INLA for Spatial Statistics

2. INLA

Daniel Simpson

Department of Mathematical Sciences
University of Bath

Outline

Statistics in space!

Spatial data comes in essentially two different forms

- Point-referenced data
  - GPS tracking
  - Fixed measuring devices
  - “High resolution” satelites
- Region-based data
  - Census data
  - Plot data
  - Region-based counts
  - Historical data

Today, we’re going to talk about regions.

Let's think about data-gathering

A reasonably common way of getting spatial data is

- Break the area of interest up into smaller regions
- Get a team to survey the region
  - Completely
  - Partially

How do we model this statistically?

NW England

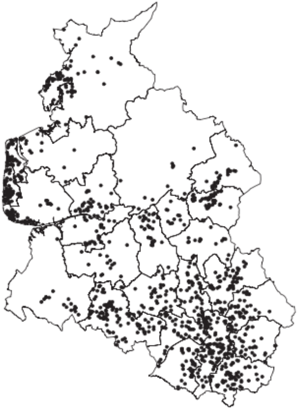


Fig. 1. Leukaemia survival data: districts of Northwest England and locations of the observations.

Outline

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model the number of animals in each region independently

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model the number of animals in each region independently
  - $\eta_i \sim N(\text{intercept} + (\text{covariates})_i, \sigma_i^2)$

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model the number of animals in each region independently
  - $\eta_i \sim N(\text{intercept} + (\text{covariates})_i, \sigma_i^2)$
  - Regional differences accounted through “random effect”

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model the number of animals in each region independently
  - $\eta_i \sim N(\text{intercept} + (\text{covariates})_i, \sigma_i^2)$
  - Regional differences accounted through "random effect"
  - But... what if the distribution is inhomogeneous?

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model the number of animals in each region independently
  - $\eta_i \sim N(\text{intercept} + (\text{covariates})_i, \sigma_i^2)$
  - Regional differences accounted through "random effect"
  - But... what if the distribution is inhomogeneous?
  - If there's an area where the animal is rare, we'll get lots of zero counts

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model some dependence across regions

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model some dependence across regions
    - "Nearby regions" should have similar counts

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model some dependence across regions
    - "Nearby regions" should have similar counts
    - $\eta_i = \text{intercept} + (\text{covariates})_i + u_i$

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model some dependence across regions
  - "Nearby regions" should have similar counts
  - $\eta_i = \text{intercept} + (\text{covariates})_i + u_i$
  - Now the random effect $u_i \sim N(0, Q^{-1})$ is *correlated*

How do we model this?

Imagine we have animal counts in each region. We can model them as Poisson

$$y_i = Po\left(e^{\eta_i}\right).$$

How do we model the *linear predictor* η_i?

- We could model some dependence across regions
  - "Nearby regions" should have similar counts
  - $\eta_i = \text{intercept} + (\text{covariates})_i + u_i$
  - Now the random effect $u_i \sim N(0, Q^{-1})$ is *correlated*
  - How should we do this?

Modelling spatial similarity

The easiest model of spatial similarity is the *Besag* model, which says that

$$x_i - x_j \sim N(0, \sigma^2)$$

if i and j are “neighbours”.

- This really does say nearby things are similar
- It says that the value at neighbouring sites is most probably not more than 3σ apart
- We need to choose neighbours.

Everybody needs good neighbours

How do we choose which points should be neighbours?

- Physical nearest points are often a good place to start
- Physical neighbours are not necessarily the best
- This is *modelling*, so you should consider your process
- Consider, for instance, the problem of Tromsø...

A theory diversion: The Markov property

Models based on neighbourhood have a name in statistics: they are *Markovian models*

- Markovian models are specified entirely through "neighbourhood structures"
- It is easier to than specifying a full covariance
- For a first example, let's consider time

Example: AR(1) process

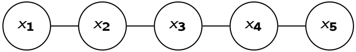
$$x_t \mid x_{t-1} = \phi x_{t-1} + \epsilon_t, \qquad t > 1, \epsilon_t \sim \mathcal{N}(0, \tau^{-1})$$

$$x_1 \sim \mathcal{N}\left(0, \frac{1}{1-\phi^2}\right)$$

- The values at t is proportional to the value at t plus some extra variability
- ϕ is the *lag-one autocorrelation*
- ϵ_t is the innovation noise
- τ is the precision of the innovation
- The distribution for x_1 ensures the process is stationary.

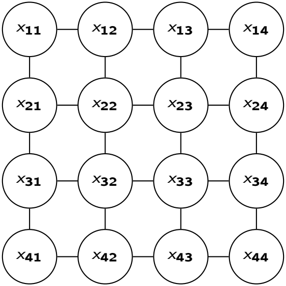
The AR(1) process in pictures

AR(1):



- The circles represent the values of x at individual time points
- There is a line between them if they are *conditionally dependent*

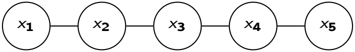
Markov in Space!



- The model above is called a *first order conditional autoregressive model* or a CAR(1) model.
- Every node is conditionally dependent on its *four nearest neighbours*
- This is also called a *First Order Random Walk* or RW(1) model.

(Informal) definition of a GMRF

- A GMRF is a Gaussian distribution where the non-zero elements of the precision (inverse covariance) matrix are defined by the graph structure.
- In the previous example the precision matrix is tridiagonal since each variable is connected only to its predecessor and successor.



Uses for the simple 1-dimensional processes in R-INLA

- The AR(1) process can be used for time simple time effects
- A random walk (RW) process for "smooth effects"

$$x_i - x_{i-1} \sim N(0, \sigma^2)$$

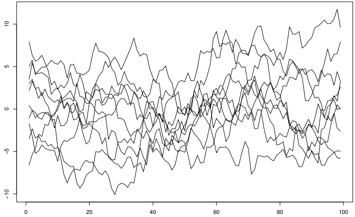
- A second-order random walk (RW2) for even "smoother" effects

$$(x_i - x_{i-1})^2 \sim N(0, \sigma^2)$$

Random walk

Can be used with a

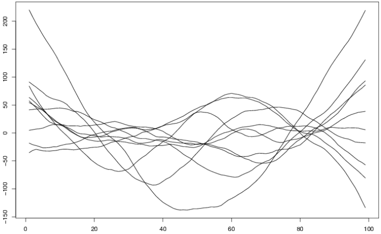
```
formula = Y ~  ... + f(covariate, model="rw1")
```



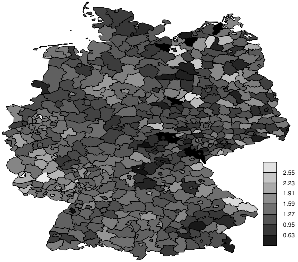
Second-order random walk

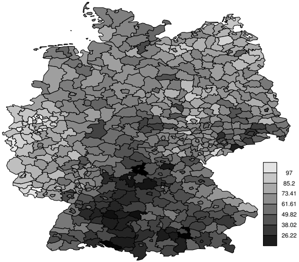
Can be used with a

```
formula = Y ~  ... + f(covariate, model="rw2")
```



Larynx cancer relative risk





Larynx cancer relative risk

Use a simple count model

$$y_i \sim \text{Poisson}(E_i \mathrm{e}^{\nu_i}),$$

where the log-relative risk ν_i is modelled as

$$\nu_i = \text{Covariates} + \text{Spatial} + \text{Noise}.$$

In R-INLA

```
inla(formula = Y~...+f(region, model="besag",
                            graph.file=g),
        family="poisson",...)
```

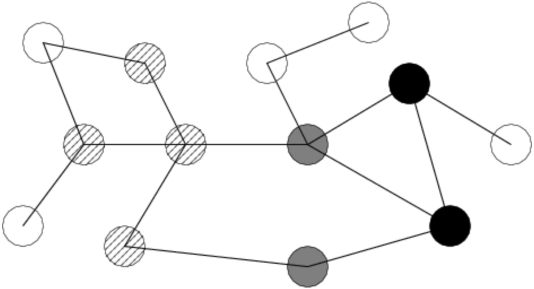
The Markov property on a Graph

Let $\boldsymbol{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

The global Markov property:

$$\boldsymbol{x}_A \perp \boldsymbol{x}_B \mid \boldsymbol{x}_C$$

for all disjoint sets A, B and C where C separates A and B, and A and B are non-empty.



Use a (undirected) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the CI properties,

- $\mathcal{V}$ Vertices: $1, 2, \ldots, n$.
- $\mathcal{E}$ Edges $\{i, j\}$
  - No edge between i and j if $x_i \perp x_j \mid \boldsymbol{x}_{-ij}$.
  - An edge between i and j if $x_i \not\perp x_j \mid \boldsymbol{x}_{-ij}$.

Key point: A graph defines the sparsity structure of $\boldsymbol{Q}$!

Definition of a GMRF

Definition (GMRF)

A random vector $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is called a GMRF wrt the graph $\mathcal{G} = (\mathcal{V} = \{1, \ldots, n\}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$, iff its density has the form

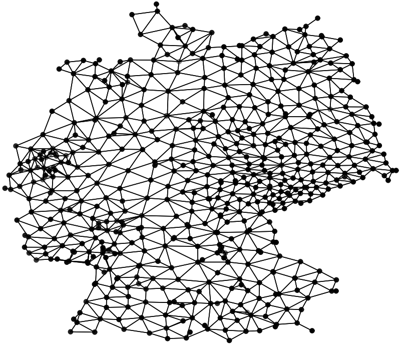
$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$$

and

$$Q_{ij} \neq 0 \quad \Longleftrightarrow \quad \{i, j\} \in \mathcal{E} \quad \text{for all} \quad i \neq j.$$

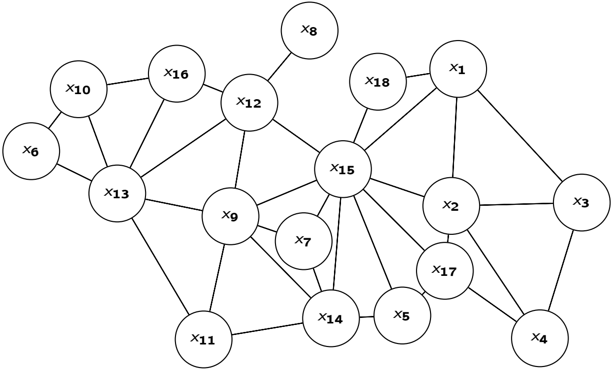
Full graph

Connecting all the neighbouring areas give the following graph



Sub graph

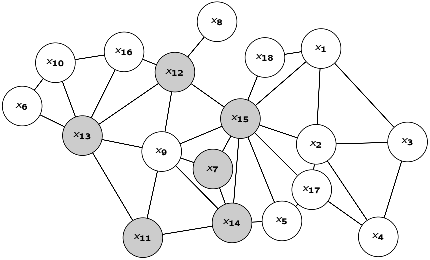
Let us focus on one small part of the graph



Besag model

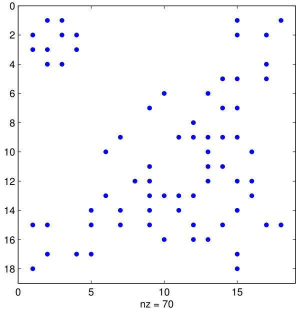
We apply a Besag model where each region conditionally has a Gaussian distribution with mean equal to the average of the neighbours and a precision proportional to the number of neighbours

$$x_9 | \boldsymbol{x}_{-9} \sim \mathcal{N}\left(\frac{1}{6}(x_7 + x_{11} + x_{12} + x_{13} + x_{14} + x_{15}), \frac{1}{6\tau}\right)$$



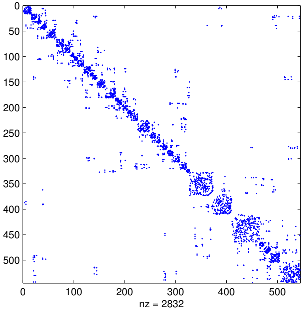
Precision matrix of sub graph

The sub graph leads to a precision matrix with 21.6% non-zero elements.



Precision matrix of full graph

The full graph leads to a precision matrix with 0.1% non-zero elements.



Intrinsic GMRFs

- The Besag model is not proper
- There are linear combinations of the variables that have infinite variance or zero precision.
- This is not allowed in a proper distribution.
- In the Besag model it is caused by the fact that the conditional distributions give no information about the “mean”.

Intrinsic GMRFs

- Distributions of this type (usually) become proper when one introduces observations
- **Identifiability issues**: for a Besag model with an intercept in the model introduce a constraint to stop the Besag from stealing the effect of the intercept.
- R-INLA uses $\sum_i x_i = 0$.

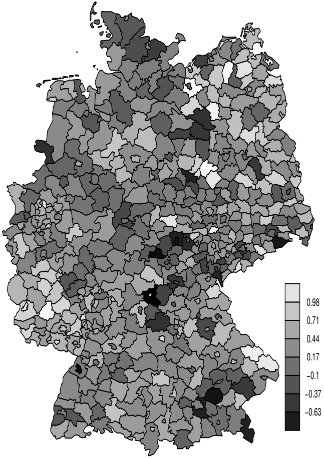
Outline

It turns out the Besag model doesn't fit very well!

- The problem is that it only accounts for similarities between regions
- But it doesn't take into account that every region will have a little bit of individual spice
- The solution is to add an i.i.d. random effect in each region (a random intercept)
- This was the work of Besag, York and Mollié, so we call this the BYM model.

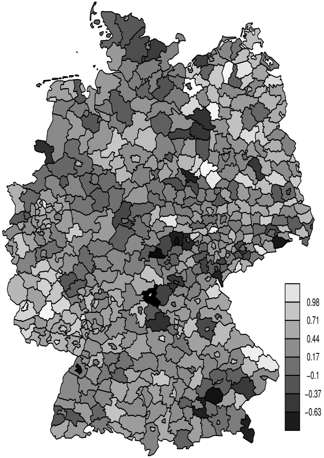
Disease mapping: The BYM-model

- Data $y_i \sim \text{Poisson}(E_i exp(\eta_i))$



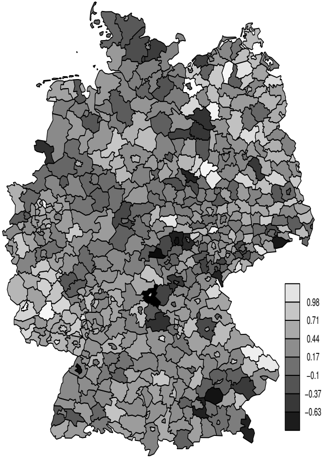
Disease mapping: The BYM-model

- Data $y_i \sim \text{Poisson}(E_i exp(\eta_i))$
- Log-relative risk
  $\eta_i = \mu + u_i + v_i + f(c_i)$



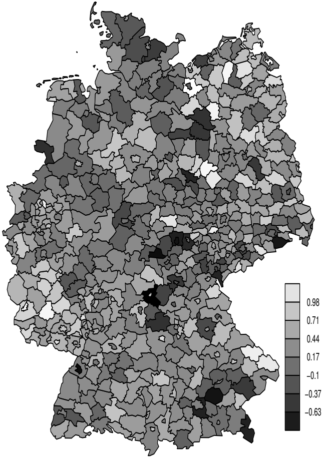
Disease mapping: The BYM-model

- Data $y_i \sim \text{Poisson}(E_i exp(\eta_i))$
- Log-relative risk
  $\eta_i = \mu + u_i + v_i + f(c_i)$
- Structured/spatial component $\boldsymbol{u}$



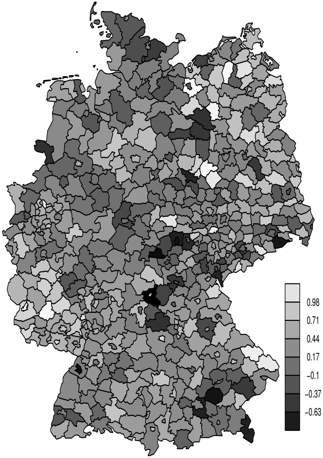
Disease mapping: The BYM-model

- Data $y_i \sim$ Poisson$(E_i exp(\eta_i))$
- Log-relative risk $\eta_i = \mu + u_i + v_i + f(c_i)$
- Structured/spatial component $\boldsymbol{u}$
- Unstructured component $\boldsymbol{v}$



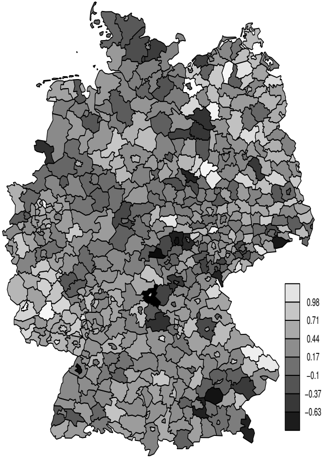
Disease mapping: The BYM-model

- Data $y_i \sim \text{Poisson}(E_i exp(\eta_i))$
- Log-relative risk
  $\eta_i = \mu + u_i + v_i + f(c_i)$
- Structured/spatial component $\boldsymbol{u}$
- Unstructured component $\boldsymbol{v}$
- $f(c)$ is the non-linear effect of a covariate c.



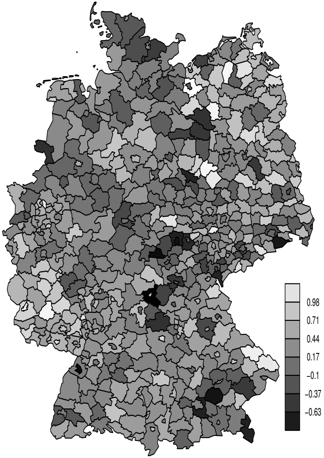
Disease mapping: The BYM-model

- Data $y_i \sim \text{Poisson}(E_i exp(\eta_i))$
- Log-relative risk $\eta_i = \mu + u_i + v_i + f(c_i)$
- Structured/spatial component $\boldsymbol{u}$
- Unstructured component $\boldsymbol{v}$
- $f(c)$ is the non-linear effect of a covariate c.
- Precisions τ_u and τ_v; smoothing parameter τ_f

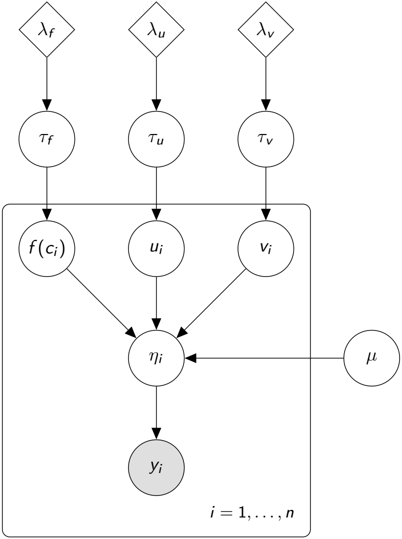


Disease mapping: The BYM-model

- Data $y_i \sim \text{Poisson}(E_i exp(\eta_i))$
- Log-relative risk
  $\eta_i = \mu + u_i + v_i + f(c_i)$
- Structured/spatial component $\boldsymbol{u}$
- Unstructured component $\boldsymbol{v}$
- $f(c)$ is the non-linear effect of a covariate c.
- Precisions τ_u and τ_v; smoothing parameter τ_f
- Common to use independent Gamma-priors



Complicated model components



Does this make sense?

Think of the variance

- The variance not explained by the covariate is modelled with u_i and v_i
- This amount of variance we can have is controlled by the independent precision parameters τ_u and τ_v
- This is ugly!
- It would be much easier to have one parameter controlling the scale of the random effect, and another controlling its makeup
- This is implemented as the `bym2` model in INLA

Disease mapping (II)

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\gamma}\mathbf{v} + \sqrt{\gamma}\mathbf{u}\right)$$

- Marginal precisions τ.

Disease mapping (II)

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}} \left(\sqrt{1-\gamma} v + \sqrt{\gamma} u \right)$$

- Marginal precisions τ.
- γ gives it interpretation: independence ($\gamma = 0$), maximal dependence ($\gamma = 1$)]

Disease mapping (II)

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\gamma}v + \sqrt{\gamma}u\right)$$

- Marginal precisions τ.
- γ gives it interpretation: independence ($\gamma = 0$), maximal dependence ($\gamma = 1$)]
- PC prior on γ depends on the graph!

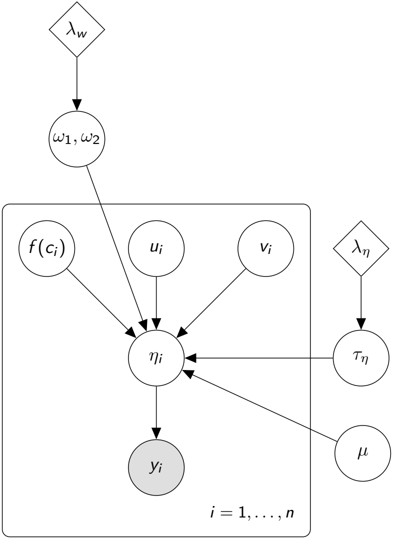
Disease mapping (II)

Rewrite the model as

$$\eta = \frac{1}{\sqrt{\tau}} \left(\sqrt{1-\gamma} v + \sqrt{\gamma} u \right)$$

- Marginal precisions τ.
- γ gives it interpretation: independence ($\gamma = 0$), maximal dependence ($\gamma = 1$)]
- PC prior on γ depends on the graph!
- Parameters control different features. Use the PC priors (later!) for τ and γ separately.

Building a better BYM



This re-parameterisation in terms of "meaningful" parameters makes it easier to set priors and leads to more stable inference.

Outline

Model choice

Chose/compare various model is important but difficult

- Bayes factors (general available)
- Deviance information criterion (DIC) (hierarchical models)
- Conditional predictive ordinances (CPO)

Never forget

Your model doesn't fit!

"All models are wrong, some models are useful" — George Box

Bayesian model comparison

- There is *no gold standard*
- It depends on what you want to do
- Basically two types
  - Ones that look at the posterior probability of the data under the model
  - Ones that look at how model the data fits the data
- The best hope is to have a model that represents data that wasn't used to fit it...

Marginal likelihood

Marginal likelihood is the normalising constant for $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\widetilde{\pi}(\boldsymbol{y}) = \int \left. \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\theta)\pi(\boldsymbol{y}|\boldsymbol{x},\theta)}{\widetilde{\pi}_{\mathsf{G}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^{\star}(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{1}$$

Marginal likelihood

Marginal likelihood is the normalising constant for $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\widetilde{\pi}(\boldsymbol{y}) = \int \left.\frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\theta)\pi(\boldsymbol{y}|\boldsymbol{x},\theta)}{\widetilde{\pi}_{\mathsf{G}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^{\star}(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{1}$$

I many hierarchical GMRF models the prior is intrinsic/improper, so this is difficult to use.

Deviance Information Criteria

Based on the *deviance*

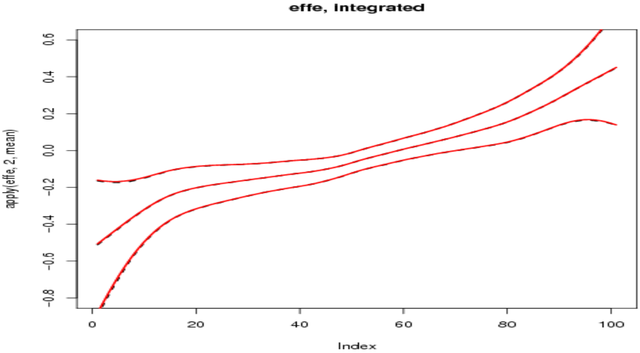
$$D(\boldsymbol{x}; \boldsymbol{\theta}) = -2 \sum_i \log(y_i \mid x_i, \boldsymbol{\theta})$$

and

$$DIC = 2 \times \text{Mean}\left(D(\boldsymbol{x}; \boldsymbol{\theta})\right) - D(\text{Mean}(\boldsymbol{x}); \boldsymbol{\theta}^*)$$

This is quite easy to compute

Example



Will a linear effect be sufficient?

Bayesian Cross-validation

Easy to compute using the INLA-approach

$$\pi(y_i \mid \boldsymbol{y}_{-i}) = \int_{\boldsymbol{\theta}} \left\{ \int_{x_i} \pi(y_i \mid x_i, \boldsymbol{\theta})\, \pi(x_i \mid \boldsymbol{y}_{-i}, \boldsymbol{\theta})\, dx_i \right\} \pi(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})\, d\boldsymbol{\theta}$$

where

$$\pi(x_i \mid \boldsymbol{y}_{-i}, \boldsymbol{\theta}) \propto \frac{\pi(x_i|\boldsymbol{y}, \boldsymbol{\theta})}{\pi(y_i|x_i, \boldsymbol{\theta})}$$

- If it is very small, this point may be an "outlier" under the model
- We can use this to define a score (bigger is better)

$$LCPO = \sum_i \log(\pi(y = y_i|y_{-1}))$$

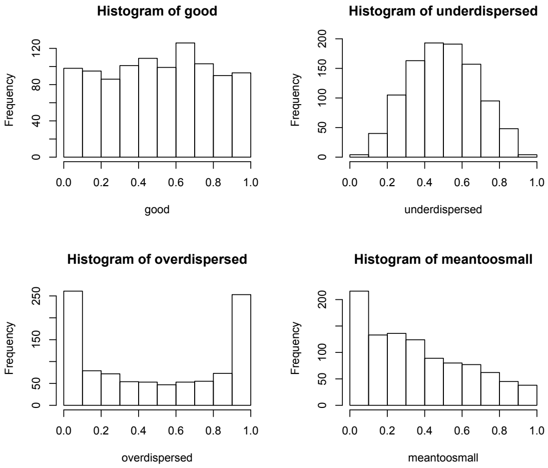
Automatic detection of “surprising” observations

Compute

$$pit_i = \text{Prob}(y_i^{\text{new}} \leq y_i \mid \boldsymbol{y}_{-i})$$

- pit_i shoew how well the ith data point is predicted by the rest of the data
- If the model is true, these PIT values are uniformly distributed
- We can use this to inspect the model fit

Good and Bad PIT plots


Histogram of good
Frequency
good
Histogram of underdispersed
Frequency
underdispersed
Histogram of overdispersed
Frequency
overdispersed
Histogram of meantoosmall
Frequency
meantoosmall