NTNU
Norwegian University of
Science and Technology

Prior choices: Penalized complexity priors

Andrea Riebler <andrea.riebler@math.ntnu.no>
Department of Mathematical Sciences, NTNU
November 6, 2014

Outline

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 precision or correlation parameters.

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 precision or correlation parameters.

What is the current practice?

About the choice of prior distributions

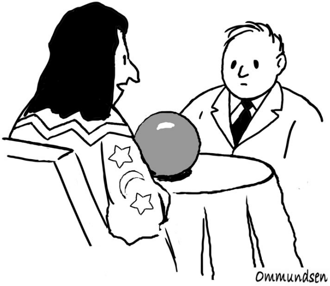
The issue of setting prior distributions on model parameters is a difficult issue in applied Bayesian statistics, in particular for parameters further down the model hierarchy, such as precision or correlation parameters.

What is the current practice?

— Choose priors based on computational convenience.

— Choose priors used in the literature and hope to avoid criticism.

— Ignore the problem and hope that the data will dominate the prior.



"Is this needed for a Bayesian analysis?"

About prior choices

Martins, Simpson, Riebler, Rue and Sørbye (2014)

"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:

Martins, T. G., Simpson, D. P., Riebler, A., Rue, H. and Sørbye, S. H. (2014). Penalising model component complexity: A principled practical approach to constructing priors. arXiv:1403.4630.

Assignment of hyperpriors

The scaling problem of intrinsic model components

— Models for splines (`rw1`, `rw2`)

Assignment of hyperpriors

The scaling problem of intrinsic model components

- Models for splines (`rw1`, `rw2`)
- Thin-plate splines (dimension > 1, `rw2d`)

Assignment of hyperpriors

The scaling problem of intrinsic model components

- Models for splines (`rw1`, `rw2`)
- Thin-plate splines (dimension > 1, `rw2d`)
- The "CAR" model/Besag-model for area/regional models (`besag`)

Assignment of hyperpriors

The scaling problem of intrinsic model components

- Models for splines (`rw1`, `rw2`)
- Thin-plate splines (dimension > 1, `rw2d`)
- The “CAR” model/Besag-model for area/regional models (`besag`)
- and others...

Assignment of hyperpriors

The scaling problem of intrinsic model components

— Models for splines (`rw1`, `rw2`)

— Thin-plate splines (dimension > 1, `rw2d`)

— The “CAR” model/Besag-model for area/regional models (`besag`)

— and others...

Assignment of hyperpriors

The scaling problem of intrinsic model components

- Models for splines (`rw1`, `rw2`)
- Thin-plate splines (dimension > 1, `rw2d`)
- The “CAR” model/Besag-model for area/regional models (`besag`)
- and others...

Problem:

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

Sørbye and Rue, 2014, Spat Stat

Illustration RW1: Marginal variance

Consider the characteristic marginal variance

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

```
> rw1(5)
[1,]  1 -1  .  .  .
[2,] -1  2 -1  .  .
[3,]  . -1  2 -1  .
[4,]  .  . -1  2 -1
[5,]  .  .  . -1  1
> geom.mean(diag(ginv(rw1(5))))
[1] 0.73
> geom.mean(diag(ginv(rw1(50))))
[1] 7.55
> geom.mean(diag(ginv(rw1(500))))
[1] 75.580
```

IGMRFs need to be scaled

That means:

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

$\Rightarrow$ Scale the IGMRF such that $\sigma^2_\tau = 1/\tau$.

In R-INLA

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

How to choose our parameters?

- Assume $\tau \sim \text{Gamma}(a, b)$ where $\text{E}(\tau) = a/b$.
- 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

- This isn't enough: Why are we using a Gamma distribution, why not half-Cauchy . . . ?

Penalised complexity (PC) priors

Martins et al. (2014) 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 for a more complex model.

Our background: `R-INLA`

Building models adding up model components

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

— Many model components represent a flexible extension of a base model.

Our background: `R-INLA`

Building models adding up model components

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

— 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.

Our background: `R-INLA`

Building models adding up model components

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

— 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.

Our background: `R-INLA`

Building models adding up model components

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

- 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 $\boldsymbol{x}$ we define a flexible model

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

  where ξ is interpreted as a flexibility parameter.
- f is a flexible version of a base model

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

Examples for base models

Case	Parameter	ξ	Base model
IID	τ (precision)	$\xi = 1/\tau$	$\xi = 0$ (no random effect)

Examples for base models

Case	Parameter	ξ	Base model
IID	τ (precision)	$\xi = 1/\tau$	$\xi = 0$ (no random effect)
Student-t	ν (dof)	$\xi = 1/\nu$	$\xi = 0$ (Gaussian)

Examples for base models

Case	Parameter	ξ	Base model
IID	τ (precision)	$\xi = 1/\tau$	$\xi = 0$ (no random effect)
Student-t	ν (dof)	$\xi = 1/\nu$	$\xi = 0$ (Gaussian)
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)
Correlation matrix	$\boldsymbol{Q}$	$\boldsymbol{\xi} = \boldsymbol{Q}$	$\boldsymbol{\xi} = \boldsymbol{I}$ (no correlation)

Examples for base models

Case	Parameter	ξ	Base model
IID	τ (precision)	$\xi = 1/\tau$	$\xi = 0$ (no random effect)
Student-t	ν (dof)	$\xi = 1/\nu$	$\xi = 0$ (Gaussian)
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)
Correlation matrix	$\boldsymbol{Q}$	$\boldsymbol{\xi} = \boldsymbol{Q}$	$\boldsymbol{\xi} = \boldsymbol{I}$ (no correlation)

Side comment: In a BYM model we would have nested base models:

Base model $= 0 \rightarrow$ iid $\rightarrow$ dependence $=$ more flexible model

1. Occams razor

— The prior for $\xi \geq 0$ should penalise the complexity introduced by ξ

1. Occams razor

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

1. Occams razor

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

1. Occams razor

- The prior for $\xi \geq 0$ should penalise the complexity introduced by ξ
- The prior should be decaying with increasing measure by the 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 $\xi > 0$,

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

for flexible model f and base model g.

2. Principle: Measure of complexity

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

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

for flexible model f and base model g.

Example

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

$$\mathrm{KLD}(f \| g) = \frac{\xi^2}{2}$$

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"

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

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"

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

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

$$\pi(d(\xi)) = \lambda \exp\left(-\lambda d(\xi)\right), \qquad \lambda > 0$$

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

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”

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

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

$$\pi(d(\xi)) = \lambda \exp\left(-\lambda d(\xi)\right), \qquad \lambda > 0$$

  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.

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.

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):

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

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

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

Example: Precision of a Gaussian

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

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

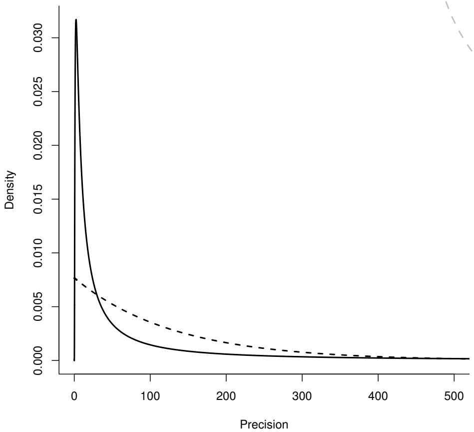
where $\mathrm{Prob}(\sigma > U) = \alpha$ gives

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

Alternative interpretation

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

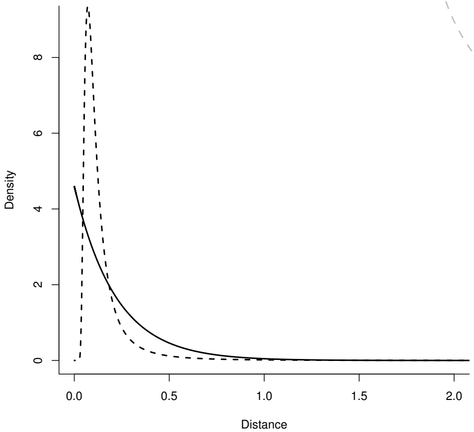
Comparison to a similar gamma prior



PC-prior with $U = 0.3/0.31$, $\alpha = 0.01$ (solid).

Gamma prior with shape 1 and rate a, with $a = 0.0076$, to get same marginal variance (dashed).

Comparison to a similar gamma prior



PC-prior with $U = 0.3/0.31$, $\alpha = 0.01$ (solid).

Gamma prior with shape 1 and rate a, with $a = 0.0076$, to get same marginal variance (dashed).

How to do this in INLA

Specifying the pc-prior in the f-function:

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

Documentation:

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

Discussion: PC priors

— The new principled constructive approach to construct priors seems very promising.

Discussion: PC priors

- The new principled constructive approach to construct priors seems very promising.
- Easy and very natural interpretation + a well defined shrinkage.

Discussion: PC priors

- The new principled constructive approach to construct priors seems very promising.
- Easy and very natural interpretation $+$ a well defined shrinkage.
- We can chose the degree of “informativeness”.

Discussion: PC priors

— The new principled constructive approach to construct priors seems very promising.

— Easy and very natural interpretation + a well defined shrinkage.

— We can chose the degree of “informativeness”.

— Exciting extensions will grow out this (not discussed)

Discussion: PC priors

- The new principled constructive approach to construct priors seems very promising.
- Easy and very natural interpretation $+$ a well defined shrinkage.
- We can chose the degree of “informativeness”.
- Exciting extensions will grow out this (not discussed)
- Not all cases are easy...

Discussion: PC priors

— The new principled constructive approach to construct priors seems very promising.

— Easy and very natural interpretation + a well defined shrinkage.

— We can chose the degree of “informativeness”.

— Exciting extensions will grow out this (not discussed)

— Not all cases are easy...

— A lot of work to integrate this into `R-INLA`

Other (theoretical) things...

— Good large-sample behaviour (via BvM theorem)

— Very good risk results in Stein-type situations

— Strong links to shrinkage priors, although you may consider a heavier tail...

Thank you for your attention!

Thanks goes:

— to you for coming.

Thank you for your attention!

Thanks goes:

— to you for coming.

— to the whole INLA + friends team.

Thank you for your attention!

Thanks goes:

- to you for coming.
- to the whole INLA + friends team.

Thank you for your attention!

Thanks goes:

— to you for coming.

— to the whole INLA + friends team.

If you have any doubts or questions, please write me:
andrea.riebler@math.ntnu.no

Thank you for your attention!

Thanks goes:

— to you for coming.

— to the whole INLA + friends team.

If you have any doubts or questions, please write me:

andrea.riebler@math.ntnu.no