Lab 4: Model selection and validation

Gaussian Markov random fields

David Bolin
Chalmers University of Technology
February 16, 2015

Model selection and validation

Model validation is an important part of a statistical analysis, we may want to

- measure how good a model is for answering a certain question
- compare different models to see which model that fits the best

How to do this

- depends on what we are interested in.
- Really deserves more time than what we spend on it here

Popular methods include AIC, BIC, and other measures of model fit based on asymptotic arguments

More modern (and better) measures of fit, based on cross-validation, include the continuous ranked probability score and the energy score

Model selection IN R-INLA

R-INLA has some built-in model selection tools that are easy to use:

- DIC: The deviance information criterion.
- CPO: conditional predictive ordinate
- Log-score
- PIT: cross-validated probability integral transform

Model selection: DIC

DIC is defined as

$$\text{DIC} = 2\text{E}(D(\mathbf{x}, \boldsymbol{\theta})) - D(\text{E}(\mathbf{x}, \boldsymbol{\theta}))$$

where $D(\mathbf{x}, \boldsymbol{\theta}) = -2\sum_i \log \pi(y_i|x_i, \theta)$ is the deviance.

- $\text{E}(D(\mathbf{x}, \boldsymbol{\theta})) - D(\text{E}(\mathbf{x}, \boldsymbol{\theta}))$ corresponds to the effective number of parameters.
- $\text{E}(D(\mathbf{x}, \boldsymbol{\theta}))$ favors a good fit
- We choose models with small DIC
- It can be seen as a hierarchical modeling generalization of AIC and BIC
- It is based on asymptotic arguments and may underpenalize complex models with many random effects.
- It requires approximate normality, INLA "fixes" this by evaluating posterior mode of $\boldsymbol{\theta}$ instead of the posterior mean

Model selection: CPO

The conditional predictive ordinate (CPO) is a leave-one-out cross-validation score

$$\text{CPO}_i = \pi(y_i^{obs}|y_{-i})$$

where y_{-i} denotes the observations y with the ith component removed.

It expresses the posterior probability of observing the value of y_i when the model is fitted to all data except y_i.

- A high value implies a better fit of the model to y_i.
- A low value suggest that y_i is an outlier and an influential observation.

The CPO is connected with the frequentist studentized residual test for outlier detection.

Model selection: log-score

Based on the CPO-values, we can calculate the logarithmic score

$$\text{logscore} = -\sum_i \log \text{CPO}_i$$

A smaller value of the logarithmic score indicates a better prediction quality of the model

The log-score can be seen as an estimator of the logarithm of the marginal likelihood, and is therefore sometimes called the log pseudo marginal likelihood (PsML).

A ratio of PsMLs is a surrogate for the Bayes factor, sometimes known as the pseudo Bayes factor (PsBF).

Model selection: PIT

The cross-validated probability integral transform (PIT) is also a leave-one-out cross-validation score

$$\text{PIT}_i = \text{P}(y_i < y_i^{obs} | y_{-i})$$

where y_{-i} denotes the observations y with the ith component removed.

For a well-calibrated model, the PIT values should be uniformly distributed. Histograms of the PIT values can therefore be used to assess the calibration of the model.

Example: Binomial regression with random effects

This is a Winbugs/ Openbugs example.

- Two types of seeds were planted and treated with one of two root extracts on one of 21 plates arranged in a 2×2 factorial design.
- The number that germinated was measured.
- The sampling model is $y_i|\eta_i, n_i \sim Bin(n_i, p_i)$
- The probabilities p_i are modelled through a logit link

$$\text{logit}(p_i) = \mu + \beta_1 x_1 + \beta_w x_2 + \beta_3 x_1 x_2 + f(\text{plate}),$$

where x_1 is the seed type and x_2 is the root extract.

- The random effect $f(\text{plate}_i)|\tau \sim N(0, \tau^{-1})$... A random intercept model

Estimate the model using R-INLA

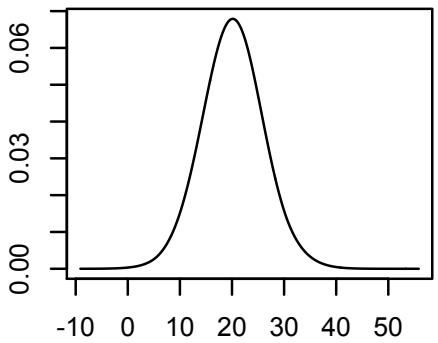
```
require(INLA)

#Load up the data - in the INLA package
data(Seeds)

# Define your formula
formula = r~ x1 + x2 +x1*x2 + f(plate,model="iid")

#Run INLA.
mod.seeds = inla(formula,family="binomial",
                 Ntrials=n, data=Seeds)

#View the results
summary(mod.seeds)
plot(mod.seeds)
```

Results

```
> summary(hyp.seeds)

Call:
"inla(formula = formula, family = \"binomial\", data = Seeds, Ntrials = n)"

Time used:
 Pre-processing    Running inla Post-processing           Total
         1.1072          0.1966          0.0843          1.3881

Fixed effects:
               mean     sd 0.025quant 0.5quant 0.975quant    mode kld
(Intercept) -0.5581 0.1261    -0.8080  -0.5573    -0.3127 -0.5557   0
x1           0.1461 0.2233    -0.2940   0.1467     0.5826  0.1479   0
x2           1.3206 0.1776     0.9745   1.3197     1.6714  1.3179   0
x1:x2       -0.7793 0.3066    -1.3806  -0.7796    -0.1773 -0.7800   0

Random effects:
Name   Model
 plate   IID model

Model hyperparameters:
                     mean     sd       0.025quant 0.5quant 0.975quant mode
Precision for plate 19549.17 19815.75   357.69   13407.51 73021.89      85.25

Expected number of effective parameters(std dev): 4.014(0.0114)
Number of equivalent replicates : 5.231

Marginal Likelihood:  -72.07
```

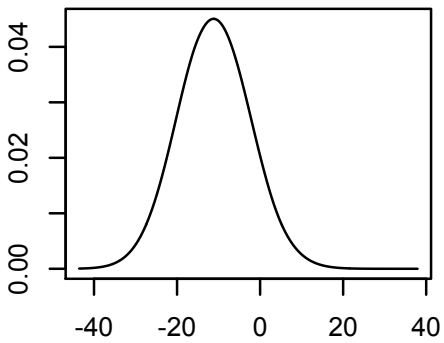
Results

PostDens [(Intercept)]



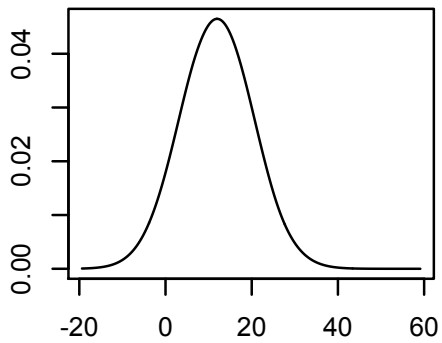
Mean = 20.137 SD = 5.965

PostDens [x1]



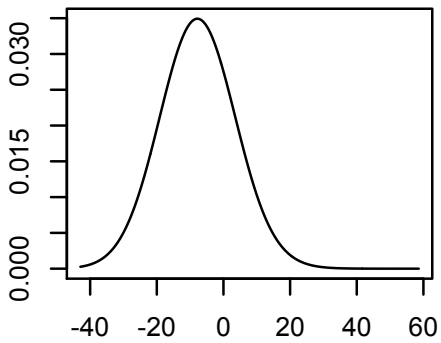
Mean = -11.106 SD = 8.773

PostDens [x2]



Mean = 11.894 SD = 8.5

PostDens [x1:x2]



Mean = -7.634 SD = 11.331

Is it necessary to include the random effect for the plates?

```
formula = r~ x1 + x2 + x1*x2

r1 = inla(formula,family="binomial",
          Ntrials=n, data=Seeds,
          control.compute=list(dic=TRUE,cpo=TRUE))

formula = r~ x1 + x2 + x1*x2 + f(plate,model="iid")

r2 = inla(formula,family="binomial",
          Ntrials=n, data=Seeds,
          control.compute=list(dic=TRUE,cpo=TRUE))

cat(r1$dic$dic, r2$dic$dic)
cat(-sum(log(r1$cpo$cpo)),-sum(log(r2$cpo$cpo)))
```