Session 4.1: Hierarchical Models, Priors and Model Checking

Imperial College London

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 Blangiardo and Cameletti (2015).

Outline

1. What are hierarchical models
2. Different modelling assumptions
3. Parameter interpretation
4. Hierarchical regression
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 (SMR or SIR)

$$\begin{align*} \text{SMR}_i &=\frac{y_i}{E_i}\\ \hbox{Var}(\hbox{SMR}_i) &= \frac{y_i}{E_i^2} \end{align*}$$

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 (SMR or SIR)

$$\begin{align*} \text{SMR}_i &=\frac{y_i}{E_i}\\ \hbox{Var}(\hbox{SMR}_i) &= \frac{y_i}{E_i^2} \end{align*}$$

- very imprecise: areas with small $\class{red}{E_i}$ 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:

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

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

Expected numbers of cases - definition

- Expected number of cases if the population had the same stratum-specific mortality/incidence rates as in a reference area
- Adjustments (strata): age, gender ...

Indirect standardisation: $E_i = \sum_k n_{ik} r_k$ with

- r_k: disease rate for stratum k in the reference population
- n_{ik}: population at risk in area i, stratum k

If internal comparison: $\color{red}{\sum_{i=1}^N O_i = \sum_{i=1}^N E_i}$

Direct standardisation: apply the disease rate in the population of interest (e.g. UK) to a standard population e.g. European standard population

External comparison: if the reference population is not the population of the study of interest. For example, to calculate the expected numbers in London, risks in England could be used.

Expected numbers of cases - calculation

Strata	Reference area=EW			Ward A		
Age group	Population n_k	Observed O_k	Age-specific rate per 100,000 males $r_k = \frac{O_k}{n_k}$	Population n_{ik}	Observed O_{ik}	Expected $E_{ik} = \frac{n_{ik} * r_k}{100000}$
0–4	41,400,692	15	0.04	11,438	0	0.00
5–9	41,143,722	6	0.01	9,697	0	0.00
10–14	41,469,696	9	0.02	9,026	0	0.00
15–19	43,087,823	39	0.09	8,650	0	0.01
20–24	45,441,353	79	0.17	12,409	0	0.02
25–29	46,873,725	172	0.37	16,963	0	0.06
30–34	46,927,658	518	1.10	17,303	0	0.19
35–39	46,936,367	1,465	3.12	13,847	0	0.43
40–44	45,304,711	4,136	9.13	11,843	1	1.08
45–49	41,657,557	9,835	23.61	9,457	5	2.23
50–54	38,451,416	20,929	54.43	8,561	3	4.66
55–59	35,842,426	40,427	112.79	7,613	8	8.59
60–64	32,480,032	68,230	210.07	6,968	5	14.64
65–69	28,231,499	95,794	339.32	6,290	15	21.34
70–74	23,315,240	110,371	473.39	5,098	27	24.13
75–79	17,297,264	102,038	589.91	4,049	22	23.89
80–84	10,498,214	68,273	650.33	2,616	20	17.01
85+	6,289,452	38,748	616.08	1,312	12	8.08
TOTAL	632,648,846	561,084		163,140	118	126.38

$$\hbox{SIR}_A=\frac{118}{126.38}=0.93$$

- Fewer incident cases of lung cancer for males in ward A than expected in EW after adjusting for differences in age.
- In R we can perform indirect standardization using the package `SpatialEpi` (we will see it in the Practical in week 6).

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:

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

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:

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

- Need to specify hyperprior distributions for:
- σ^2_v (between-area variance), e.g. $1/\sigma^2_v \sim \text{Gamma}(1,0.001)$
- 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:

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

- Need to specify hyperprior distributions for:
- σ^2_v (between-area variance), e.g. $1/\sigma^2_v \sim \text{Gamma}(1,0.001)$
- 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^2_v

→ *global smoothing* of the area RR

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

Such models are called *Hierarchical* or *Random effects* or *Multilevel* models

Interpretation

Parameter interpretation and useful quantities

- ρ_i is the log-relative risk for the area i compared to the average area with the same structure in the expected values.
- 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)
- $\text{exp}(v_{i})$ relative risk in area i compared to the risk for the whole study region
- The variance of the random effects σ^2_v reflects the amount of extra-Poisson variation in the data

Parameter interpretation and useful quantities

- ρ_i is the log-relative risk for the area i compared to the average area with the same structure in the expected values.
- 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)
- $\text{exp}(v_{i})$ relative risk in area i compared to the risk for the whole study region
- The variance of the random effects σ^2_v reflects 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 5th and 95th percentiles of the area relative risk distribution
- let $q_{5\%} = \rho_{5\%}$ denote the log relative risk of outcome for the area ranked at the 5th percentile
- let $q_{95\%} = \rho_{95\%}$ denote the log relative risk of outcome for the area ranked at the 95th percentile

Quantile ratio

$$\text{QR}_{90} = \text{exp}(q_{95\%}-q_{5\%})$$ 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 model selection indexes, as well as to draw samples from the joint posterior

Results for lip cancer in Scotland example

- \(\text{exp}(b_0 + v_i)\) is the relative risk of lip cancer in area \(i\) relative to the average area with the same age/sex structure (see map)
- \(\sigma_v\) 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 standard deviaiton using

```
> sigma.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),sigma.v)
```

```
[1] 0.5018776 0.6746031 0.7255274 0.7731827 0.8332256 1.1426164
```

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
   1.696743    1.322882    1.250297
```

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 \(\bm{\eta}\) in `joint.post$latent` (row 1 to N)
3. joint posterior distribution for the random effects \(\bm{v}\) 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] 10.66425
```

- 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: We consider a dataset which contains measurements on whether Swiss Willow Tits were found on three separate visits in several locations across Switzerland. We want to understand the relationship between the presence of the birds, elevation and forest cover.

Regression in INLA

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

Example: We consider a dataset which contains measurements on whether Swiss Willow Tits were found on three separate visits in several locations across Switzerland. We want to understand the relationship between the presence of the birds, elevation and forest cover.

We specify a random effect logistic model

$$\begin{eqnarray*} y_i &\sim& \text{Bernoulli}(\pi_i , 1)\\ \text{logit}(\pi_i) &=& b_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + v_i\\ v_i &\sim& \text{Normal}(0, \sigma^2_v) \end{eqnarray*}$$

where x_{1i} , x_{2i} are the elevation and forest cover for the i $-th$ visit. Note that b_0 , β_1 , β_2, $1/\sigma^2_v$ are given independent "noninformative" priors.

R code to prepare the dataset

```
> swt = read.csv('swt.csv', stringsAsFactors = TRUE)
```

```
> head(swt)
```

```
  rep.1 rep.2 rep.3 c.2 c.3 elev forest dur.1 dur.
1     0     0     0   0   0  420      3   240    5
2     0     0     0   0   0  450     21   160    3
3     0     0     0   0   0 1050     32   120    4
4     0     0     0   0   0 1110     35   180    4
5     0     0     0   0   0  510      2   210    5
6     0     0     0   0   0  630     60   150    5
```

```
> # Add ID
> swt$ID <- seq(1,length(swt[,1]))
```

```
> # Move from wide to long format
> library(tidyr)
> swt_long <- gather(swt, outings, presence, rep.
> head(swt_long)
```

```
  c.2 c.3 elev forest dur.1 dur.2 dur.3 length alt
1   0   0  420      3   240    58    73    6.2 Lov
2   0   0  450     21   160    39    62    5.1 Lov
3   0   0 1050     32   120    47    74    4.3 Mec
4   0   0 1110     35   180    44    71    5.4 Mec
5   0   0  510      2   210    56    73    3.6 Lov
6   0   0  630     60   150    56    73    6.1 Lov
```

R-INLA code

```
> formula<- presence ~ elev + forest + f(ID,model
> model.regression <- inla(formula,data=swt_long,
> model.regression$summary.fixed
```

```
                    mean           sd  0.025quant
(Intercept) -5.787771531 0.4732679434 -6.71580341
elev         0.002336135 0.0002224772  0.00189988
forest       0.042872372 0.0046380369  0.03377765
```

```
> plot(model.regression$summary.random$ID$mean)
```

Choice of prior

How to specify priors?

- *Relatively* easy to specify priors on regression parameters
  – Typical choice is a Normal distribution
  – Tuning the variance it can be more or less informative
  – The scale of the variable it represents need to be considered

How to specify priors?

- *Relatively* easy to specify priors on regression parameters
  – Typical choice is a Normal distribution
  – Tuning the variance it can be more or less informative
  – The scale of the variable it represents need to be considered

- Variances are more complex (and a bit more controversial)
- In small area studies we usually work with Poisson/Binomial distribution on data - no variance parameter; the main interest is on random effect variance.

How to specify priors?

- *Relatively* easy to specify priors on regression parameters
  – Typical choice is a Normal distribution
  – Tuning the variance it can be more or less informative
  – The scale of the variable it represents need to be considered

- Variances are more complex (and a bit more controversial)
- 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 but inference could be sensitive to choice of ϵ. Typically to ensure vague priors small ϵ are specified (e.g. 0.1, 0.01). However, this prior has also been criticised (e.g. (Gelman, 2006)) as it has a spike for values around 0.

How to specify priors?

- *Relatively* easy to specify priors on regression parameters
  – Typical choice is a Normal distribution
  – Tuning the variance it can be more or less informative
  – The scale of the variable it represents need to be considered

- Variances are more complex (and a bit more controversial)
- 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 but inference could be sensitive to choice of ϵ. Typically to ensure vague priors small ϵ are specified (e.g. 0.1, 0.01). However, this prior has also been criticised (e.g. (Gelman, 2006)) as it has a spike for values around 0.
- Careful as "non informative" prior distributions are sensitive to changes of scale.

Changing the scale

- For instance starting with a Uniform on the 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

Which model?

All models are wrong, some models are useful.

G. Box

Which model?

All models are wrong, some models are useful.

G. Box

So the question is: how is my model doing?

1. in terms of model assumption
2. compared to other models

Which model?

All models are wrong, some models are useful.

G. Box

So the question is: how is my model doing?

1. in terms of model assumption
2. compared to other models

We can answer the first question using the posterior predictive distribution

We can answer the first question using methods based on the trade-off between a measure of model fit and of model complexity

Posterior predictive distribution

Main idea: If the combined model assumptions are reasonable, then our posterior model should be able to simulate data that's similar to the original one

- Let's assume we want to find the relationship between air pollution and asthma attacks and we collect data of the outcome over 500 days in a London Local Authority

Posterior predictive distribution

Main idea: If the combined model assumptions are reasonable, then our posterior model should be able to simulate data that's similar to the original one

- Let's assume we want to find the relationship between air pollution and asthma attacks and we collect data of the outcome over 500 days in a London Local Authority

We propose the following model (y=number of asthma attacks, x=level of \(PM_{10}\) in the previous 3 days \(i=1,\ldots,500\)):

\begin{align} y_i &\sim \text{Poisson}(E \rho_i)\\ \text{log}(\rho_i) &= b_0 + \beta x_i + v_i\\ b_0, \beta &\sim N(0,0.001)\\ v_i &\sim N(0,\sigma^2_{v})\\ \text{log}(1/\sigma^2_{v}) &\sim \text{logGamma}(1,0.00005) \end{align}

The assumptions are

1. that the data are distributed as Poisson
2. that there is a linear relationship between air pollution and the log risk of asthma attacks
3. that the days are similar (we include a random effect)

Posterior predictive distribution

- We run the model and predict observations \(y^*_1,\ldots,y^*_{500}\) based on the posterior distribution of the parameters (note that we need to include `control.predictor` in the `inla` function to access these):

```
> asthma_formula1 <- y ~  x + as.factor(dow) + f(ID, model="iid")
> asthma_model1 <- inla(asthma_formula1, data=data,family="poisson",E = E, control.predictor=list(compute
```

To get the fitted values we run:

```
                          mean        sd 0.025quar
fitted.Predictor.001  7.721553 0.5540704   6.63286
fitted.Predictor.002  7.437117 0.5248997   6.43883
fitted.Predictor.003 12.164226 0.8383416  10.51568
fitted.Predictor.004  4.455800 0.3615330   3.79550
fitted.Predictor.005 10.801745 0.7428501   9.40185
```

Posterior predictive distribution

Now let's assume we run a different model

$$\begin{align} y_i &\sim \text{Normal}(E\theta_i, \tau)\\ \theta_i &= b_0 + \beta x_i\\ b_0, \beta &\sim N(0,0.001)\\ \text{log}(1/\sigma^2_{v}) &\sim \text{logGamma}(1,0.00005) \end{align}$$

The assumptions are

1. that the data are distributed as Gaussian
2. that there is a linear relationship between air pollution and the risk of asthma attacks

Posterior predictive distribution

- We run the model and predict observations \(y^*_1,\ldots,y^*_{500}\) based on the posterior distribution of the parameters (note that we need to include `control.predictor` and `control.compute` in the `inla` function to access these):

```
> asthma_formula2 <- y ~ x
> asthma_model2 <- inla(asthma_formula2, data=data,family="gaussian", E=E,
+                       control.predictor=list(link=1,compute=TRUE),
+                       control.compute=list(config=TRUE,dic=TRUE, waic=TRUE, return.marginals.predictor=
```

To get the fitted values we run:

```
                          mean        sd 0.025quar
fitted.Predictor.001  6.560510 0.2189444   6.13103
fitted.Predictor.002  9.121423 0.1472643   8.83255
fitted.Predictor.003 10.513605 0.2060418  10.10944
fitted.Predictor.004  4.596248 0.3530703   3.90367
fitted.Predictor.005  9.383114 0.1539703   9.08109
```

Comparison

- Both models seem reasonable (the predicted values are in line with the observed ones), but there is more of a shift on the right for the Gaussian model (as expected given its symmetric property)

- Which one is better?

Model comparison: Bayesian p-value

- We can use the posterior predictive distribution to compare to the observed one through a *p-value*
- Let's go back to `asthma_model1` and `asthma_model2` as output of running the `inla` function and use `inla.pmarginal`

```
> # Model 1
> Bayesian_p1 <- c()
> model1_sample <- matrix(NA,500,1000)
> model1_pois <- matrix(NA,500,1000)
>
> for(i in 1:500){
+ model1_sample[i,] <- inla.rmarginal(1000,asthma
+ for(j in 1:1000){
+ model1_pois[i,j] <- rpois(1, model1_sample[i,j
+ }
+
+ Bayesian_p1[i] <- sum(model1_pois[i,]<y[i])/100
+ }
```

- Ideally we would expect a uniform distribution of the p-values which would tell us there is no pattern of over(under) estimation in the prediction
- Here model 1 seems a bit better than model 2

Model comparison: fit vs complexity

- 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 $\bm{y}$ with distribution $p(\bm{y}\mid \theta)$, the deviance of the model is defined as: $$\begin{eqnarray*} D(\theta) = -2 \hbox{log} p(\bm{y} \mid \theta) \end{eqnarray*}$$ where θ identifies the parameter of the likelihood

Model comparison: fit vs complexity

- 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 $\bm{y}$ with distribution $p(\bm{y}\mid \theta)$, the deviance of the model is defined as: $\begin{eqnarray*} D(\theta) = -2 \hbox{log} p(\bm{y} \mid \theta) \end{eqnarray*}$ where θ identifies the parameter of the likelihood

- Ex. $y_i \sim \hbox{Bernoulli}(\theta) \rightsquigarrow p(\mathbf{y}\mid \theta) = \prod_{i=1}^n \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \theta^{y_i} (1-\theta)^{n_i-y_i}$

$\displaystyle D(\theta) = -2\left[\sum_i y_i \log \theta_i + (n_i-y_i)\log(1-\theta_i)+ \log\left(\begin{array}{c}n_i\\y_i \end{array}\right)\right]$

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 $(\overline{D}) = E_{\theta\mid y} [D(\theta)]$ 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

$$D(\theta) = -2 \hbox{log} p(\bm{y} \mid \theta)$$

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

$$D(\theta) = -2 \hbox{log} p(\bm{y} \mid \theta)$$

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

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

$$D(\theta) = -2 \hbox{log} p(\bm{y} \mid \theta)$$

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

- The DIC is then defined analogously to AIC as

$$\hbox{DIC} = D(\textsf{E}_{\theta\mid y}\left[\theta \right]) + 2p_D$$

- 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.

Back to our example...

We run the model adding the `dic` (here for model 1, it is the same for model 2):

```
> asthma_model1 <- inla(asthma_formula1, data=data,family="poisson",control.predictor=list(link=1,compute
+                       control.compute=list(dic=TRUE))
```

And now check the value of the DIC

```
> # Poisson data distribution
> asthma_model1$dic$dic
```

```
[1] 2425.228
```

```
> # Normal data distribution
> asthma_model2$dic$dic
```

```
[1] 2581.008
```

The first model is without any doubt preferred as the DIC is smaller.

DIC: some drawbacks

The 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
3. Issues when there are missing data

See for a complete description of the criticisms.

DIC: some drawbacks

The 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
3. Issues when there are missing data

See for a complete description of the criticisms.

What is the alternative?

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 ith unit
- The effective model size is $$p_W = \sum_{i=1}^nv_i$$
- The criteria is $$WAIC = -2\sum_{i=1}^nm_i + 2p_W$$
- The WAIC is readily available in `INLA` using `control.compute=list(waic=TRUE)`

Back to our example...

We run the model adding the `waic` (here for model 1, it is the same for model 2):

```
> asthma_model1 <- inla(asthma_formula1, data=data,family="poisson",control.predictor=list(link=1,compute
+                       control.compute=list(waic=TRUE))
```

And now check the value of the DIC

```
> # Poisson data distribution
> asthma_model1$waic$waic
```

```
[1] 2422.167
```

```
> # Normal data distribution
> asthma_model2$waic$waic
```

```
[1] 2581.84
```

There is accordance between DIC and WAIC as the first model is still preferred as the WAIC is smaller.

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
- posterior predictive distribution is useful to check if a model is in line with the data under study
- DIC/WAIC are useful tools 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

Blangiardo, M. and M. Cameletti (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models". In: *Bayesian Analysis* 1".3", pp. 515-534.

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