\frac{f_i}{f_i+\varepsilon}} \sim \text{PC}_0(0.25)$ $\left(\omega_{\frac{\alpha}{f_i}}, \omega_{\frac{\beta}{f_i}}, 1 - \omega_{\frac{\alpha}{f_i}} - \omega_{\frac{\beta}{f_i}}\right) \sim \text{Dirichlet}(3)$ $\omega_{\frac{\gamma^{(1)}}{\gamma^{(1)}+\gamma^{(2)}}} \sim \text{PC}_0(0.25)$

Table 3: Tree structures and the corresponding parameters for the prior used in the latin square model. $f_i = \alpha_i + \beta_i + \gamma_1^{(1)} + \gamma_1^{(2)}$.

We show how to create the prior distributions in Table 3. We want to avoid overfitting of the model, and use a prior with shrinkage towards the residuals in the top split with median giving 75% residual effect. We do not have any preference for the attribution of the row, column and treatment effects, and use an ignorant Dirichlet prior for the middle split. In the bottom split we again we want to avoid overfitting, and use a prior with shrinkage towards the unstructured treatment effect and a median corresponding to 75% unstructured treatment effect. At last we do not want to say anything about the scale of the total variance, and use the default Jeffreys' prior.

The dataset is included in `makemyprior` as `latin_data`. It is a simulated dataset, following Fuglstad et al. (2020, Section 5.2), where we have used $\sigma_\alpha = \sigma_\beta = \sigma_{\gamma^{(2)}} = \sigma_\varepsilon = 0.1$ and true treatment effect $\gamma_i^{(1)} = 0.02 \cdot ((i - 5)^2 - 20/3)$. The following will fit this model and produce the plots in Figure 7:

```

R> formula <- ~ -1 + mc(row) + mc(col) + mc(iid) +
+   mc(rw2, model = "rw2", constr = T, lin_constr = T)
R> prior <- make_prior(
+   formula, latin_data,
+   prior = list(tree = "s1 = (rw2, iid); s2 = (row, col, s1);
+     s3 = (s2, eps)",
+     w = list(
+       s1 = list(prior = "pc0", param = 0.25),
+       s2 = list(prior = "dirichlet"),
+       s3 = list(prior = "pc0", param = 0.25)))

```

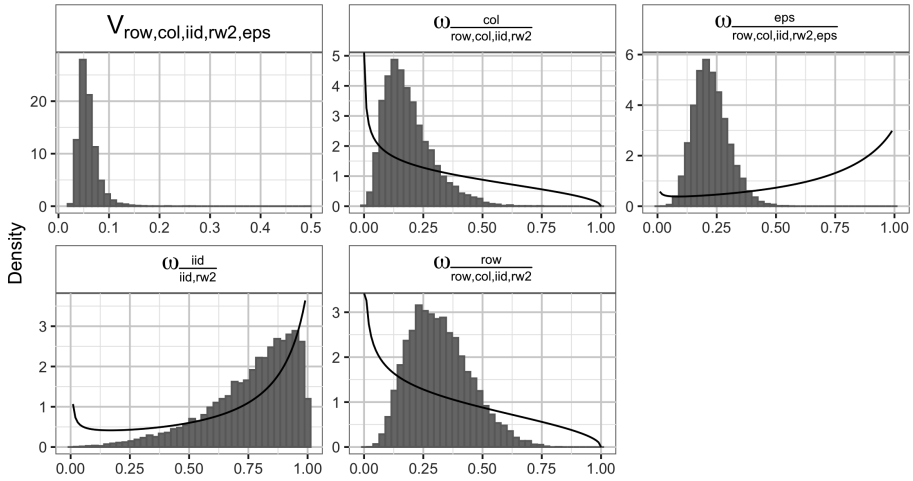


Figure 7: Prior and posterior distribution of the random effect parameters for the latin square example.

```
R> posterior <- inference_stan(
+ prior, iter = 15000, warmup = 5000, seed = 1, init = "0",
+ chains = 1, control = list(adapt_delta = 0.9))
R> plot_posterior_stan(posterior, param = "prior", prior = T)
```

```
Tree structure: iid_rw2 = (iid,rw2); row_col_iid_rw2 =
(row,col,iid_rw2); eps_row_col_iid_rw2 = (eps,row_col_iid_rw2)
```

Weight priors:

```
w[iid/iid_rw2] ~ PC1(0.75)
(w[row/row_col_iid_rw2],
 w[col/row_col_iid_rw2]) ~ Dirichlet(3)
w[eps/eps_row_col_iid_rw2] ~ PC1(0.75)
```

Total variance priors:

```
V[eps_row_col_iid_rw2] ~ Jeffreys'
```

SAMPLING FOR MODEL 'full_file' NOW (CHAIN 1).

Chain 1:

Chain 1: Gradient evaluation took 0.00015 seconds

Chain 1: 1000 transitions using 10 leapfrog steps per transition

```

would take 1.5 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:      1 / 15000 [  0%] (Warmup)
Chain 1: Iteration:    1500 / 15000 [ 10%] (Warmup)
Chain 1: Iteration:    3000 / 15000 [ 20%] (Warmup)
Chain 1: Iteration:    4500 / 15000 [ 30%] (Warmup)
Chain 1: Iteration:    5001 / 15000 [ 33%] (Sampling)
Chain 1: Iteration:    6500 / 15000 [ 43%] (Sampling)
Chain 1: Iteration:    8000 / 15000 [ 53%] (Sampling)
Chain 1: Iteration:    9500 / 15000 [ 63%] (Sampling)
Chain 1: Iteration:   11000 / 15000 [ 73%] (Sampling)
Chain 1: Iteration:   12500 / 15000 [ 83%] (Sampling)
Chain 1: Iteration:   14000 / 15000 [ 93%] (Sampling)
Chain 1: Iteration:   15000 / 15000 [100%] (Sampling)
Chain 1:
Chain 1: Elapsed Time: 16.4578 seconds (Warm-up)
Chain 1:                37.089 seconds (Sampling)
Chain 1:                53.5468 seconds (Total)
Chain 1:

```

Figure 7 shows the prior and posterior together on the parameterization of the prior. The posterior distribution of the bottom split, attributing the treatment effect to the random noise and smooth signal, is only slightly different from the prior, indicating that there is no strong signal about the smooth treatment effect in the data. By using a prior with shrinkage towards only random noise treatment effect, we avoid overfitting. The model has learned about the three other variance proportions, and we see that even though the prior on the amount of total variance going to the residual effect has shrinkage towards 1, the model is not restricted by this.

5.2 Binomial responses: Neonatal mortality

This example is based on a study carried out by Fuglstad et al. (2020). Child mortality is an important indicator of health and well-being in a country. We define neonatal mortality as the number of deaths of infants the first month of life per live birth, which can be estimated using national household surveys from Demographic and Health Surveys (Kenya National Bureau of Statistics et al., 2015). From such surveys we can extract the number of live births $b_{i,j}$ and the number of neonatal deaths $y_{i,j}$ in cluster j in county i , and use an indicator $x_{i,j}$

Tree structure	Parameters, priors
<pre> graph TD A["u + v + nu"] --> B["u + v"] A --> C["nu"] B --> D["u"] B --> E["v"] </pre>	$\sigma_{u+v+\nu}^2 \sim \text{PC}_0(3.35, 0.05)$ $\omega_{\frac{u+v}{u+v+\nu}} \sim \text{PC}_1(0.75)$ $\omega_{\frac{u}{u+v}} \sim \text{PC}_0(0.25)$

Table 4: Tree structures and the corresponding parameters for the neonatal mortality model.

for classifying cluster j in county i as rural ($x_{i,j} = 0$) or urban (1). We model $y_{i,j} | b_{i,j}, p_{i,j} \sim \text{Binomial}(b_{i,j}, p_{i,j})$ with the linear predictor

$$\eta_{i,j} = \text{logit}(p_{i,j}) = \mu + x_{i,j}\beta + u_i + v_i + \nu_{i,j}, \quad i = 1, \dots, n, \quad j = 1, \dots, m_i, \quad (6)$$

where $v_i \sim \mathcal{N}(0, \sigma_v^2)$ and $\nu_{i,j} \sim \mathcal{N}(0, \sigma_\nu^2)$ are i.i.d. random effects for counties and clusters, respectively, and \mathbf{u} is a Besag effect on county with variance σ_u^2 and a sum-to-zero constraint. In the Besag model, the spatial effect of each county depends on the effects in the neighboring regions (see e.g. Besag et al. (1991) for details), and when combining it with an i.i.d. effect on the same level in the hierarchy, we get a BYM (Besag, York and Mollié) model (Besag et al., 1991). We want to investigate whether or not there is a spatial effect present.

We simulated a dataset following the description in Fuglstad et al. (2020, Section 6.2) with the 47 counties in Kenya (see Figure 9 for a map). We used 6, 7 or 8 clusters in each county which gave in total 327 clusters, and thus also 327 observations, $b_{i,j} = 25$ live births in each cluster, and parameters $\mu = -4$, $\beta = 0.1$, $\sigma_v^2 = 0.2$, $\sigma_\nu^2 = 0.1$, and $\sigma_u^2 = 0.5$.

This dataset is available in `makemyprior` as `neonatal_data`, as well as other necessary files for fitting the model.

We prefer coarser over finer unstructured effects, and unstructured over structured effects. That means that we prefer \mathbf{v} over \mathbf{u} and $\mathbf{v} + \mathbf{u}$ over $\boldsymbol{\nu}$ in the prior. The BYM model is intuitively represented with a dual split in the prior tree, where one leaf node represents a Besag effect and the other represents an i.i.d. effect. We achieve this with a prior that distributes the between-county variance with shrinkage towards the unstructured county effect, which gives the BYM2 model of Riebler et al. (2016), and the total variance towards the county effects. Following Fuglstad et al. (2020), we induce shrinkage on the total variance such that we have a 90% credible interval of (0.1, 10) for the effect of $\exp(v_i + u_i + \nu_{i,j})$.

We use the function `find_pc_prior_param()` in `makemyprior` to find the parameters for the PC prior:

```
R> set.seed(1)
R> find_pc_prior_param(lower = 0.1, upper = 10,
+                      prob = 0.9, N = 2e5)
```

```
U = 3.353132
Prob(0.09866969 < exp(eta) < 9.892902) = 0.9
```

This gives a $PC_0(3.35, 0.05)$ prior. The tree structure and a summary of the prior distributions can be found in Table 4. We fit the model with Stan:

```
R> graph_path <- paste0(path.package("makemyprior"),
+                       "/neonatal.graph")
R> formula <- y ~ mc(nu) + mc(v) +
+   mc(u, model = "besag", graph = graph_path,
+     scale.model = TRUE)
R> prior <- make_prior(
+   formula, neonatal_data, family = "binomial",
+   prior = list(
+     tree = "s1 = (u, v); s2 = (s1, nu)",
+     w = list(s1 = list(prior = "pc0", param = 0.25),
+              s2 = list(prior = "pc1", param = 0.75)),
+     V = list(s2 = list(prior = "pc", param = c(3.35, 0.05)))
+   ))
R> posterior <- inference_stan(
+   prior, iter = 15000, warmup = 5000, seed = 1, init = "0",
+   chains = 1, control = list(adapt_delta = 0.85))
```

```
Tree structure: v_u = (v,u); nu_v_u = (nu,v_u)
```

Weight priors:

```
w[v/v_u] ~ PC1(0.75)
w[nu/nu_v_u] ~ PC0(0.25)
```

Total variance priors:

```
sqrt(V)[nu_v_u] ~ PC0(3.35, 0.05)
```

```
SAMPLING FOR MODEL 'full_file' NOW (CHAIN 1).
```

```
Chain 1:
```

```
Chain 1: Gradient evaluation took 0.000567 seconds
```

```

Chain 1: 1000 transitions using 10 leapfrog steps per transition
would take 5.67 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:      1 / 15000 [  0%] (Warmup)
Chain 1: Iteration:    1500 / 15000 [ 10%] (Warmup)
Chain 1: Iteration:    3000 / 15000 [ 20%] (Warmup)
Chain 1: Iteration:    4500 / 15000 [ 30%] (Warmup)
Chain 1: Iteration:    5001 / 15000 [ 33%] (Sampling)
Chain 1: Iteration:    6500 / 15000 [ 43%] (Sampling)
Chain 1: Iteration:    8000 / 15000 [ 53%] (Sampling)
Chain 1: Iteration:    9500 / 15000 [ 63%] (Sampling)
Chain 1: Iteration:   11000 / 15000 [ 73%] (Sampling)
Chain 1: Iteration:   12500 / 15000 [ 83%] (Sampling)
Chain 1: Iteration:   14000 / 15000 [ 93%] (Sampling)
Chain 1: Iteration:   15000 / 15000 [100%] (Sampling)
Chain 1:
Chain 1: Elapsed Time: 79.4016 seconds (Warm-up)
Chain 1:                157.281 seconds (Sampling)
Chain 1:                236.683 seconds (Total)
Chain 1:

```

Note that for inference with INLA, the `Ntrials` argument must be provided to `inference_inla()`. The following produce the plots in Figure 8 and gives some key information on the posterior:

```

R> plot_fixed_posterior(posterior)
R> plot_posterior_stan(posterior, param = "prior", prior = T)
R> posterior

```

```

Model: y ~ urban + mc(nu) + mc(v) + mc(u, model = "besag",
graph = graph_path, scale.model = T)
Tree structure: v_u = (v,u); nu_v_u = (nu,v_u)

```

Inference done with Stan.

Param.	mean	median	sd
V[nu_v_u]	0.662	0.630	0.244
w[v/v_u]	0.655	0.715	0.263
w[nu/nu_v_u]	0.333	0.330	0.190

```

intercept    -4.157 -4.153 0.135
urban        0.469  0.469 0.169

```

Figure 9 shows the posterior spatial effect e^{u_i} plotted in a map. We see a spatial variation between the counties. The necessary data for creating the spatial map are not included, but can be obtained from <https://gadm.org/>. The samples for effects and variances can easily be extracted with `extract_posterior_effects()` and `extract_posterior_variance()`, respectively, which both take the arguments `obj` from `inference_stan()` and the name of the effect `effname`:

```

R> u <- extract_posterior_effects(posterior, "u")
R> u_var <- extract_posterior_variance(posterior, "u")

```

We have only used `u` for the plot in Figure 9.

The fixed effects in Figure 8a show that the intercept is not contributing much to the linear predictor, while the effect of urban/rural shows that there is a higher mortality in urban areas (which is the case also for real data, see Kenya National Bureau of Statistics et al. (2015)). From Figure 8b we see that the model has learned about the total variance from the data and about the amount of total (latent) variance to the cluster effect (ν), but there is not enough information in the data about the amount of county variance to the unstructured county effect (u). The following fits the model without the likelihood (sampling from the prior) and produces the plots of the prior and posterior on standard deviation scale in Figure 10:

```

R> prior_samps <- inference_stan(
+   prior, use_likelihoood = FALSE, iter = 15000,
+   warmup = 5000, seed = 1, init = "0", chains = 1)
R> plot_several_posterior_stan(
+   list(Prior = prior_samps, Posterior = posterior), "stdev")

```

```
Tree structure: v_u = (v,u); nu_v_u = (nu,v_u)
```

Weight priors:

```

w[v/v_u] ~ PC1(0.75)
w[nu/nu_v_u] ~ PC0(0.25)

```

Total variance priors:

```
sqrt(V)[nu_v_u] ~ PC0(3.35, 0.05)
```

```
SAMPLING FOR MODEL 'full_file' NOW (CHAIN 1).
```

```
Chain 1:
```

```

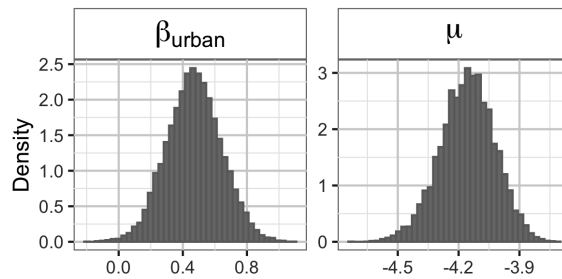
Chain 1: Gradient evaluation took 0.000118 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition
would take 1.18 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:      1 / 15000 [  0%] (Warmup)
Chain 1: Iteration:    1500 / 15000 [ 10%] (Warmup)
Chain 1: Iteration:    3000 / 15000 [ 20%] (Warmup)
Chain 1: Iteration:    4500 / 15000 [ 30%] (Warmup)
Chain 1: Iteration:    5001 / 15000 [ 33%] (Sampling)
Chain 1: Iteration:    6500 / 15000 [ 43%] (Sampling)
Chain 1: Iteration:    8000 / 15000 [ 53%] (Sampling)
Chain 1: Iteration:    9500 / 15000 [ 63%] (Sampling)
Chain 1: Iteration:   11000 / 15000 [ 73%] (Sampling)
Chain 1: Iteration:   12500 / 15000 [ 83%] (Sampling)
Chain 1: Iteration:   14000 / 15000 [ 93%] (Sampling)
Chain 1: Iteration:   15000 / 15000 [100%] (Sampling)
Chain 1:
Chain 1: Elapsed Time: 6.33609 seconds (Warm-up)
Chain 1:                19.7902 seconds (Sampling)
Chain 1:                26.1263 seconds (Total)
Chain 1:

```

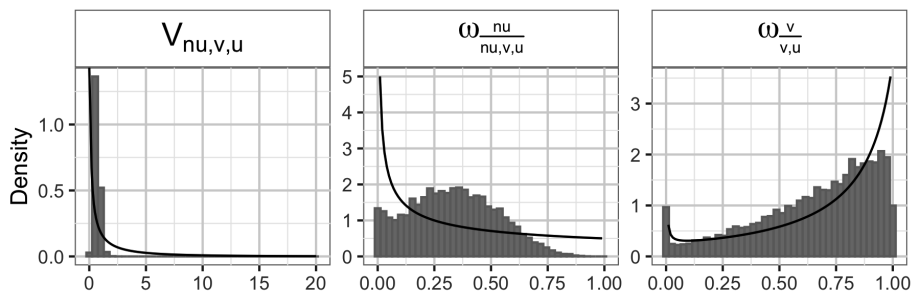
From these graphs we see that the posterior of the standard deviations are clearly different from the prior. We saw in Figure 8b that the model did not learn much about the amount of county variation accounted for by the Besag effect (\mathbf{u}), but we cannot see this from plots of the posterior standard deviations, and they do not show the whole picture. This is another advantage of the HD prior; it is easy to see that even though we get the impression that the model has learned from the data, that knowledge is not necessarily about the whole model. This shows, as Fuglstad et al. (2020) points out, that one should be careful before drawing conclusions on first impressions about the results, and more investigation should be done.

6 Summary and discussion

The `makemyprior` package offers an intuitive and transparent way of choosing and visualizing prior distributions. It is easy to utilize expert knowledge, and clear what prior distributions are used, also when the default settings are chosen.



(a) Posterior of the effect of urban/rural and the intercept.



(b) Random effect parameters.

Figure 8: Prior and posterior distribution of a) coefficients of the fixed effects and b) total variance and variance proportions of the random effects for the neonatal mortality example.

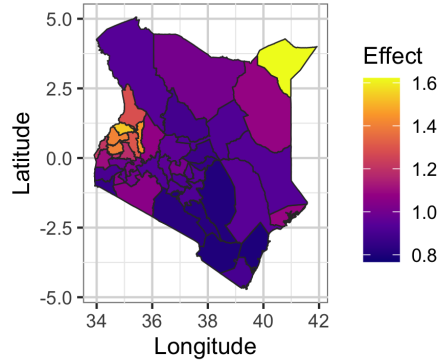


Figure 9: Posterior median of e^{u_i} for each county in Kenya. Note that this is based on simulated data.

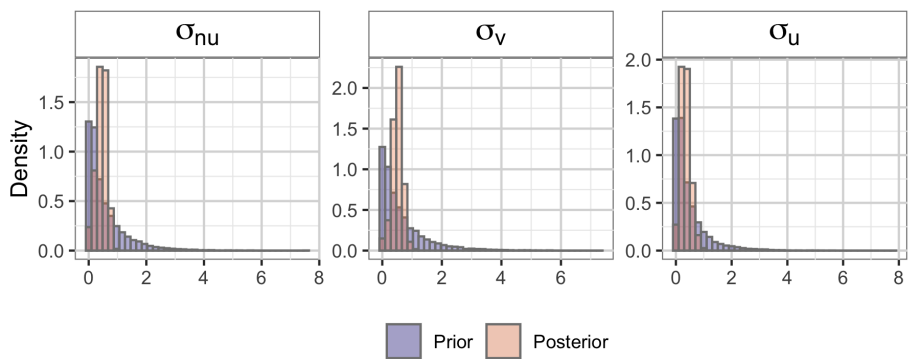


Figure 10: Prior and posterior distribution for the neonatal mortality example on standard deviation scale.

There are naturally some limitations. The hierarchical decomposition (HD) prior itself is restricted to latent Gaussian models (Fuglstad et al., 2020). `makemyprior` only handles a selection of latent models, and it does not open for using models that are specified by other parameters than variances, such as an auto-regressive processes of order 1 with a correlation parameter. However, this is not a big restriction, as other latent models and models with e.g. correlations can however be implemented by the user with a custom Stan-code. Then the correlation parameter can be assigned an independent prior (in the same way an individual variance parameter or fixed effect get independent priors). The package is limited to Gaussian, binomial and Poisson likelihoods, covering a range of applications, and other likelihoods can be implemented in a custom Stan-code.

If the inference for a model is carried out with another software than INLA and `rstan`, the `makemyprior` package is still useful. `makemyprior` computes the hierarchical decomposition (HD) prior, and the usage of the prior is not limited to inference carried out with `rstan` or INLA. The package can be used to investigate the prior choices, and the user can simulate from the prior with Stan and look at the prior distributions on different parameterizations. In this way, crucial misunderstandings of what prior distributions are used can be discovered and corrected, and thus increase the understanding and meaning of the prior.

To include fixed effects in the HD prior framework has been discussed by Fuglstad et al. (2020), but has not yet been investigated or done. To see how the individual fixed effects contribute to the total data variation would be interesting, but fixed effects are often correlated, and the variance that is explained by each single fixed effect is not well defined. The perhaps most intuitive way to include fixed effects is to assign one variance parameter to each effect. However, this can quickly increase the amount of variance parameters to a level where inference become computationally hard. Gelman and Hill (2007) and Zhang et al. (2020) have proposed prior distributions related to the coefficient of determination, R^2 , which measures the amount of variance explained by the model. The generalized R^2 proposed by Gelman and Hill (2007) measures this at each level in the hierarchical model.

The idea behind the HD prior can be extended to models outside the class of latent Gaussian models, or to models where the hyperparameters of the priors will get prior distributions. This will however require further development of the framework, and be highly computational expensive, as the penalized complexity (PC) prior cannot be pre-computed in the same way as we do now with the conditioning on the hyperparameters.

Other ideas include to open for customized specification of prior distributions, for both variance and variance proportion parameters, in the `makemyprior`

package (the HD prior framework is already open for this). The PC prior can be computationally hard to calculate for large models, and the possibility of including for example customized beta distributions for each split could be helpful. However, the prior is only computed once as we approximate it by conditioning on the medians and base models for lower splits, and the model size will only increase computation time in the computation if the prior, not during inference. In addition, this will complicate the intuition behind the prior, and it will be more difficult to use prior and expert knowledge in a transparent way. It will require more thoughtful prior choices, and we lose one of the large advantages with the easy-to-use and intuitive way of making priors with `makemyprior`, in addition to the shrinkage properties of the PC prior. Including more latent models will increase the amount of applications the package can be used for directly, without specifying custom Stan-code. To open for easy integration into other softwares for inference, such as the Template Model Builder (TMB, Kristensen et al., 2016), can be done, and can be useful for models that are very complex and will be highly time consuming and difficult to fit with `rstan` or `INLA`.

In conclusion, `makemyprior` is a valuable addition to the range of packages that can be used to carry out inference. It is easy to include prior knowledge in an intuitive and transparent way, can be used to verify prior choices, and allows direct inference in a simple way. It increases the awareness of what prior is used, which is important when doing inference, to ensure that the model fitted is indeed the intended one.

Computational details

The results in this paper were obtained with:

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Catalina 10.15.7
```

Package version of dependent, imported and suggested packages are:

```
ggplot2_3.3.2
Matrix_1.2.18
knitr_1.29
shiny_1.5.0
shinyjs_2.0.0
shinyBS_0.61
visNetwork_2.0.9
```


rmarkdown_2.3
testthat_2.3.2
splines_4.0.2
MASS_7.3.51.6
ggpubr_0.4.0
rstan_2.21.2
INLA_20.3.17
ggplot2_3.3.2

References

- 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.
- Caldwell, A., Kollar, D., and Kröninger, K. (2009). BAT—The Bayesian analysis toolkit. *Computer Physics Communications*, 180(11):2197–2209.
- 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).
- Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). shiny: Web application framework for r. <https://CRAN.R-project.org/package=shiny>. R package version 1.5.0.
- Depaoli, S., Winter, S. D., and Visser, M. (2020). The importance of prior sensitivity analysis in Bayesian statistics: Demonstrations using an interactive Shiny app. *Frontiers in Psychology*, 11.
- Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., Hearne, S. J., Gonen, S., and Hickey, J. M. (2016). AlphaSim: Software for breeding program simulation. *The Plant Genome*, 9(3):1–14.
- Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.
- Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*, 15(4):1109–1137.
- Gabry, J. (2018). shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models. <https://CRAN.R-project.org/package=shinystan>. R package version 2.5.0.

- Gaynor, C. (2019). Alphasimr: Breeding program simulations. <https://CRAN.R-project.org/package=AlphaSimR>. R package version 0.10.0.
- Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., Mackay, I. J., and Hickey, J. M. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, 57(5):2372–2386.
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multi-level/Hierarchical Models*, volume 1. Cambridge University Press, New York, New York.
- 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.
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808 [stat.ME]*.
- Goel, P. K. and Degroot, M. H. (1981). Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association*, 76(373):140–147.
- Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. <https://mc-stan.org/rstanarm>. R package version 2.21.1.
- 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.
- Hem, I. G., Selle, M., Gorjanc, G., Fuglstad, G.-A., and Riebler, A. (2020). Supplemental material for "Robust modelling of additive and non-additive variation with intuitive inclusion of expert knowledge". <https://doi.org/10.6084/m9.figshare.12040716>.
- Hem, I. G., Selle, M. L., Gorjanc, G., Fuglstad, G.-A., and Riebler, A. (2021). Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge. *Genetics*. iyab002.
- Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.
- Holand, A. M., Steinsland, I., Martino, S., and Jensen, H. (2013). Animal models and integrated nested Laplace approximations. *G3: Genes, Genomes, Genetics*, 3(8):1241–1251.

- 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.
- Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21.
- 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.
- Plummer, M. (2017). JAGS version 4.3.0 user manual [Computer software manual]. sourceforge.net/projects/mcmc-jags/files/Manuals/4.x.
- 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.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC press, Boca Raton, Florida.
- 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 (Statistical Methodology)*, 71(2):319–392.
- 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*, 4:395–421.
- Selle, M. L., Steinsland, I., Hickey, J. M., and Gorjanc, G. (2019). Flexible modelling of spatial variation in agricultural field trials with the R package INLA. *Theoretical and Applied Genetics*, 132(12):3277–3293.
- 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.
- Smid, S. C. and Winter, S. D. (2020). Dangers of the defaults: A tutorial on the impact of default priors when using Bayesian SEM with small samples. *Frontiers in Psychology*, 11:3536.

- Sørbye, S. H. and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38(6):923–935.
- Sorensen, D. and Gianola, D. (2007). *Likelihood, Bayesian, and MCMC methods in Quantitative Genetics*. Springer Science & Business Media.
- Stan Development Team (2020). RStan: the R interface to Stan. <http://mc-stan.org/>. R package version 2.21.2.
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. <https://mc-stan.org/loo/>. R package version 2.4.1.
- Wakefield, J. (2006). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.
- Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). Bayesian regression using a prior on the model fit: The R2-D2 shrinkage prior. *Journal of the American Statistical Association*, 0(0):1–13.
- Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., and de Schoot, R. V. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development*, 14(4):305–320.

A Code

We include all R code used in the paper.

```

library(makemyprior)

#### Example model for Section 4 ####

formula <- y \texttildelow x + mc(a) + mc(b)

p <- 10
m <- 10
n <- m*p

set.seed(1)
data <- list(a = rep(1:p, each = m),
             b = rep(1:m, times = p),
             x = runif(n))
data$y <- data$x + rnorm(p, 0, 0.5)[data$a] +
  rnorm(m, 0, 0.3)[data$b] + rnorm(n, 0, 1)

prior <- make_prior(formula, data, family = "gaussian",
                    intercept_prior = c(0, 1000),
                    covariate_prior = list(x = c(0, 100)))

new_prior <- makemyprior_gui(prior)

summary(new_prior)

prior <- make_prior(
  formula, data,
  prior = list(
    tree = "s1 = (a, b); s2 = (s1, eps)",
    w = list(s1 = list(prior = "pcM", param = c(0.7, 0.5)),
             s2 = list(prior = "pc0", param = 0.25)),
    V = list(s2 = list(prior = "pc0", param = c(3, 0.05)))
  ),
  covariate_prior = list(x = c(0, 100)))

prior

```

```

#### Genomic model for wheat breeding from Section 5.1 ####

library(makemyprior)

wheat_data_scaled <- wheat_data
wheat_data_scaled$Q_a <- scale_precmat(wheat_data$Q_a)
wheat_data_scaled$Q_d <- scale_precmat(wheat_data$Q_d)
wheat_data_scaled$Q_x <- scale_precmat(wheat_data$Q_x)

formula <- y \texttt{d} \texttt{e} \texttt{l} \texttt{o} \texttt{w}
  mc(a, model = "generic0", Cmatrix = Q_a, constr = T) +
  mc(d, model = "generic0", Cmatrix = Q_d, constr = T) +
  mc(x, model = "generic0", Cmatrix = Q_x, constr = T)

prior <- make_prior(formula, wheat_data_scaled, prior = list(
  tree = "s1 = (d, x); s2 = (a, s1); s3 = (s2, eps)",
  w = list(s1 = list(prior = "pcM", param = c(0.67, 0.8)),
    s2 = list(prior = "pcM", param = c(0.85, 0.8)),
    s3 = list(prior = "pc0", param = 0.25)))

# the first time you do inference with Stan, we recommend to run:
compile_stan(save = T)

posterior <- inference_stan(prior, iter = 15000,
  warmup = 5000, seed = 1,
  init = "0", chains = 1)

plot_posterior_stan(posterior, param = "prior", prior = T)

#### Latin square experiment from Section 5.1 ####

formula <- y \texttt{d} \texttt{e} \texttt{l} \texttt{o} \texttt{w} -1 + mc(row) + mc(col) + mc(iid) +
  mc(rw2, model = "rw2", constr = T, lin_constr = T)

prior <- make_prior(
  formula, latin_data,
  prior = list(tree = "s1 = (rw2, iid); s2 = (row, col, s1);
    s3 = (s2, eps)",
  w = list(

```

```

s1 = list(prior = "pc0", param = 0.25),
s2 = list(prior = "dirichlet"),
s3 = list(prior = "pc0", param = 0.25)))

posterior <- inference_stan(prior, iter = 15000, warmup = 5000,
  seed = 1, init = "0", chains = 1,
  control = list(adapt_delta = 0.9))

plot_posterior_stan(posterior, param = "prior", prior = T)

#### Neonatal mortality from Section 5.2 ####

set.seed(1)
find_pc_prior_param(lower = 0.1, upper = 10, prob = 0.9, N = 2e5)

graph_path <- paste0(path.package("makemyprior"),
  "/neonatal.graph")

formula <- y \texttt{delow urban} + mc(nu) + mc(v) +
  mc(u, model = "besag", graph = graph_path, scale.model = T)

prior <- make_prior(
  formula, neonatal_data, family = "binomial",
  prior = list(tree = "s1 = (u, v); s2 = (s1, nu)",
    w = list(s1 = list(prior = "pc0", param = 0.25),
      s2 = list(prior = "pc1", param = 0.75)),
    V = list(s2 = list(prior = "pc",
      param = c(3.35, 0.05)))))

posterior <- inference_stan(prior, iter = 15000, warmup = 5000,
  seed = 1, init = "0", chains = 1,
  control = list(adapt_delta = 0.85))

plot_fixed_posterior(posterior)
plot_posterior_stan(posterior, param = "prior", plot_prior = TRUE)
posterior

u <- extract_posterior_effects(posterior, "u")
u_var <- extract_posterior_variance(posterior, "u")

```

```
prior_samps <- inference_stan(prior, use_likelihood = FALSE,  
                             iter = 15000, warmup = 5000,  
                             seed = 1, init = "0", chains = 1)  
  
plot_several_posterior_stan(  
  list(Prior = prior_samps, Posterior = posterior), "stdev")  
  
####
```


ISBN 978-82-326-6915-8 (printed ver.)
ISBN 978-82-326-6411-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)