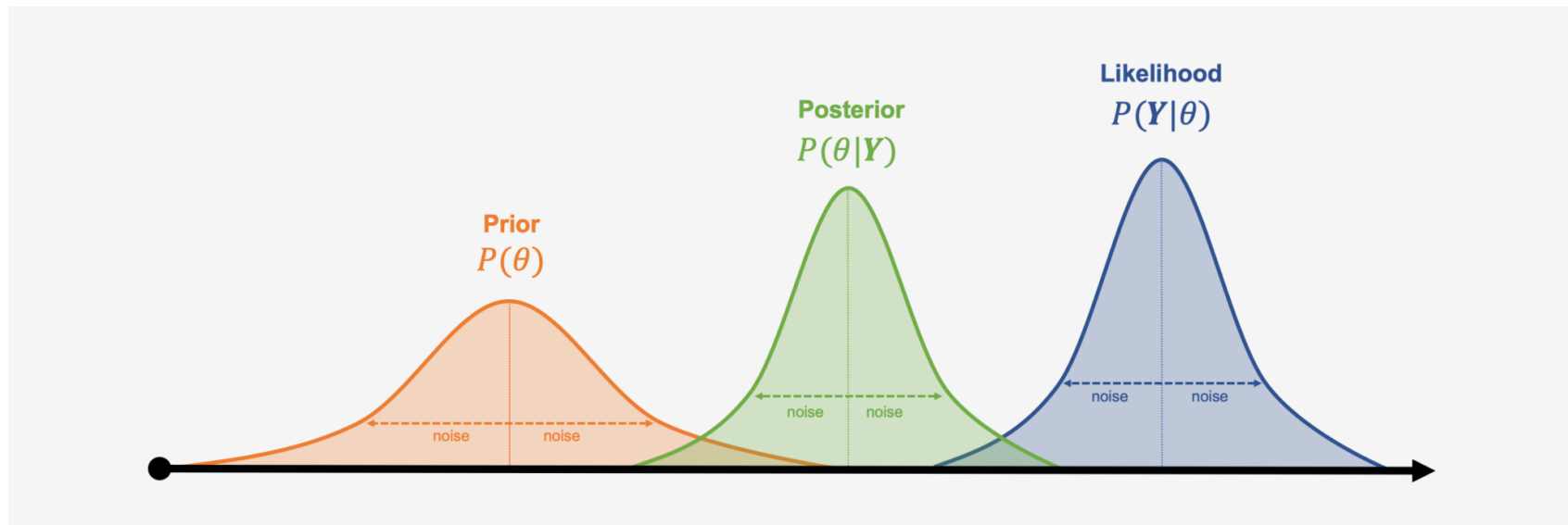


Prior distributions for variance parameters



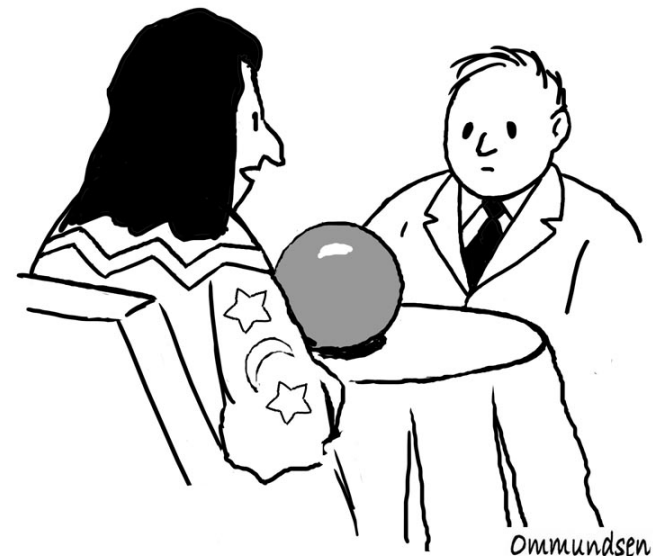
<https://hudsonthames.org/bayesian-portfolio-optimisation-the-black-litterman-model/> (15.03.22)

About the choice of prior distributions

The issue of **setting prior distributions on model parameters** is a difficult **issue** in applied Bayesian statistics, in particular for parameters further down the model hierarchy, such as **variance (or precision) parameters**.

What is often done in practice?

- **Ignore the problem** and use the default priors specified in a **software package**.
- Choose **priors used in the literature** and hope to avoid criticism.



**“Is this needed for a
Bayesian analysis?”**

Setting: Bayesian hierarchical models (BHM)

We consider BHMs where the observations $\mathbf{y}$ are conditionally independent $\pi(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}_L) = \prod_{i=1}^n \pi(y_i \mid \eta_i, \boldsymbol{\theta}_L)$ and the linear predictor is described as

$$\eta_i = \underbrace{\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}}_{\text{Fixed effects}} + \underbrace{\sum_{j=1}^N u_{j,k_j[i]} }_{\text{Random effects}}$$

Here, $\boldsymbol{\theta}_L$ denotes the likelihood parameters. A latent model is specified for each random effect $\pi(\mathbf{u}_j \mid \sigma_j^2)$. Let $\boldsymbol{\theta}_M = (\sigma_1^2, \dots, \sigma_M^2)$. The BHM is completed by specifying a prior for $(\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_L, \boldsymbol{\theta}_M)$.

Our focus is on $\boldsymbol{\theta}_M$.

There is a vast literature on prior distributions

- uniform priors
- Jeffreys' priors
- reference priors
- weakly informative priors
- ...

Prior distributions encode prior knowledge but also stabilize inference.

About prior choices

Simpson, Rue, Martins, Riebler and Sørbye (2015)

“Prior selection is the fundamental issue in Bayesian statistics. Priors are the Bayesian’s greatest tool, but they are also the greatest point for criticism: the arbitrariness of prior selection procedures and the lack of realistic sensitivity analysis are a serious argument against current Bayesian practice.”

Reference:

Simpson, D. P., Rue, H., Martins, T. G., Riebler, A. and Sørbye, S. H. (2017). Penalising model component complexity: A principled practical approach to constructing priors. *Statistical Science* (32) 1–28.

Assignment of hyperpriors

The scaling problem of intrinsic (improper) model components

- Models for splines (rw1 , rw2)
- The conditional autoregressive models for area/regional models
e.g. used in disease mapping.
- and others...

Problem:

- These models are **unscaled** and their **properties change** with
locations/dimension/graph.

Illustration RW1: Marginal variance

```
1 > rw1(5)
2 [1,]  1 -1  .  .  .
3 [2,] -1  2 -1  .  .
4 [3,]  . -1  2 -1  .
5 [4,]  .  . -1  2 -1
6 [5,]  .  .  . -1  1
7 > geom.mean(diag(ginv(rw1(5))))
8 [1] 0.73
9 > geom.mean(diag(ginv(rw1(50))))
10 [1] 7.55
11 > geom.mean(diag(ginv(rw1(500))))
12 [1] 75.580
```

Consider the generalised variance

$$\sigma_{\text{GV}}^2 = \exp \left(\frac{1}{n} \sum_{i=1}^n \log \left(\frac{1}{\tau_u} [\mathbf{R}^{-1}]_{ii} \right) \right) = \frac{1}{\tau} \exp \left(\frac{1}{n} \sum_{i=1}^n \log([\mathbf{R}^{-1}]_{ii}) \right)$$

IGMRFs need to be scaled

That means:

- An **uninformative prior** on τ could be very informative on σ^2 .

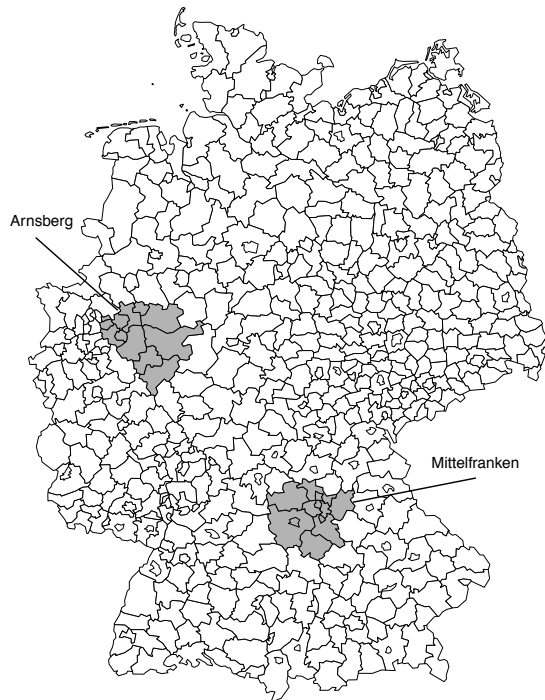
⇒ Scale the IGMRF such that $\sigma_\tau^2 = 1/\tau$.

In R-INLA

1

```
formula = f(., model="..", hyper=..., scale.model=T)
```

Illustration: Besag model



We consider two administrative regions:

- Mittelfranken
- Arnsberg

Both regions have 12 districts.

Can we use the same hyperpriors?

What is the issue?

[Sørbye and Rue (2014), Spat Stat]

The marginal variances $\tau_u^{-1}[\mathbf{R}^-]_{ii}$ depend on the graph structure!

Consider the generalised variance

$$\sigma_{\text{GV}}^2 = \frac{1}{\tau_u} \exp \left(\frac{1}{n} \sum_{i=1}^n \log ([\mathbf{R}^-]_{ii}) \right)$$

If $\tau_u = 1$ we get

- $\sigma_{\text{GV}}^2 \approx 0.40$ for Arnsberg
- $\sigma_{\text{GV}}^2 \approx 0.29$ for Mittelfranken
- $\sigma_{\text{GV}}^2 \approx 0.56$ for whole Germany

A prior for τ_u will not mean the same thing for every problem!

How do we scale?

In order to unify the interpretation of the prior for τ_u and make it transferable, we need to scale the model so that

$$\sigma_{GV}^2 = \frac{1}{\tau_u}$$

Now, τ_u is the precision of the (marginal) deviation from a constant level, independently from the underlying graph.

Caution: This scaling issues applies to all GMRFs that have not a full-rank precision matrix (intrinsic GMRFs)!

```
1 formula = f(. , model = "..", hyper = ..., scale.model = TRUE)
```

The R-package INLA contains a function to scale any singular precision matrix `R`, see `?inla.scale.model`

How to choose our parameters?

- Assume $\tau \sim \text{Gamma}(a, b)$ where $E(\tau) = a/b$.
- We can say something about the **scale** of the effect with

$$\sigma = \sqrt{1/\tau}$$

For example:

$$\text{Prob}(\sigma > U) = \alpha$$

From this we can derive parameter b , if we fix a value for a , say.

[Sørbye and Rue, 2014, Spat Stat](#); [Papoila et al., 2014, Biom J](#)

- **However:** Why are we using a Gamma distribution, why not half-Cauchy ...?

Penalised complexity (PC) priors

Simpson et al. (2015) introduced a new concept of defining priors that are **robust**, **invariant to reparameterisations** and **principle based**.

Main idea: Occam's razor—a principle of parsimony

Simpler model formulations should be preferred until there is enough support in the data for a more complex model.

Consider INLA

Building models adding up model components

$$\eta = \mathbf{X}\boldsymbol{\beta} + f_1(\dots; \boldsymbol{\theta}_1) + f_2(\dots; \boldsymbol{\theta}_2) + \dots$$

- Many model components represent a flexible extension of a base model.
- Put a prior on the *distance* between the flexible model and the base model.
- Important: Mode should be at a distance equal to zero.
- Transform the prior back to the parameter of interest.

1. Principle: Occam's razor

- Many model components represent a flexible extension of a base model. For each model component $\mathbf{x}$ we define a flexible model

$$f = \pi(\mathbf{x}|\xi)$$

where ξ is interpreted as a flexibility parameter.

- f is a flexible version of a base model

$$g = \pi(\mathbf{x}|\xi = \xi_0)$$

Examples for base models

Case	Parameter	ξ	Base model
IID	τ (precision)	$\xi = 1/\tau$	$\xi = 0$ (no random effect)
IGMRF	τ (precision)	$\xi = 1/\tau$	$\xi = 0$ (constant, line, plane)
AR1	ρ (correlation)	$\xi = \rho$	$\xi = 0$ (no time-dependence) $\xi = 1$ (no change in time)
Student-t	ν (dof)	$\xi = 1/\nu$	$\xi = 0$ (Gaussian)
Correlation matrix	$\mathbf{Q}$	$\xi = \mathbf{Q}$	$\xi = \mathbf{I}$ (no correlation)

1. Occams razor

- The prior for $\xi \geq 0$ should penalise the complexity introduced by ξ
- The prior should be decaying with increasing complexity (the mode should be at the base model)

A prior will cause overfitting (force complexity) if, loosely,

$$\pi_{\xi}(\xi = 0) = 0$$

2. Principle: Measure of complexity

Use **Kullback-Leibler discrepancy** to measure the increased complexity introduced by $\xi > 0$,

$$\text{KLD}(f\|g) = \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx$$

for flexible model f and base model g .

Example

Assume that the flexible model f is a $\mathcal{N}(\mu; 1)$ where $\mu > 0$. The base model g refers to $\mu = 0$. Then

$$\text{KLD}(f\|g) = \frac{\mu^2}{2}$$

3. Principle: Constant-rate penalisation

Main idea

Assign priors to “distances” between models, instead of assigning priors to the parameters.

- Define the (uni-directional) “distance”

$$d(\xi) = \sqrt{2 \text{KLD}(\xi)}$$

- Assign an exponential distribution to $d(\xi)$:

$$\pi(d(\xi)) = \lambda \exp(-\lambda d(\xi)), \quad \lambda > 0$$

which has mode at $d(\xi) = 0$.

- Do the change-of-variables to get a prior for the parameter of interest.

4. Principle: User-defined scaling

- Determine λ based on some knowledge of the model component, for example in terms of prior mass in the tail.
- A natural criterion for IGMRFs is

$$P(\sigma > U) = P\left(\frac{1}{\sqrt{\tau}} > U\right) = \alpha$$

where U is an upper limit for the standard deviation and α is a small probability.

- The scale U determines the magnitude of the effect of a model component and how **informative** the prior will be.

Example: Precision of a Gaussian

Analytic result in this case (type-2 Gumbel):

$$\pi(\tau) = \frac{\theta}{2} \tau^{-3/2} \exp(-\theta/\sqrt{\tau}), \quad E(\tau) = \infty,$$

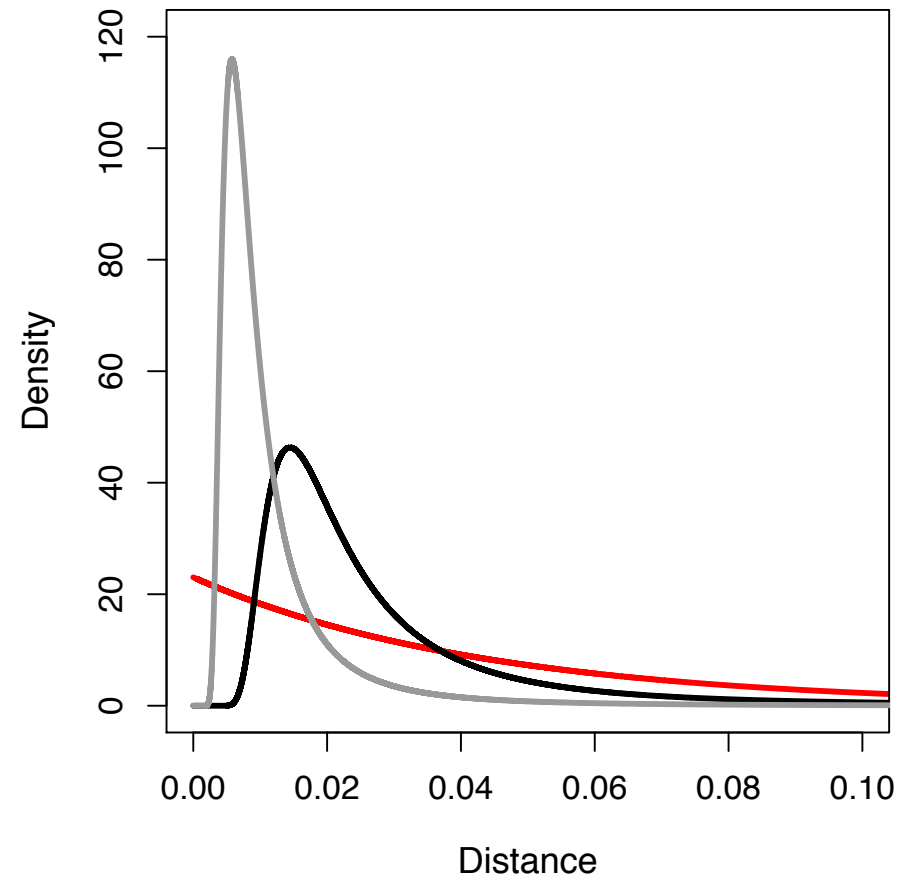
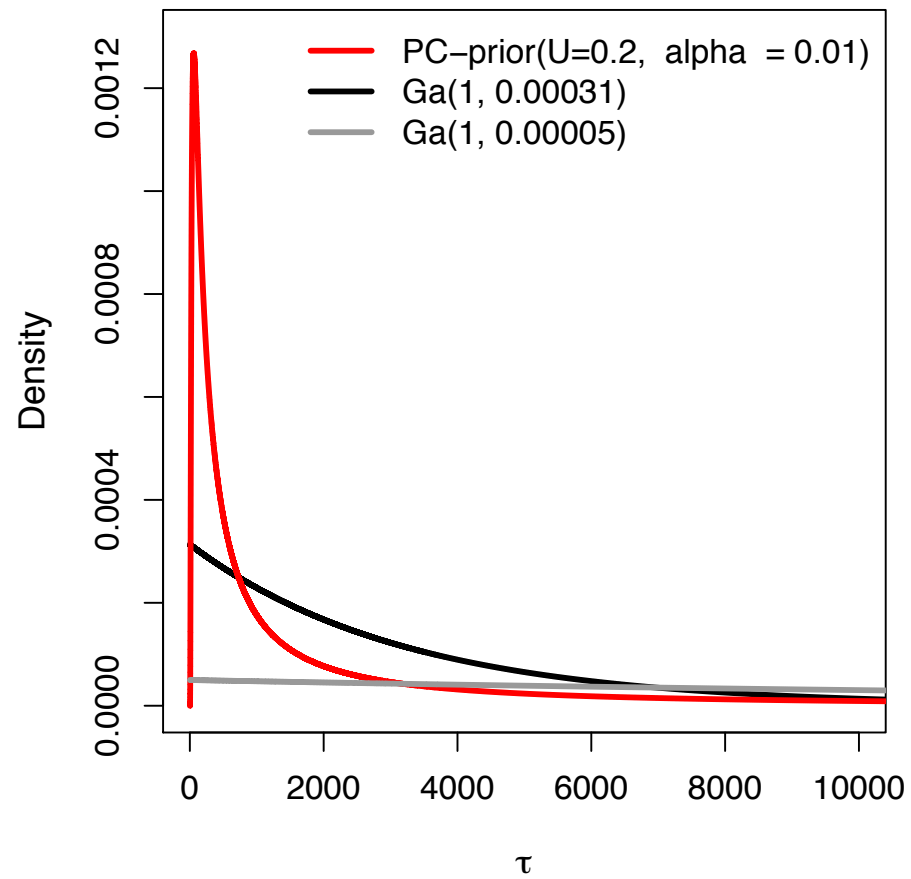
where $\text{Prob}(\sigma > U) = \alpha$ gives

$$\theta = -\frac{\ln(\alpha)}{U}$$

Alternative interpretation

$$\pi(\sigma) = \theta \exp(-\theta\sigma)$$

Comparison to a gamma prior



How to do this in INLA

Specifying the pc-prior in the f-function:

```
1 hyper = list(precision =  
2   list(prior = "pc.prec",  
3     param = c(u, alpha)))
```

Documentation:

```
1 inla.doc("pc.prec")
```

Extensions

This is just a very simple example. So far PC priors have been proposed for a lot of models, for example

- Disease mapping models: Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4), 1145-1165.
- Autoregressive processes: Sørbye, S. H., & Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38(6), 923-935.
- Gaussian random fields: Fuglstad, G. A., Simpson, D., Lindgren, F., & Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525), 445-452.
- As building block for a joint prior for the variance components of all random effects Fuglstad, G. A., Hem, I. G., Knight, A., Rue, H., & Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*, 15(4), 1109-1137.

Hierarchical decomposition (HD) priors (Fuglstad et al. 2020, Bayesian Anal.)

Idea:

- Distribute the total variance along a tree structure to the individual random effect components.
- For each split in the tree, the analyst decides to
 - ▶ be ignorant $\Rightarrow$ Dirichlet prior
 - ▶ have a sound intuition
 $\Rightarrow$ penalised complexity prior for robust shrinkage

(Simpson et al. 2017, Statist. Sci.)

on how to attribute variance to the branches.

The main benefit of the HD priors is their combination of intuitive graphical representation with robust inference.

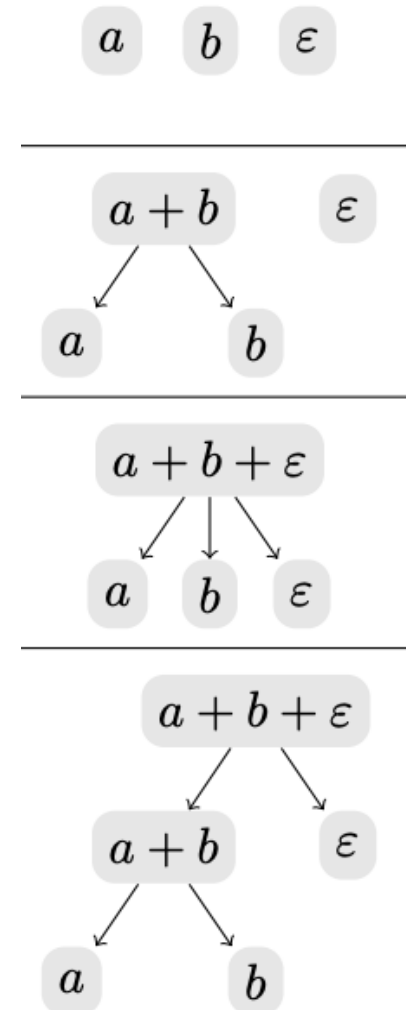
A simple example

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

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

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

Tree structure



The R-package `makemyprior` (Hem et al. 2021, <https://arxiv.org/abs/2105.09712>)

- forces users to be aware of prior distribution choices
- offers a GUI with a guide helping to incorporate the available prior knowledge
- is designed for latent Gaussian models
- supports Gaussian, Poisson and binomial likelihoods
- visualizes the prior distributions for inspection
- allows direct inference via `rstan` (<https://mc-stan.org/users/interfaces/rstan>) and `INLA` (www.r-inla.org).
- available from CRAN:

<https://cran.r-project.org/web/packages/makemyprior>

Modify tree structure

DETACH ATTACH
MERGE REMOVE

Change priors

Prior on split

- ☒ Dirichlet
- ☐ PC prior

No node chosen

- ☒ Jeffreys'
- ☐ PC prior
- ☐ Inverse gamma
- ☐ Half-Cauchy (st.dev.)

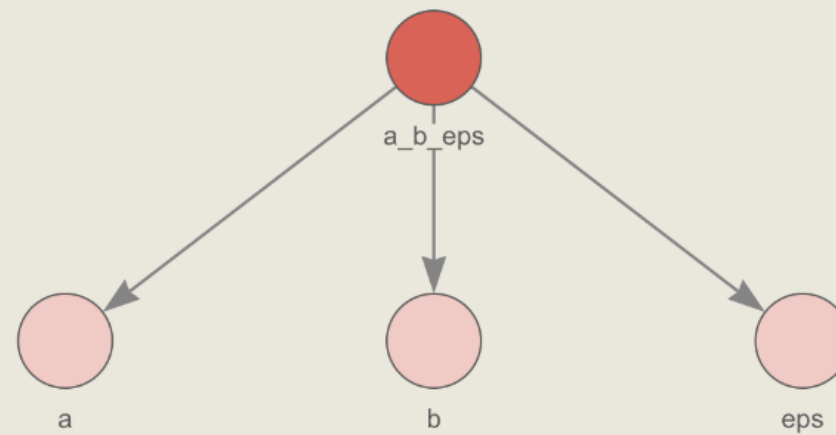
Parameterization

Priors

BEGIN GUIDE

RESET CLOSE

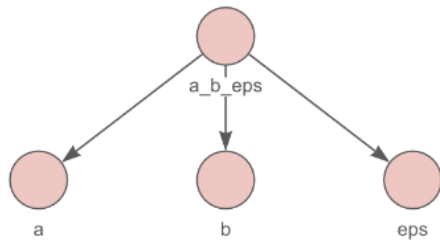
Gaussian likelihood. $y = \mu + x\beta_x + a + b + eps$, $\eta = g(\mu) = \mu$



PLOT PRIORS

Welcome!

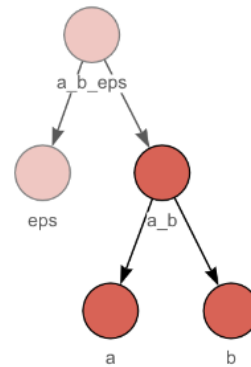
You will now be guided through your prior specification. First you specify the model structure using a node tree, then you choose prior distributions for each model components based on your knowledge about the model, and at last you get to see the resulting prior distributions based on your choices. To quit the guide, click 'Quit guide' at any time (your changes may be lost if you do this before the end).



BEGIN GUIDE

QUIT GUIDE

Step 2: Choose split prior



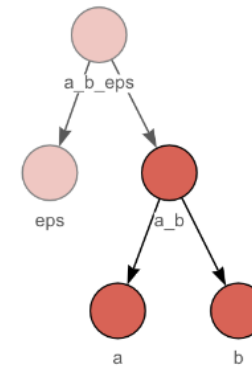
Do you have an intuition on how the variance of a_b is distributed, and do you want to use this knowledge?

YES

NO

QUIT GUIDE

Step 2: Choose split prior



How much of the variance of a_b do you think should go to a ? (This is the same as choosing the **median** of the prior.)

0.7

OK

QUIT GUIDE

Modify model structure

DETACH	ATTACH
MERGE	REMOVE

Change priors

Parameterization

$$V_{a,d,x,eps} = \sigma_a^2 + \sigma_d^2 + \sigma_x^2 + \sigma_{eps}^2$$

$$\omega_{\frac{eps}{a,d,x}} = \frac{\sigma_{eps}^2}{\sigma_a^2 + \sigma_d^2 + \sigma_x^2 + \sigma_{eps}^2}$$

$$\omega_{\frac{a}{a,d,x}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_d^2 + \sigma_x^2}$$

$$\omega_{\frac{d}{d,x}} = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_x^2}$$

☒ Function of variances

Priors

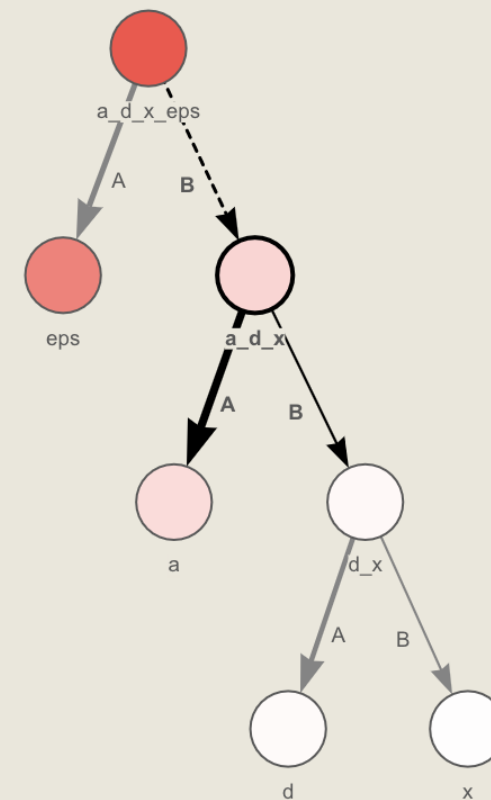
$$V_{a,d,x,eps} \sim \text{Jeffreys}'$$

$$\omega_{\frac{eps}{a,d,x}} \sim \text{PC}_1(0.75)$$

$$\omega_{\frac{a}{a,d,x}} \sim \text{PC}_M(0.85)$$

$$\omega_{\frac{d}{d,x}} \sim \text{PC}_M(0.67)$$

Model equation: $y \sim \mu + a + d + x + \text{eps}$



Chosen node: a_d_x

PLOT PRIORS

Wheat breeding

- **Purpose:** More grain yield (weight of grain from each strain)
- **How:** Find which strains of wheat are genetically best
- **Challenges:**
 - ▶ Separate environmental noise from genetic variance (variation)
 - ▶ Separate non-additive variance from additive and environmental variance
 - ▶ Obtain robust inference

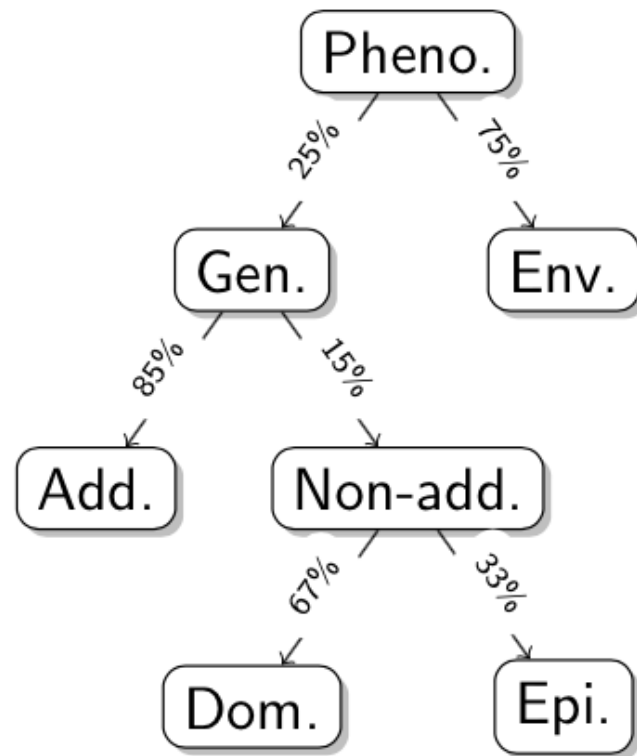
Model

$$y_i = \mu + \underbrace{a_i + d_i + x_i}_{g_i} + e_i, \quad i = 1, \dots, 100$$

- μ : intercept, we use $\mu \sim \mathcal{N}(0, \sigma^2 = 1000)$
- Genetic effects $g_i = a_i + d_i + x_i$:
 - ▶ $\mathbf{a} = (a_1, \dots, a_{100}) \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 \mathbf{V}_a)$: additive
 - ▶ $\mathbf{d} = (d_1, \dots, d_{100}) \sim \mathcal{N}(\mathbf{0}, \sigma_d^2 \mathbf{V}_d)$: dominance
 - ▶ $\mathbf{x} = (x_1, \dots, x_{100}) \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbf{V}_x)$: epistasis
- $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I})$: environmental effect

The covariance matrices $\mathbf{V}_a$, $\mathbf{V}_d$, $\mathbf{V}_x$ come from the SNP matrix with genetic markers for each strain of wheat

Parameterisation and priors (Hem et al. 2021, *Genetics*, 217(3))



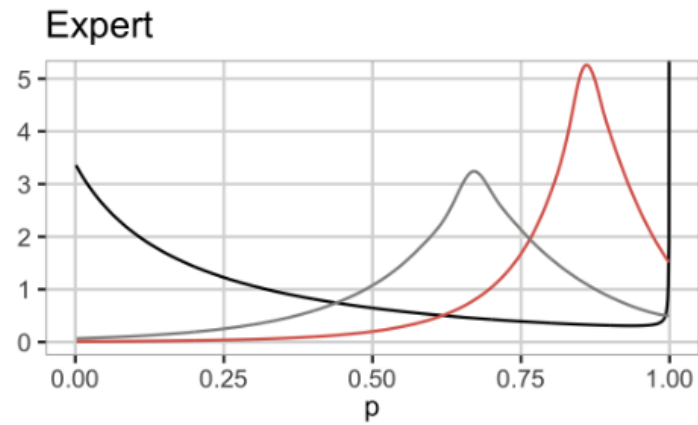
$$V = \sigma_a^2 + \sigma_d^2 + \sigma_x^2 + \sigma_e^2$$

$$p_{\frac{g}{g+e}} = \frac{\sigma_a^2 + \sigma_d^2 + \sigma_x^2}{V} \leftarrow \text{broad-sense heritability}$$

$$p_{\frac{a}{g}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_d^2 + \sigma_x^2}$$

$$p_{\frac{d}{d+x}} = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_x^2}$$

Priors



- Proportion of phenotypic variance to genetic effect (25%)
- Proportion of genetic variance to additive effects (85%)
- Proportion of non-additive variance to dominance effect (67%)

Specifying the model and priors without the GUI

[illegible]

Summary

The R-package `makemyprior` has the goal to make prior choice an active part of the Bayesian workflow. The package:

- makes it easy to include prior knowledge in an intuitive and transparent way
- can be used to verify prior choices
- allows direct inference in a simple way