Session 1.2: Hierarchical Models, Priors, Prediction and Model Checking

VIBASS, University of Valencia

20 July 2022

Learning Objectives

After this session you should be able to:

- Understand the different modelling assumptions for hierarchical data
- Be able to specify a hierarchical model for Poisson data
- Be able to perform prediction in a Bayesian approach
- Distinguish and choose between several prior distributions for the precision/variance parameter
- Use the DIC/WAIC as tools for model selection.

The topics treated in this lecture are covered in Chapter 5 of the book **Spatial and Spatio-Temporal Bayesian models with R-INLA**

Outline

1. What are hierarchical models
2. Different modelling assumptions
3. Parameter interpretation
4. Hierarchical regression
5. Prediction
6. Choice of prior
7. Model selection

What are hierarchical models

What are hierarchical models?

Hierarchical model is a very broad term that refers to wide range of model set-ups

- Multilevel models
- Random effects models
- Random coefficient models
- Variance-component models
- Mixed effect models

Key feature: Hierarchical models are statistical models that provide a formal framework for analysis with a complexity of structure that matches the system being studied.

The hierarchical approach

- Attempt to capture (model) and understand the structure of the data

The hierarchical approach

- Attempt to capture (model) and understand the structure of the data
- Is flexible:
  - all sources of correlation and heterogeneity can be incorporated in a modular fashion, in particular by the introduction of unit-specific parameters
  - can be combined with other types of models, e.g. for missing data or measurement error

The hierarchical approach

- Attempt to capture (model) and understand the structure of the data
- Is flexible:
  – all sources of correlation and heterogeneity can be incorporated in a modular fashion, in particular by the introduction of unit-specific parameters
  – can be combined with other types of models, e.g. for missing data or measurement error
- We wish to make inference on models with many parameters $(\lambda_1, \ldots, \lambda_N)$ measured on N units (individuals, areas, time-points, etc.) which are related or connected by the structure of the problem.

The hierarchical approach

- Attempt to capture (model) and understand the structure of the data
- Is flexible:
  – all sources of correlation and heterogeneity can be incorporated in a modular fashion, in particular by the introduction of unit-specific parameters
  – can be combined with other types of models, e.g. for missing data or measurement error
- We wish to make inference on models with many parameters $(\lambda_1, \ldots, \lambda_N)$ measured on N units (individuals, areas, time-points, etc.) which are related or connected by the structure of the problem.
- Unit specific parameters will borrow strength from corresponding parameters associated with the other units

Motivating example: Disease mapping

- To summarise spatial and spatio-temporal variation in disease risk
- **Question**: Which areas have particularly high or low disease rates?
- **Question**: Can we explain some of the variation in disease rates by area-level covariates?

Motivating example: Disease mapping

- To summarise spatial and spatio-temporal variation in disease risk
- **Question**: Which areas have particularly high or low disease rates?
- **Question**: Can we explain some of the variation in disease rates by area-level covariates?
- Data are the observed (y_i) and expected number of cases in area i: $E_i = \sum_k n_{ik} r_k$, where r_k reference rate for stratum k (age, sex,...)
- Rare disease and/or small areas: Poisson framework

$$y_i \sim \text{Poisson}(\rho_i E_i)$$

where ρ_i is the **unknown RR** in area i

Non smoothed estimates of the RR

$$\text{SMR}_i = \frac{y_i}{E_i}$$

$$\hat{\text{Var}}(\text{SMR}_i) = \frac{y_i}{E_i^2}$$

Motivating example: Disease mapping

- To summarise spatial and spatio-temporal variation in disease risk
- **Question**: Which areas have particularly high or low disease rates?
- **Question**: Can we explain some of the variation in disease rates by area-level covariates?
- Data are the observed (y_i) and expected number of cases in area i: $E_i = \sum_k n_{ik} r_k$, where r_k reference rate for stratum k (age, sex,...)
- Rare disease and/or small areas: Poisson framework

$$y_i \sim \text{Poisson}(\rho_i E_i)$$

where ρ_i is the **unknown RR** in area i

Non smoothed estimates of the RR

$$\text{SMR}_i = \frac{y_i}{E_i}$$

$$\hat{\text{Var}}(\text{SMR}_i) = \frac{y_i}{E_i^2}$$

- very imprecise: areas with small E_i have high associated variance
- estimated independently: makes no use of risk estimates in other areas of the map

Motivating example: Disease mapping

Example:

- observed cases of lip cancer y_i diagnosed in Scotland in 1975-1980 at county level $i = 1, \ldots, 56$ areas
- expected number of cases E_i are also available using age/sex standardised reference rates and population counts:

Motivating example: Disease mapping

Example:

- observed cases of lip cancer y_i diagnosed in Scotland in 1975-1980 at county level $i = 1, \ldots, 56$ areas
- expected number of cases E_i are also available using age/sex standardised reference rates and population counts:

Assume a Poisson likelihood for the disease counts in each area:

$$y_i \sim \text{Poisson}(\lambda_i) \qquad \lambda_i = \rho_i E_i \qquad i = 1, \ldots, 56$$

- We have 56 parameters ρ_i (one for each area). What prior do we specify on ρ_i?

Modelling assumptions

Different modelling assumptions

Identical parameters

- Assume $\rho_i = \rho$

$\rightsquigarrow$ all the data can be pooled and the individual areas ignored.

- Assume a prior $\rho \sim \text{Gamma}(1, 1)$

$\rightsquigarrow$ conjugate prior

Different modelling assumptions

Identical parameters

- Assume $\rho_i = \rho$

$\rightsquigarrow$ all the data can be pooled and the individual areas ignored.

- Assume a prior $\rho \sim \text{Gamma}(1, 1)$

$\rightsquigarrow$ conjugate prior

- One parameter generates all the observations
- Very easy to implement as it is conjugate (no need for INLA) and all the data are pooled to produce one estimate of the parameter of interest
- Can be unrealistic (it does not take into account differences in the areas)

Different modelling assumptions

Independent parameters

- All the ρ_i are unrelated, meaning that the areas are analysed independently
- Assume a prior $\rho_i \sim \text{Gamma}(1, 1); \qquad i = 1, \ldots, 56$

$\rightsquigarrow$ individual estimates of ρ_i are likely to be highly variable (unless very large sample sizes)

Different modelling assumptions

Independent parameters

- All the ρ_i are unrelated, meaning that the areas are analysed independently
- Assume a prior $\rho_i \sim \text{Gamma}(1, 1); \qquad i = 1, \ldots, 56$

$\rightsquigarrow$ individual estimates of ρ_i are likely to be highly variable (unless very large sample sizes)

- Every area is treated separately (No exchange of information between these). Estimates close to SMR $(\rho_i \approx y_i/E_i)$.
- Again no need for INLA, conjugacy can be exploited.

Different modelling assumptions

Similar (exchangeable) parameters

- All the ρ_i are assumed to be *similar*

$\rightsquigarrow$ they come from the same distribution (are generated by the same parameters)

- Assume a hierarchical prior $\rho_i \sim \text{Gamma}(a, b)$

where a and b are unknown parameters and need to be estimated.

Different modelling assumptions

Similar (exchangeable) parameters

- All the ρ_i are assumed to be *similar*

$\rightsquigarrow$ they come from the same distribution (are generated by the same parameters)

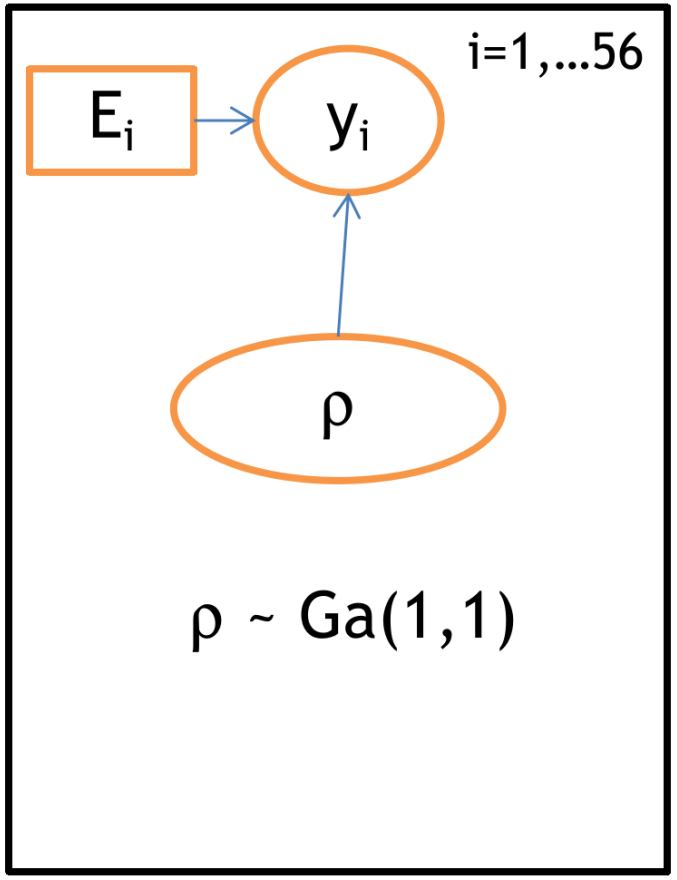
- Assume a hierarchical prior $\rho_i \sim \text{Gamma}(a, b)$

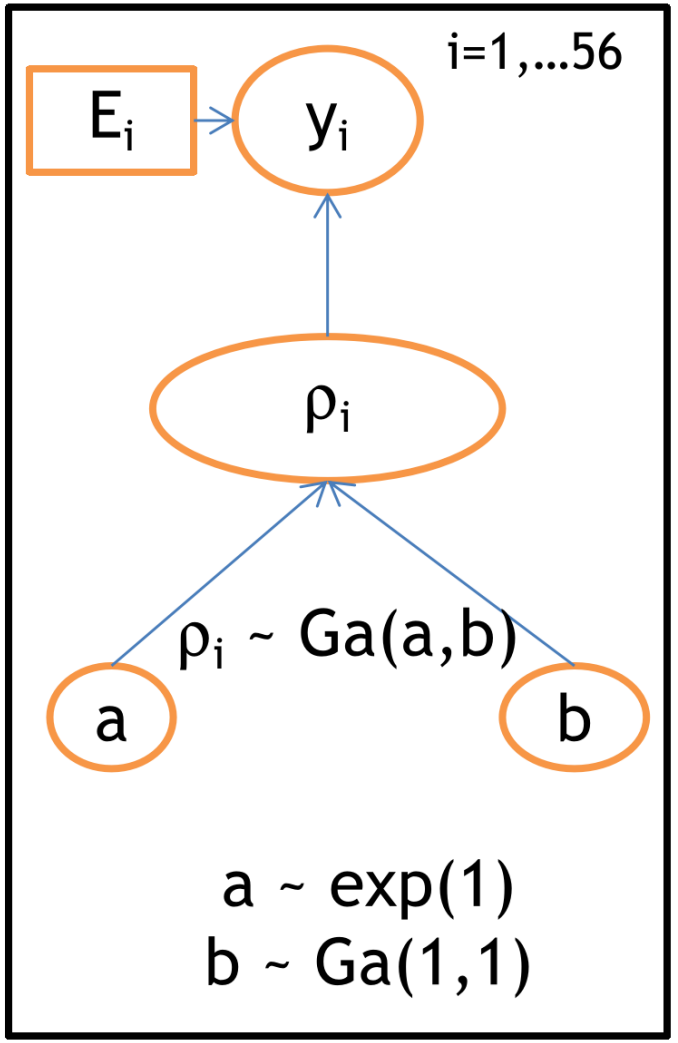
where a and b are unknown parameters and need to be estimated.

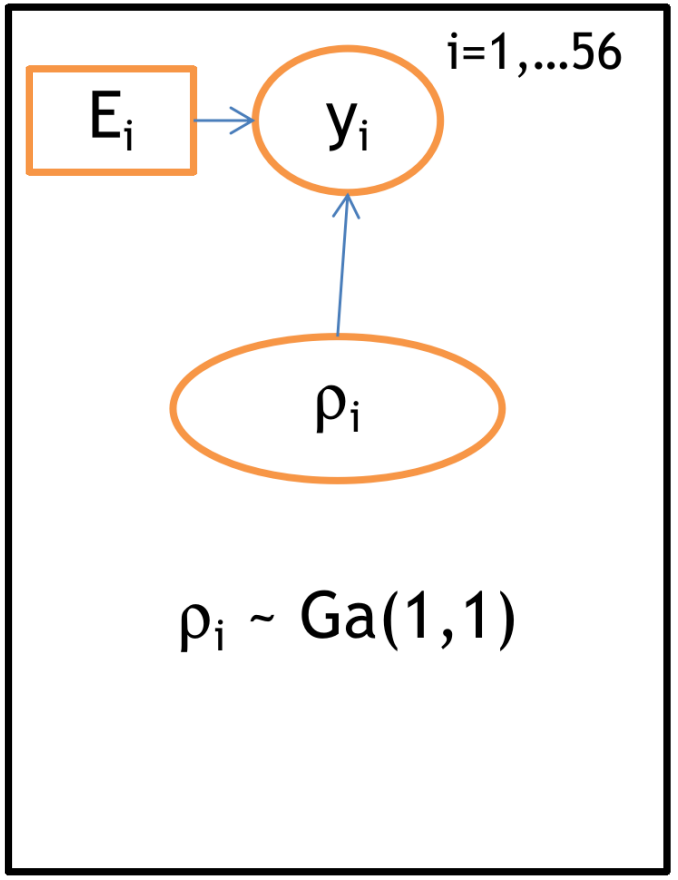
- Different levels of analysis
- Allow the exchange of information between different levels as they are all connected to each other
- Assign hyperprior distribution to a and b, for instance

$$a \sim \text{Exp}(1); b \sim \text{Gamma}(1, 1)$$

Graphical representation of lip cancer hierarchical model







A more flexible hierarchical prior for the relative risks

- A gamma random effect prior for the ρ_i is mathematically convenient, but might be restrictive:
  - Covariate adjustment is difficult
  - Not possible to allow for spatial correlation between risks in nearby areas

A more flexible hierarchical prior for the relative risks

- A gamma random effect prior for the ρ_i is mathematically convenient, but might be restrictive:
  - Covariate adjustment is difficult
  - Not possible to allow for spatial correlation between risks in nearby areas
  - A Normal random effect prior on the $\log \rho_i$ is more flexible:

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i = \rho_i E_i) \\
\eta_i &= \log \rho_i = b_0 + v_i \\
v_i &\sim \text{Normal}(0, \sigma_v^2)
\end{aligned}
$$

A more flexible hierarchical prior for the relative risks

- A gamma random effect prior for the ρ_i is mathematically convenient, but might be restrictive:
  - Covariate adjustment is difficult
  - Not possible to allow for spatial correlation between risks in nearby areas
  - A Normal random effect prior on the logρ_i is more flexible:

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i = \rho_i E_i) \\
\eta_i &= \log \rho_i = b_0 + v_i \\
v_i &\sim \text{Normal}(0, \sigma_v^2)
\end{aligned}
$$

- Need to specify hyperprior distributions for:
- σ_v^2 (between-area variance), e.g. $1/\sigma_v^2 \sim \text{Gamma}(1, 0.001)$
- b_0 (mean log relative risk), e.g. $b_0 \sim \text{Normal}(0, 0.0001)$

A more flexible hierarchical prior for the relative risks

- A gamma random effect prior for the ρ_i is mathematically convenient, but might be restrictive:
  - Covariate adjustment is difficult
  - Not possible to allow for spatial correlation between risks in nearby areas
  - A Normal random effect prior on the $\log\rho_i$ is more flexible:

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i = \rho_i E_i) \\
\eta_i &= \log\rho_i = b_0 + v_i \\
v_i &\sim \text{Normal}(0, \sigma_v^2)
\end{aligned}
$$

- Need to specify hyperprior distributions for:
- σ_v^2 (between-area variance), e.g. $1/\sigma_v^2 \sim \text{Gamma}(1, 0.001)$
- b_0 (mean log relative risk), e.g. $b_0 \sim \text{Normal}(0, 0.0001)$

Advantages of this approach:
Posterior for each v_i

- *borrows strength* from the likelihood contributions of **all** the areas, via their joint influence on the estimate of the unknown population (prior) parameter σ_v^2

→ *global smoothing* of the area RR

→ reflects our *full uncertainty* about the true values of σ_v^2

Interpretation

Parameter interpretation and useful quantities

- v_i are the random effects. It can also be seen as the latent variable which captures the effect of unknown or unmeasured area level covariates.
- If area level covariates are spatially structured we should take this into account when modelling v_i (we will see it later)
- $\exp(v_i)$ relative risk in area i compared to the risk for the whole study region
- The variance of the random effects σ_v^2 reflect the amount of extra-Poisson variation in the data

Parameter interpretation and useful quantities

- v_i are the random effects. It can also be seen as the latent variable which captures the effect of unknown or unmeasured area level covariates.
- If area level covariates are spatially structured we should take this into account when modelling v_i (we will see it later)
- $\exp(v_i)$ relative risk in area i compared to the risk for the whole study region
- The variance of the random effects σ_v^2 reflect the amount of extra-Poisson variation in the data
- A useful summary of among unit variability in a Poisson hierarchical model is to rank the random effects and calculate the difference between two units at opposite extremes
- Suppose we consider the 5^{th} and 95^{th} percentiles of the area relative risk distribution
- let $q_{5\%} = \lambda_{5\%}$ denote the log relative risk of outcome for the area ranked at the 5^{th} percentile
- let $q_{95\%} = \lambda_{95\%}$ denote the log relative risk of outcome for the area ranked at the 95^{th} percentile

Quantile ratio

$$\text{QR}_{90} = \exp(q_{95\%} - q_{5\%})$$

is the relative risk of outcome between the top and bottom 5% of areas

Lip cancer dataset

```
> LipCancer <- read.csv("scotlip.csv")
> LipCancer
```

```
# A tibble: 6 × 11
  CODENO       AREA PERIMETER RECORD_ID DISTRICT NAME          CODE      y    POP     E     x
   <int>      <dbl>     <dbl>     <int>    <int> <chr>         <chr> <int>  <int> <dbl> <int>
1   6126  974002000    184951         1        1 Skye-Lochalsh w6126     9  28324  1.38    16
2   6016 1461990000    178224         2        2 Banff-Buchan  w6016    39 231337  8.66    16
3   6121 1753090000    179177         3        3 Caithness     w6121    11  83190  3.04    10
4   5601  898599000    128777         4        4 Berwickshire  w5601     9  51710  2.53    24
5   6125 5109870000    580792         5        5 Ross-Cromarty w6125    15 129271  4.26    10
6   6554  422639000    118433         6        6 Okney         w6554     8  53199  2.4     24
```

- `DISTRICT` identifies the area
- `y` identifies the counts of cancer cases
- `E` identifies the expected cases of cancer using the entire region under study as reference
- `x` identifies the exposure to sun (percentage of agriculture , farming and fishery works)

In R-INLA

We first populate the `formula` environment

```
> formula.inla <- y ~ 1 +
+   f(RECORD_ID,model="iid", hyper=list(prec=list(prior="loggamma",
+       param=c(1,0.01))))
```

- The model specification is exactly the same as in GLM;
- Anything with `f(.)` specifies a random effect; in this case `iid` represents the exchangeable structure.

Then we run the model through

```
> lipcancer.poisson <- inla(formula.inla,family="poisson",
+                           data=LipCancer, E=E,
+                           control.predictor=list(compute=TRUE),
+                           control.compute=list(config=TRUE),
+                           control.fixed=list(mean.intercept=0,prec.intercept=0.00001))
```

Note that

- `control.fixed` allows to specify the parameters of the prior for the fixed effects (intercept)
- `control.predictor` tells `INLA` to include the linear predictor estimation (the parameters of the prior for the fixed effects (intercept)) useful for prediction - see later)
- `control.compute` allows to include models election indexes, as well as to draw samples from the joint posterior

Results for lip cancer in Scotland example

- exp (v_i) is the relative risk of lip cancer in area i relative to average across Scotland (see map)
- σ_v is the between-area standard deviation of log relative risk of lip cancer
- As in INLA we get the precision we need to convert it into variance using

```
> sigma2.v<- inla.tmarginal(function(x) sqrt(1/x),
+             lipcancer.poisson$marginals.hyperpar[[1]])
```

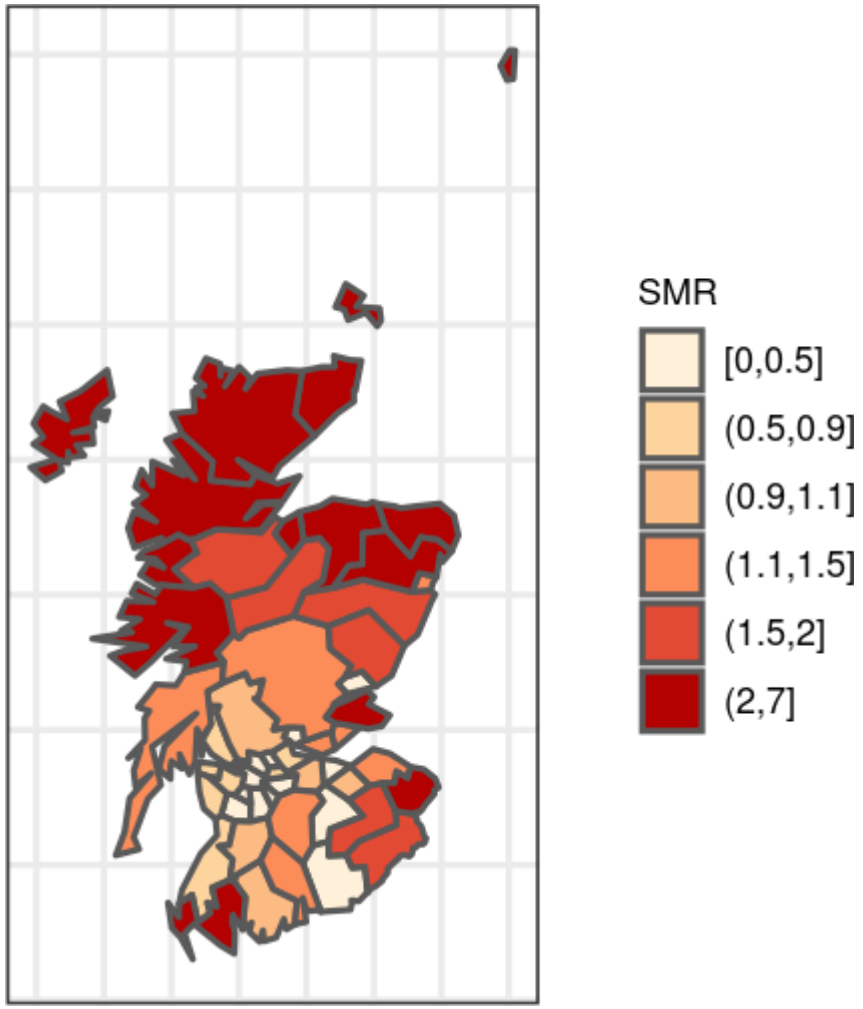
And we can calculate quintiles with

```
> inla.qmarginal(seq(0,1,0.2),sigma2.v)
```

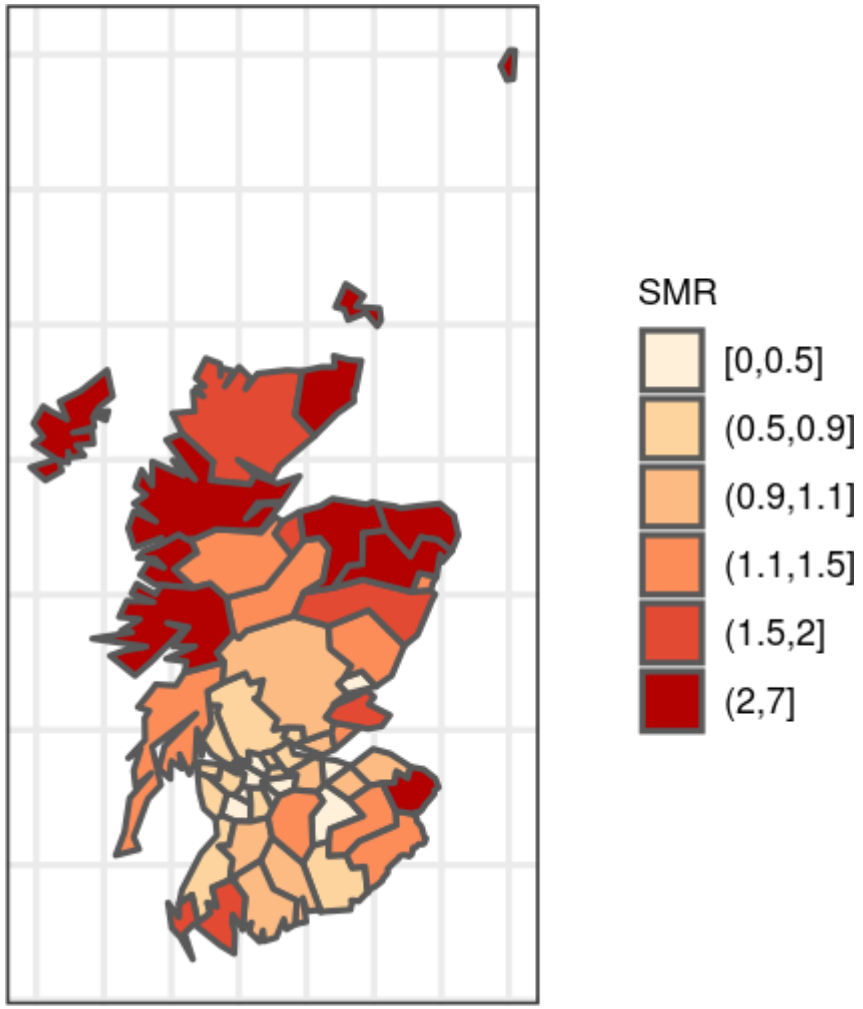
```
[1] 0.4996176 0.6744936 0.7257159 0.7735114 0.8337035 1.1555258
```

Maps: comparing SMR with smoothed estimates

SMR



Posterior mean



Quantile ratios

To obtain the quantile ratio we need to follow these steps:

1. Obtain the **join posterior distribution** for the model under consideration

```
> joint.post <- inla.posterior.sample(100,lipcancer.poisson)
> names(joint.post[[1]])
```

```
[1] "hyperpar" "latent"   "logdens"
```

```
> joint.post[[1]]$latent[1:3,]
```

```
Predictor:1 Predictor:2 Predictor:3
   1.402447    1.617122    1.218793
```

Note that:

- `joint.post` is a list of 100 elements and each element includes a value from

1. the joint posterior distribution for the hyperparameters `joint.post\$hyperpar`
2. joint posterior distribution for the linear predictor $\boldsymbol{\eta}$ in `joint.post\$latent` (row 1 to N)
3. joint posterior distribution for the random effects $\boldsymbol{v}$ in `joint.post\$latent` (N +1 to 2N)

Quantile ratios

2. For each iteration rank the areas based on their v_i values

```
> joint.v <- matrix(NA,56,100)
> for(i in 1:100){
+   joint.v[,i]<- joint.post[[i]]$latent[57:112]
+ }
```

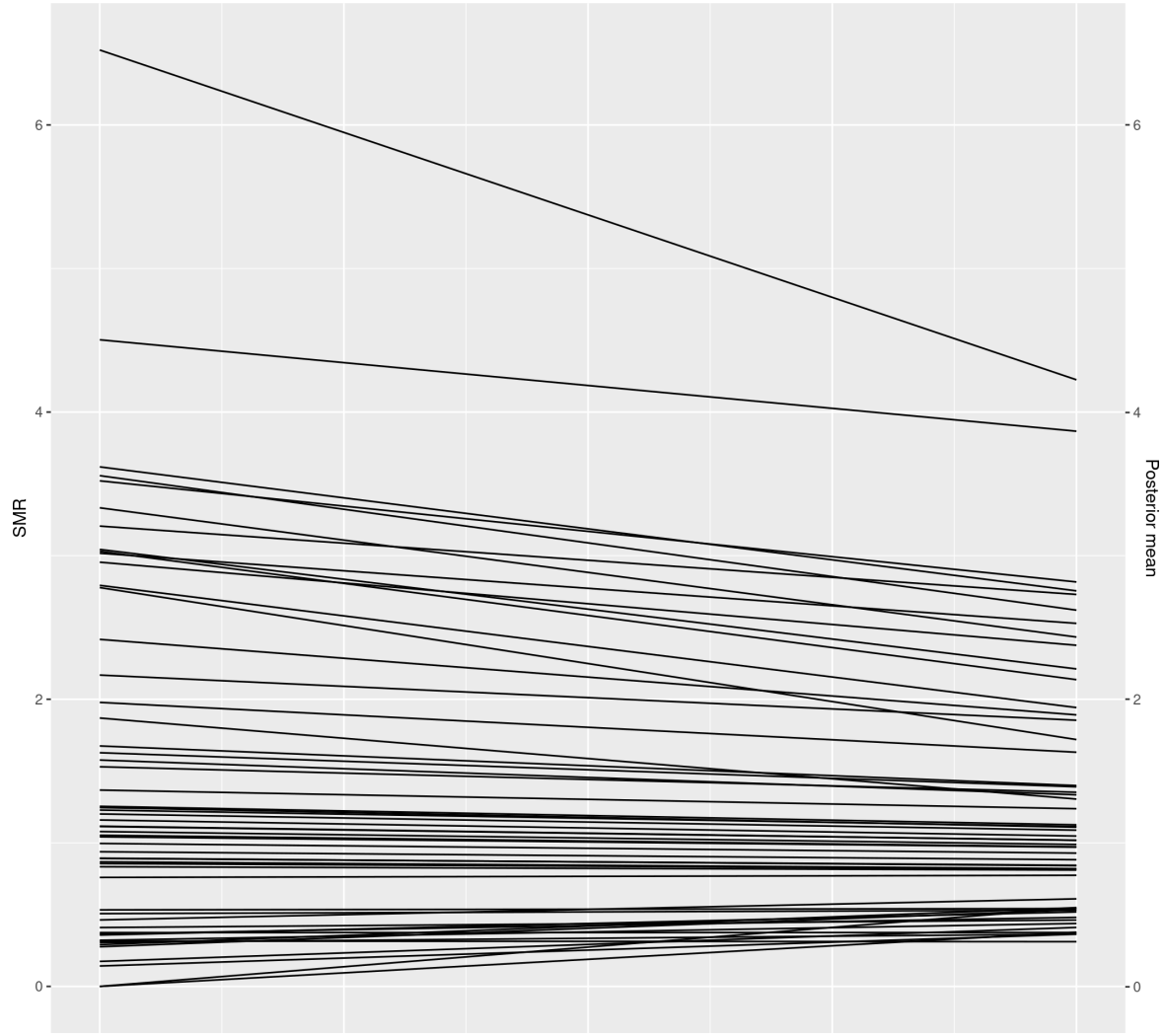
- Calculate v_3 and v_{53} (5% and 95%) and build the ratio

```
> v5perc <- apply(joint.v,2, function(x) quantile(x,0.05))
> v95perc <- apply(joint.v,2, function(x) quantile(x,0.95))
> QR90<- mean(exp(v95perc-v5perc))
> QR90
```

```
[1] 10.79859
```

- The $QR90$ points towards a large spatial variability.

SMR versus posterior mean RR for selected areas



- Comparing the SMR and the area level posterior mean from the model shows a shrinkage towards the global (national mean)

Hierarchical Regression

Regression in INLA

It is easy to move from hierarchical models to regression models with random effects.

Example: In the Seeds dataset we are interested in the proportion of seeds that germinated on each of 21 plates arranged according to a 2 by 2 factorial layout by seed and type of root extract. The data consider the number of germinated y_i and the total number of seeds n_i on the i $-th$ plate, $i = 1, \ldots, 21$.

Regression in INLA

It is easy to move from hierarchical models to regression models with random effects.

Example: In the Seeds dataset we are interested in the proportion of seeds that germinated on each of 21 plates arranged according to a 2 by 2 factorial layout by seed and type of root extract. The data consider the number of germinated y_i and the total number of seeds n_i on the i $-th$ plate, $i = 1, \ldots, 21$.

We specify a random effect logistic model

$$\begin{aligned} y_i &\sim \text{Binomial}(\pi_i, n_i) \\ \text{logit}(\pi_i) &= b_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + v_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \end{aligned}$$

where x_{1i} , x_{2i} are the seed type and root extract of the i$-th$ plate, and an interaction term $\beta_{12} x_{1i} x_{2i}$ is included. b_0 , β_1 , β_2 , β_{12} , σ_v^2 are given independent "noninformative" priors.

R-INLA code

```
> data(Seeds)
> head(Seeds)
```

```
   r  n x1 x2 plate
1 10 39  0  0     1
2 23 62  0  0     2
3 23 81  0  0     3
4 26 51  0  0     4
5 17 39  0  0     5
6  5  6  0  1     6
```

```
> formula <- r~x1 + x2 + x1*x2 + f(plate, model="iid")
> model.regression <- inla(formula, data=Seeds,
+                          family="binomial", Ntrials=n)
```

Output: Parameters

```
> model.regression$summary.fixed
```

```
                  mean        sd 0.025quant   0.5quant 0.975quant       mode          kld
(Intercept) -0.5573106 0.1290580 -0.8128318 -0.5566128 -0.3057575 -0.5551229 8.141460e-05
x1           0.1432173 0.2272850 -0.3066092  0.1443640  0.5862926  0.1463687 6.017888e-05
x2           1.3214742 0.1819023  0.9680552  1.3202582  1.6820314  1.3180323 9.645924e-05
x1:x2       -0.7815996 0.3120993 -1.3948849 -0.7814952 -0.1691147 -0.7814959 4.969306e-05
```

```
> head(model.regression$summary.random$plate)
```

```
  ID          mean         sd  0.025quant      0.5quant 0.975quant          mode        kld
1  1 -0.0103355441 0.06211783 -0.17943488 -8.454498e-04 0.04377829 -1.670834e-04 0.03280257
2  2  0.0005949997 0.04749199 -0.07794005 -2.291544e-05 0.08598236  2.026897e-05 0.02229568
3  3 -0.0123824254 0.06238287 -0.19682905 -1.137677e-03 0.03712337 -3.282143e-05 0.02335741
4  4  0.0158317164 0.07312171 -0.03334043  1.441999e-03 0.24192536  3.201181e-04 0.02754597
5  5  0.0063046878 0.05407668 -0.05320428  4.815531e-04 0.13605575  3.543913e-04 0.02878329
6  6  0.0026606361 0.05648353 -0.07252901  3.664692e-04 0.10769889 -4.851865e-05 0.04532726
```

Prediction

Predictive distribution

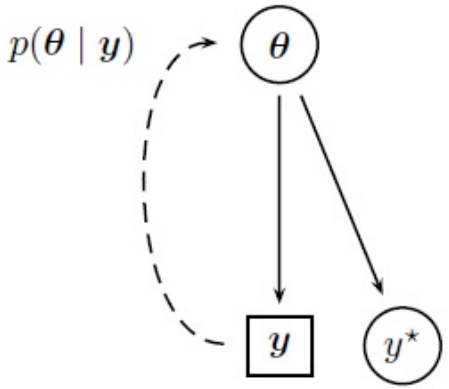
- An important consequence of the concept of exchangeability is that we can derive also a predictive result on the dependent variable
- Assume that $y^\star$ represents a future occurrence of y. If $\boldsymbol{y}$ and $y^\star$ are exchangeable, we then have that:

$$
\begin{aligned}
p(y^\star \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y}, y^\star)}{p(\boldsymbol{y})} && \text{from the conditional probability} \\
&= \frac{\int p(y^\star \mid \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{p(\boldsymbol{y})} && \text{by exchangeability} \\
&= \frac{\int p(y^\star \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) p(\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}}{p(\boldsymbol{y})} && \text{applying Bayes' Theorem} \\
&= \int p(y^\star \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}
\end{aligned}
$$

- Following the INLA notation $\boldsymbol{\theta}$ identifies the vector of all the parameters.

Predictive distribution

- The quantity $p(y^\star \mid \boldsymbol{y})$, known as *predictive distribution*, is only meaningful within the Bayesian approach
  $\rightarrow$ the posterior distribution for $\boldsymbol{\theta}$ only exists if $\boldsymbol{\theta}$ are random variables.



$$p(y^\star \mid \boldsymbol{y}) = \int p(y^\star \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}$$

- y and $y^\star$ are generated by the same random process governed by the parameters $\boldsymbol{\theta}$, associated with a suitable prior distribution, $p(\boldsymbol{\theta})$.
- When we observe the value $\boldsymbol{y}$, the uncertainty about the parameter is updated into the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, which in turns is used to infer about the future realization $y^\star$.

Example: Prediction of Missing data

- We assume that the first observation in the Seeds dataset is missing

- To predict it we simply run `INLA` with the option `control.predictor=list(link=link)` where `link` is a vector of the length equal to the number of observations with 1 only where the observation is missing

```
> link<- rep(NA, length(Seeds$r))
> link[is.na(Seeds$r)]<-1
> formula <- r~x1 + x2 + x1*x2 + f(plate, model="iid")
> model.regression <- inla(formula, data=Seeds,
+                          family="binomial", Ntrials=n, control.predictor=list(link=link))
```

Example: Prediction of missing data

- The summary statistics of the predicted values can be accessed by

```
> dim(model.regression$summary.fitted.values)
```

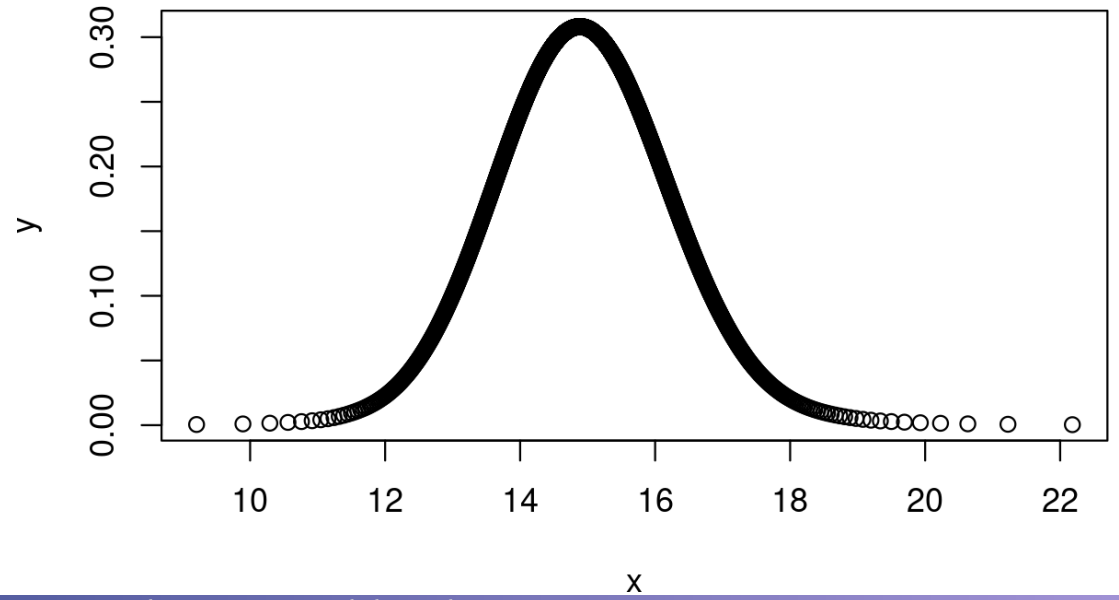
```
[1] 21  6
```

```
> model.regression$summary.fitted.values[1,]
```

```
                          mean          sd 0.025quant  0.5quant 0.975quant      mode
fitted.Predictor.01 0.3831032 0.03519607  0.3168842 0.3824702  0.4528834 0.3816362
```

- Note that we get a distribution for each of the 21 observations, but we need to consider only the first as this was the missing one
- The fitted value is on the probability scale - to go back to the scale of the observations we run

```
> pred.values <- inla.tmarginal(function(x) x*See
+                               model.regression$ma
```



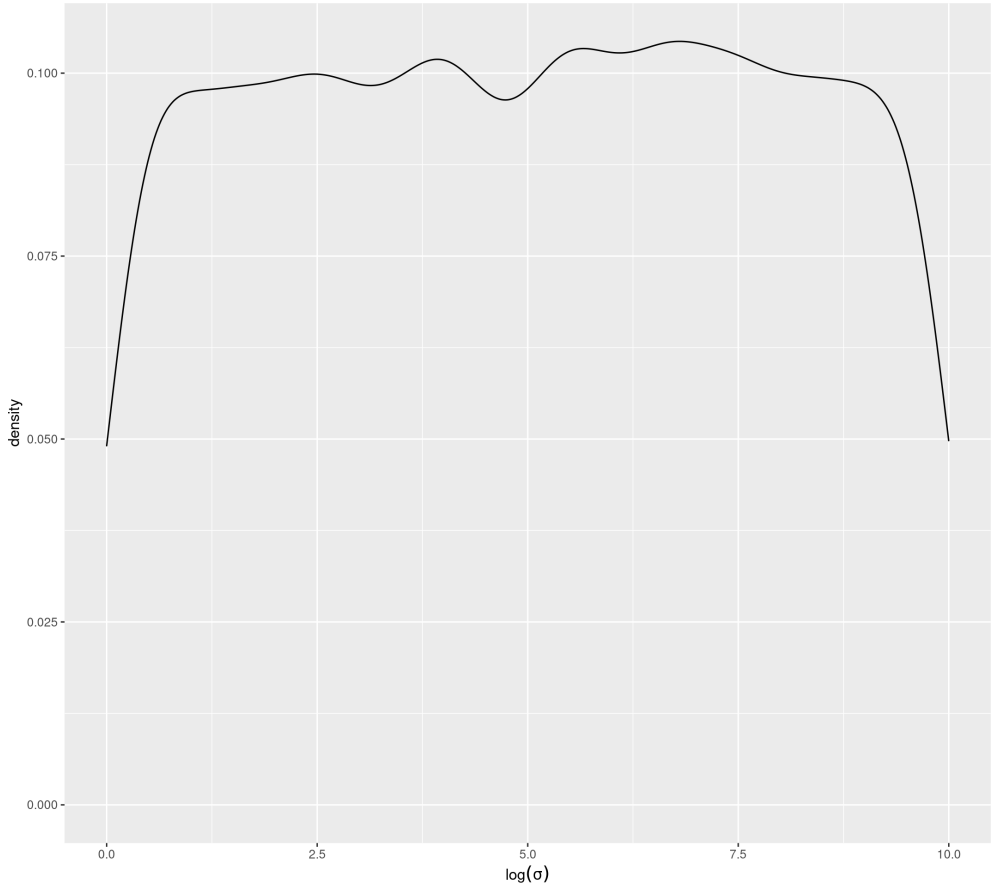
Choice of prior

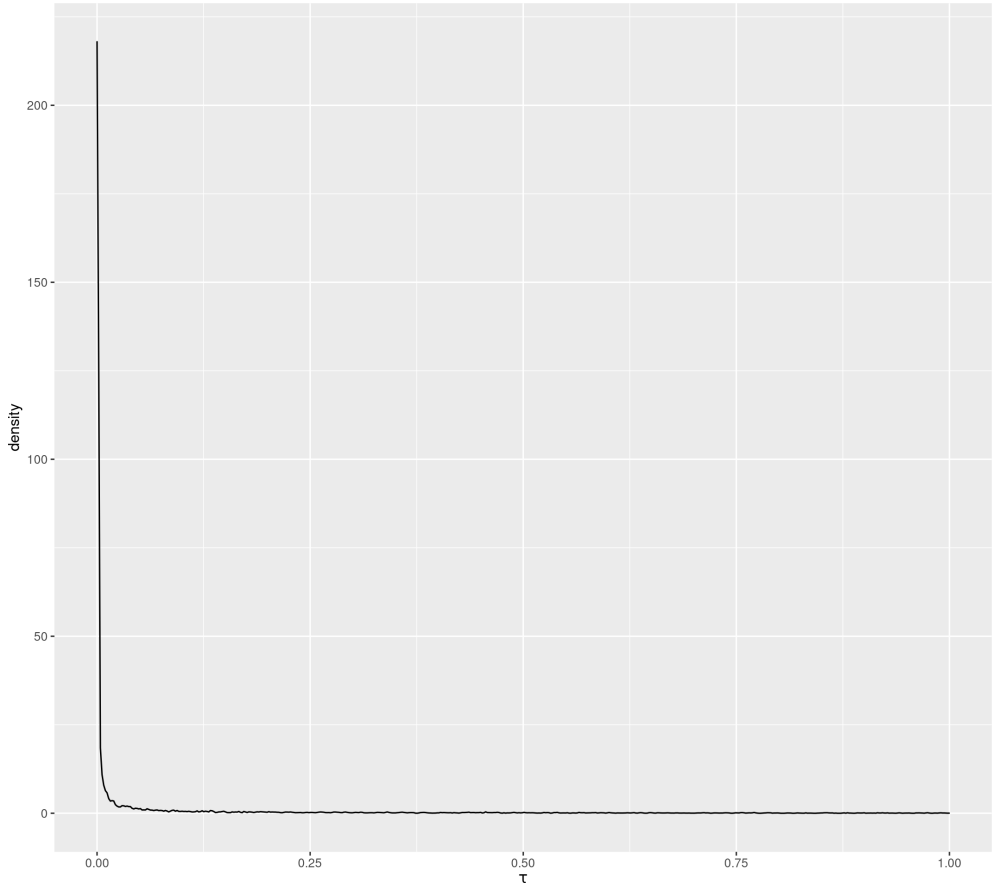
How to specify priors?

- In small area studies we usually work with Poisson/Binomial distribution on data - no variance parameter; the main interest is on random effect variance.
- A Gamma (ϵ, ϵ) can be used on the precision - nice conjugacy property with the Normal distribution of the random effects - but inference could be sensitive to choice of ϵ - particularly if little evidence of heterogeneity between areas is present in the data. It has also been criticised (e.g. (Gelman, 2006))
- A vague or weakly informative prior can be specified so that all possible values are assumed to be a priori equally likely
- Unfortunately, "non informative" prior distributions are sensitive to changes of scale

Changing the scale

- For instance starting with a Uniform on the log standard deviation we end up with a high density on low values for the precision





Remember...

- INLA parametrises the precision and the default is

$$\log\left(1/\sigma^2\right) \sim \text{logGamma}(1, 0.00005)$$

- However alternatives can be built, for instance:
  - Truncated Normal on log precision (`logtnormal`)
  - Uniform prior on the standard deviation: as it is not implemented we need to specify it through the `expression` as follows

```
UN.prior = "expression: log_dens = 0 - log(2) - theta / 2; return(log_dens);
```

In general we need to be careful to check the level of information (weakly, strong) on the scale we are interested in (e.g. variance) and see what this corresponds on the standard deviation/precision (on which prior is usually specified).

See Gómez-Rubio (2020) for more information on how to specify priors in INLA

Model selection

Model comparison: Methods based on the deviance

- When the interest lays mainly on the prior distribution or on the functional form of some parameters the deviance of the model can be used to evaluate the goodness of fit.

Given the data $\mathbf{y}$ with distribution $p(\mathbf{y} \mid \theta)$, the deviance of the model is defined as:

$$D(\theta) = -2\log p(y \mid \theta)$$

where θ identifies the parameter of the likelihood

Model comparison: Methods based on the deviance

- When the interest lays mainly on the prior distribution or on the functional form of some parameters the deviance of the model can be used to evaluate the goodness of fit.

Given the data $\mathbf{y}$ with distribution $p(\mathbf{y} \mid \theta)$, the deviance of the model is defined as:

$$D(\theta) = -2\log p(y \mid \theta)$$

where θ identifies the parameter of the likelihood

- Ex. $y_i \sim \text{Bernoulli}(\theta) \rightsquigarrow p(\mathbf{y} \mid \theta) = \prod_{i=1}^{n} \binom{n_i}{y_i} \theta^{y_i}(1-\theta)^{n_i - y_i}$

$$D(\theta) = -2\left[\sum_i y_i \log\theta_i + (n_i - y_i)\log(1-\theta_i) + \log\binom{n_i}{y_i}\right]$$

Mean deviance

- The deviance of the model measures the variability linked to the likelihood, ie the probabilistic structure used for the observation (conditional on the parameters)
- This quantity is a random variable in the Bayesian framework, so it is possible to synthesise it through several indexes (mean, median, etc.)
- Many authors suggested using posterior mean deviance $(\overline{D}) = E_{\theta|y}[D(\theta)]$ as a measure of fit

DRAWBACK: more complex models will fit the data better and so will have smaller $\overline{D}$

- Need to have some measure of *model complexity* to trade off against $\overline{D}$

Deviance Information Criterion - DIC

- Natural way to compare models is to use criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model
- Deviance Information Criterion, DIC = goodness of fit + complexity of the model

Deviance Information Criterion - DIC

- Natural way to compare models is to use criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model
- Deviance Information Criterion, DIC = goodness of fit + complexity of the model
  – The fit is measure through the deviance

$$D(\theta) = -2\log p(y \mid \theta)$$

Deviance Information Criterion - DIC

- Natural way to compare models is to use criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model
- Deviance Information Criterion, DIC = goodness of fit + complexity of the model
  – The fit is measure through the deviance

$$D(\theta) = -2\log p(y \mid \theta)$$

- Complexity measured by estimate of the "effective number of parameters":

$$p_D = \mathsf{E}_{\theta|y}\left[D(\theta)\right] + D(\mathsf{E}_{\theta|y}\left[\theta\right]) = \overline{D} - D(\bar{\theta})$$

Deviance Information Criterion - DIC

- Natural way to compare models is to use criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model
- Deviance Information Criterion, DIC = goodness of fit + complexity of the model
  – The fit is measure through the deviance

$$D(\theta) = -2\log p(y \mid \theta)$$

- Complexity measured by estimate of the "effective number of parameters":

$$p_D = \mathsf{E}_{\theta|y}\left[D(\theta)\right] + D(\mathsf{E}_{\theta|y}\left[\theta\right]) = \overline{D} - D(\bar{\theta})$$

- The DIC is then defined analogously to AIC as

$$\mathrm{DIC} = D(\bar{\theta}) + 2p_D = \overline{D} + p_D$$

- Models with smaller DIC are better supported by the data

- DIC can be monitored in INLA including `control.compute=list(dic=TRUE)` into the `inla` function.

Scottish lip cancer example

- Counts of cases of lip cancer y_i in 56 districts in Scotland: $y_i \sim \text{Poisson}(\rho_i E_i)$
- Range of models:
  1. Pooled: $\log \rho_i = b_0 + \beta_1 x_i$
  2. Random Effects 1: $\log \rho_i = b_0 + \beta_1 x_i + \theta_i$; Flat prior on $\log \sigma_v$
  3. Random Effects 2: $\log \rho_i = b_0 + \beta_1 x_i + \theta_i$; Gamma prior on $\log \tau_v$

Table: DIC elements under different models

	D	D(theta)	pD	DIC
Pooled	589	588	1	590.4
Random effects 1	270	228	270	310.0
Random effects 2	270	230	270	310.0

Scottish lip cancer example

- Counts of cases of lip cancer y_i in 56 districts in Scotland: $y_i \sim \text{Poisson}(\rho_i E_i)$
- Range of models:
  1. Pooled: $\log\rho_i = b_0 + \beta_1 x_i$
  2. Random Effects 1: $\log\rho_i = b_0 + \beta_1 x_i + \theta_i$; Flat prior on $\log\sigma_v$
  3. Random Effects 2: $\log\rho_i = b_0 + \beta_1 x_i + \theta_i$; Gamma prior on $\log\tau_v$

Table: DIC elements under different models

	D	D(theta)	pD	DIC
Pooled	589	588	1	590.4
Random effects 1	270	228	270	310.0
Random effects 2	270	230	270	310.0

DIC has been criticised over the years, specifically:

1. $p(D)$ is not invariant to reparameterization. For example, we would obtain a (slightly) different value if we parameterized in terms of σ or $log\sigma$

2. It is not based on a proper predictive criterion

4. Issues when there are missing data

See (Spiegelhalter, Best, Carlin, and Van der Linde, 2014) for a complete description of the criticisms.

Watanabe AIC - WAIC

- Considers the posterior predictive mean and variance (on the log scale)
- Linked to cross-validation
- Similarly to DIC:
  – WAIC has a model-fit and model-complexity components
  – Smaller WAIC indicates the preferred model

Watanabe AIC - WAIC

- Considers the posterior predictive mean and variance (on the log scale)
- Linked to cross-validation
- Similarly to DIC:
  – WAIC has a model-fit and model-complexity components
  – Smaller WAIC indicates the preferred model

- Let m_i and v_i be the posterior predictive mean and variance for the $ i^{th} $ unit
- The effective model size is

$$p_W = \sum_{i=1}^{n} v_i$$

- The criteria is

$$WAIC = -2 \sum_{i=1}^{n} m_i + 2p_W$$

- The WAIC is readily available in `INLA` using `control.compute=list(waic=TRUE)`

Summary

- Hierarchical models allow **borrowing of strength** across units

  → posterior distribution of the unit-parameter borrows strength from the likelihood contributions for all the units, via their joint influence on the posterior estimates of the unknown hyper-parameters

  → improved efficiency
- Judgements of exchangeability need careful assessment → units suspected a priori to be systematically different might be modelled by including relevant covariates so that residual variability more plausibly reflects exchangeability
- Subgroups of prior interest should be considered separately

Summary

- Hierarchical models allow **borrowing of strength** across units

  → posterior distribution of the unit-parameter borrows strength from the likelihood contributions for all the units, via their joint influence on the posterior estimates of the unknown hyper-parameters

  → improved efficiency
- Judgements of exchangeability need careful assessment → units suspected a priori to be systematically different might be modelled by including relevant covariates so that residual variability more plausibly reflects exchangeability
- Subgroups of prior interest should be considered separately

Careful on the prior specification

- non informative on one scale might be informative on another
- always run some sensitivity analyses changing the prior and investigating how this affect the estimates of parameters of interest
- DIC is a useful tool for model selection, easy to calculate in INLA → bear in mind that they can only be used to compare models - similarly to the AIC they do not have an absolute meaning.

References

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". In: *Bayesian analysis* 1.3, pp. 515-534.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, et al. (2014). "The deviance information criterion: 12 years on". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.3, pp. 485-493.