13 Model assessment

The first two steps in bayesian analysis are broadly:

(1) defining the probability model (conditional model of data & prior).
(2) computation of the model (if via MCMC, convergence check).

In computations, we could easily have more than one possible model already, possibly running in parallel as in some of our previous examples. Because BUGS is so 'open-ended' tool, it is fairly easy to try slightly different models. And there are several things that are possible to modify: prior distribution, conditional distribution of data, hierarchical structures, link functions, parameterizations....

Before final conclusions, some form of model assessment is recommended. Several things could be asked at this stage: how well the model describes data, how well it corresponds to what is generally known about the phenomenon we are modeling, how well it can predict future observations? And how sensitive the conclusions are to the choices we made during modeling? And of course, the BUGS code itself should be debugged: trap messages indicate some faults in the model that may be due to poor identifiability, bad numerical range of values, or logical inconsistencies. All of these are poorly detectable by the messages we get from BUGS. There are too many things that could be a problem, so the error messages cannot be very specific. Such faults just need to be resolved by trial and error. But a running BUGS simulation can still produce nonsense. It is quite possible to get something running which still doesn't make sense. See the BUGS-paper for examples (e.g. there could be superfluous parameters that could not be estimated from any data). Or recall the nonsense model `x ~ dnorm(x,1)`. Greater freedom in modeling requires greater responsibilities from the user. There is no automatic safeguard. In the following we just assume that we have a model or several models that *do make sense*.

We can (1) examine the fit of the current model, and (2) compare alternative models. Of course, in the former task we are already prepared to propose a better model in case the current model is not very good.

13.1 Sensitivity analysis

Are the results very sensitive to different model assumptions? What if I had used a slightly different model?

In Bayesian context, sensitivity analysis typically consists of checking the sensitivity of posterior inferences to different priors. (Although the conditional distribution of data, 'likelihood', can be just as suspect. By 'model', we mean both structures). When the data set is large and informative, the posterior is closely the same under a reasonable set of priors. But when the data provide very little information, the results become increasingly sensitive to the choice of prior. Standard choices of 'uninformative' priors can then lead to unrealistic results. But if the results seem unrealistic to us, it means that we do have some background information that could be used for selecting an informative prior that represents such knowledge. E.g. in the binomial model with little data and uniform prior, the posterior mean will likely be near 0.5 (the prior mean). If this seems unreasonable, then perhaps the uniform prior was not really describing our actual 'state of uncertainty'.

In hierarchical models, model improvements can be sought in several levels from hyper priors down to

conditional distributions of data. For example the choice of uninformative priors for variance parameters can be problematic when data are limited. Then, Gelman suggests uniform prior for σ instead of a Gamma-density for $\tau = 1/\sigma^2$ with small parameters [?].

13.1.1 Example: schools and SAT scores

Sensitivity to prior distributions could be checked. Also, the normal model of observations could be replaced by e.g. t-distribution. These modifications might affect the estimated school effects. Also, the prior exchangeability assumption could be questioned (are there groups of similar schools?). Based on the posterior distribution of model parameters, completely replicated observations could be predicted. These predictive distributions could be compared with the actual data. Are they in agreement? Also, we could produce predictive distribution for the minimum observation $\min(y)$ or $\max(y)$ and compare these with the actual min and max. Likewise, other less trivial functions of data could be selected to check consistency of the model predictions with the actual values.

13.1.2 Example: meta-analysis

(Example 5.13, BSM p. 188). This is a meta-analysis of studies $i = 1, \ldots, 5$ about the effect of vitamin A supplements on childhood mortality and morbidity. In each study, the number of deaths were reported in two groups: those who took the supplement, and a control group without the supplement. The study specific model is:

$$\begin{aligned} D_{\text{vita},i} &\sim \text{Bin}(N_{\text{vita},i}, p_{\text{vita},i}) \\ D_{\text{control},i} &\sim \text{Bin}(N_{\text{control},i}, p_{\text{control},i}), \end{aligned}$$

where the difference between the treatment group and control group could be parameterized as:

$$\begin{aligned} \text{logit}(p_{\text{control},i}) &= \mu_i \\ \text{logit}(p_{\text{vita},i}) &= \mu_i + e_i. \end{aligned}$$

The control group mean μ_i, and the treatment effect e_i in *each study* have the priors

$$\begin{aligned} \mu_i &\sim \text{N}(0, 1.0E+5) \\ e_i &\sim \text{N}(\beta, \sigma^2). \qquad \tau = 1/\sigma^2. \end{aligned}$$

The parameters e_i are also known as a *random effects* describing unexplained 'random' differences between different studies. A meta-analysis attempts to draw information from all studies, by specifying hyper priors for parameters β (N(0,1.0E+4)) and τ (Default prior, maybe not good anyway: Gamma(1.0E-3,1.0E-3)). As a result, we obtain posterior density of β, or some interesting function of this parameter, e.g. OR=exp(β) which is the 'common odds ratio'. (Compare: study specific $\text{OR}_i = \exp(e_i)$). Additionally, it is easy to compute the predictive effect by simulating posterior predictive distribution for: $e.\text{new} \sim \text{N}(\beta, \tau)$. This prediction depends conditionally from hyper parameters, which become estimated from all studies (and hyper prior!). The predicted study is exchangeable, *a priori*, with the other studies. The number of studies was only 5. Therefore, the results may be sensitive to the choice of hyper prior $\pi(\tau)$.

```
Data
D.Vit[] N.vit[] D.Controls[] N.Controls[]
101     12991    130          12209
 39      7076     41           7006
 37      7764     80           7755
152     12541    210          12264
138      3786    167           3411
END
```

13.2 Bayesian residual plots

In regression models, the conditional expected value of observations $Y = Y_1, \ldots, Y_n$ is given by $E(Y \mid X, \theta)$ where X denotes explanatory variables, and θ denotes unknown parameters. This expected value is thus some function $g(X, \theta)$. For a given value of x_i and θ, the value of $g(x_i, \theta)$ is the predicted value for the data point y_i. The *'realized' residual* is $y_i - g(x_i, \theta)$. Note: θ is here unknown. In contrast, the classical or estimated residual is $y_i - g(x_i, \hat{\theta})$. Classical residual plots can be thought of as approximations to the bayesian residual plots, ignoring posterior uncertainty in θ. In BUGS, these can be obtained by inserting the appropriate expression to be monitored.

13.2.1 Ice cream consumption

Recall the linear model of ice cream consumption explained by temperature.

$$\text{Cons}_i \sim \text{N}(\mu_i, \sigma^2)$$

$$\mu_i = \alpha_0 + \alpha_1 \text{Temp}_i$$

Below are model fit plots of the expected consumption μ versus temperature, and μ versus time, together with the actual consumptions y. When plotted against temperature, the fit seems reasonable, perhaps apart from e.g. the highest consumption that seems higher than expected. When plotted against time, the highest consumption is most out of pattern and it is also the last time point in the series. There also seems to be a temporal pattern that is not so well fitted. Perhaps a temporally structured random effect might be added? On the other hand, adding seasonal parameter might not be good idea because season and temperature are probably collinear variables.

Plots of residuals $y - \mu$ are also included, both versus temperature, and versus time.

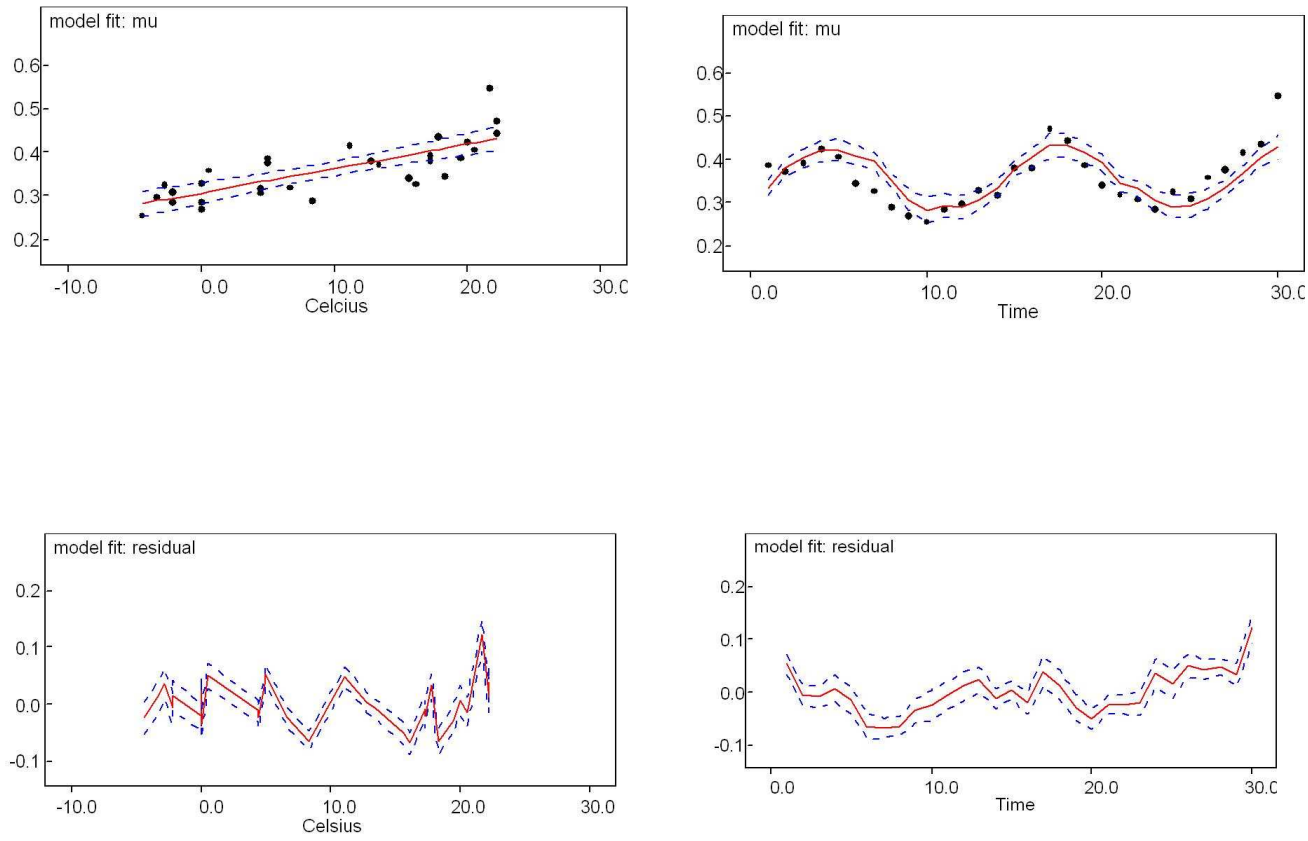

Figure 1: Posterior distribution of μ (and y as '•') versus temperature (top-left) and time (top-right). Residuals $y - \mu$ versus temperature (down-left) and time (down-right). Residual plots should be scattered around zero. Box plots for these could also be used.

13.3 Predictive model fit diagnostic

The classical p-value is defined as

$$P(T(X^{\text{pred}}) > T(X^{\text{obs}}) \mid \theta),$$

where θ is a fixed parameter of the model $\pi(X \mid \theta)$. Here, $T(X^{\text{obs}}$ is an observed value, and the distribution of $T(X^{\text{pred}})$ is determined for fixed θ, i.e. it does not depend on observations X. In classical analysis, the value of θ is typically determined by a 'null hypothesis'. In bayesian context, this can be generalized to

$$P(T(X^{\text{pred}}, \theta) > T(X^{\text{obs}}, \theta) \mid X^{\text{obs}}).$$

With fixed θ, the classical p-value is obtained as a special case. Graphically, we can compare $T(X^{\text{pred}}, \theta)$ with $T(X^{\text{obs}}, \theta)$ by plotting a scatter plot of them, taken from the MCMC sample of them. Alternatively, we could draw the histogram of their difference. The scatter plot should be symmetric about the

45° line, and the histogram should include 0. Ideally, test quantities will be chosen to reflect aspects of the model that are relevant to the scientific purposes to which the inference will be applied.

Note: bayesian predictive checks are not used to 'accept' or 'reject' a model (remember: all models are wrong anyway!) but rather to understand the limits of its applicability in realistic applications.

13.3.1 Example: normal model

Consider the model $X_i \sim \mathrm{N}(\mu, \sigma^2)$ with unknown μ and assumed (fixed) σ^2. After the data $X_1, \ldots, X_n$ has been observed, the posterior of μ is $\mathrm{N}(\bar{X}, \sigma^2/n)$, and the predictive distribution of a new X^* is $\mathrm{N}(\bar{X}, \sigma^2 + \sigma^2/n)$. Based on this, we can simulate a replicate data of n observations $X_1^*, \ldots, X_n^*$ to be compared with the original data. Is the simulated data reasonably similar to the original? We need to decide how the similarity is to be judged. For example, we could study the predictive distribution of the smallest data point. This is easily obtained from the simulations by generating the whole data set many times (iterations) and each time (iteration) recording the smallest of the n generated points. Assume the original data has 10 observations:

```
-0.7417224, -2.1873614,  1.1508363,  0.1306749, -1.1931158,
 0.2093445, -0.1040642,  1.230186,   0.910799,  0.1830353
```

The smallest of these was -2.187361, and:

$$\pi(\mu \mid X, \sigma) = \mathrm{N}(-0.04113878, \sigma^2/10)$$

$$\pi(X^* \mid X, \sigma) = \mathrm{N}(-0.04113878, \sigma^2 11/10).$$

In this case, these are easy to simulate with R by simply drawing from normal densities. The posterior predictive distribution of the smallest value among 10, assuming different σ values, is simulated as

```
sigma<-0.5
for(i in 1:10000){xmin[i]<-min(rnorm(10,mean(x),sigma*sqrt(11/10)))}
```

The models with $\sigma = 0.5$ and $\sigma = 10$ show the worst fit to the the smallest data point, whereas the model with $\sigma = 1$ performs better. The bayesian p-value $P(X^*_{\text{smallest}} > -2.2 \mid X)$ compares the predicted variable to its observed value. A p-value that is close to 0 or 1 indicates lack of fit. But there are many discrepancy functions that we could study, and they could also depend on the unknown parameter. For example:

$$T(X, \theta) = \mid X_{\text{smallest}} - \theta \mid$$

$$T(X^*, \theta) = \mid X^*_{\text{smallest}} - \theta \mid$$

and the bayesian p-value is then

$$P(T(X^*, \theta) > T(X, \theta) \mid X)$$

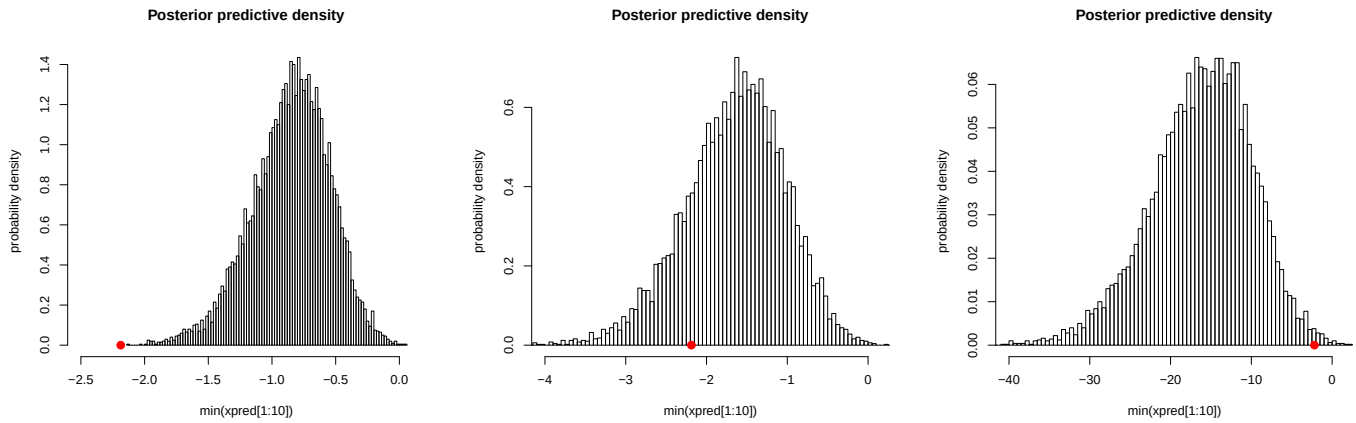

Figure 2: Posterior predictive distribution of X^*_{smallest}, based on $\sigma = 0.5$ (left), $\sigma = 1$ (middle), $\sigma = 10$ (right). Minimum data point (-2.2) shown as red bullet.

13.3.2 Example: independence of bernoulli trials

Assume that a series of 20 bernoulli trials is observed to be

$$x = c(1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0).$$

Our model could be $x_i \sim \text{Bernoulli}(p)$, with prior $p \sim \text{U}(0, 1)$. This model assumes that each x_i is conditionally independent, given p. But the data show considerably systematic pattern that does not look like it could result from independent trials. We can check the model fit with respect to the number of switches from 0 to 1, or 1 to 0, over the series. In the observed data, the number of switches was $T(x) = 3$. We need to compute $P(T(x^{\text{rep}}) > T(x) \mid x)$, from the posterior predictive distribution of $T(x^{\text{rep}})$, that is:

$$\pi(T(x^{\text{rep}}) \mid x) = \int_0^1 \pi(T(x^{\text{rep}}) \mid p)\pi(p \mid x)\mathbf{d}p$$

```
model{
p  ~ dunif(0,1)
for(i in 1:n){
x[i] ~ dbern(p);  xrep[i]~dbern(p)
}
for(i in 2:n){
test[i-1] <- 1-equals(xrep[i],xrep[i-1]) # count the switches
}
T <- sum(test[1:n-1])
P <-step(T-Tobs)
Q <-step(Tobs-T)
}
list(n=20,Tobs=3,x=c(1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0))
```

The statistic T has posterior predictive distribution with mean 8.4 and 95% CI [3,13]. Therefore, the observed value $T = 3$ seems unlikely when compared with this prediction. The simple model is not fitting well the number of switches.

13.3.3 Omnibus discrepancies

Consider the following discrepancy

$$T(x^{\text{obs}}, \theta) = \sum \frac{(X_i - E(X \mid \theta))^2}{V(X \mid \theta)}$$

$$T(x^{\text{pred}}, \theta) = \sum \frac{(X_i^{\text{pred}} - E(X \mid \theta))^2}{V(X \mid \theta)}$$

Assume then that we have observed $X_1, \ldots, X_7$ and our model is $\text{N}(\mu, \sigma^2)$ with unknown μ but with fixed σ^2. Obviously, the choice of model is subjective and there are two subjective elements: the choice of normal model in the first place, and the choice of σ^2. In this example, the observed X_i values were generated from $\text{N}(0, 1)$-model, so the normal distribution is a correct choice. But if we choose wrong variance, $\sigma^2 \neq 1$, the chosen density is either too narrow or too wide. In the data, sample mean = -0.3907726, and sample SD = 1.086697. We apply noninformative prior $\mu \sim \text{N}(0, 0.001)$ and assume specific values for σ^2. The posterior density of the unknown mean will be $\mu \sim \text{N}(\bar{X}, \sigma^2/7)$. The predictive density is thus

$$\pi(X^{\text{pred}} \mid X^{\text{obs}}) = \int_{-\infty}^{\infty} \pi(X^{\text{pred}} \mid \mu, \sigma^2) \underbrace{\pi(\mu \mid X^{\text{obs},\sigma^2})}_{\text{N}(\bar{X}, \sigma^2/7)} \mathbf{d}\mu$$

which can be simulated by sequentially drawing values of μ from $\text{N}(\bar{X}, \sigma^2/7)$, and then values of X^{pred} from $\text{N}(\mu, \sigma^2)$. In this way, we can produce a simulated data of seven values $X_1^{\text{pred}}, \ldots, X_7^{\text{pred}}$ which can be compared with the actual seven values. A good model produces predictions that are similar to the actual data.

```
model{
 # model 1: too small variance
 tau[1]<-1/(s[1]*s[1]);  s[1]<-0.5; mu[1] ~ dnorm(0,0.0001)
 # model 2: variance = observed sample variance
 tau[2]<-1/(s[2]*s[2]); s[2]<-1.086697; mu[2] ~ dnorm(0,0.0001)
 # model 3: too large variance
 tau[3]<-1/(s[3]*s[3]); s[3]<-400; mu[3] ~ dnorm(0,0.0001)

 for(m in 1:3){
 for(i in 1:N){
  x[m,i] ~ dnorm(mu[m],tau[m])
  xpred[m,i] ~ dnorm(mu[m],tau[m])
  T1[m,i] <- pow((x[m,i]-mu[m])/s[m],2)
  T2[m,i] <- pow((xpred[m,i]-mu[m])/s[m],2) }
  TT[m,1]<-sum(T1[m,]);
  TT[m,2]<-sum(T2[m,]) P[m]<-step(TT[m,2]-TT[m,1]) } }
```

```
 list(N=7,x=structure(.Data=c(
-0.7417224, -2.1873614, 1.1508363, 0.1306749, -1.1931158, 0.2093445, -0.1040642,
-0.7417224, -2.1873614, 1.1508363, 0.1306749, -1.1931158, 0.2093445, -0.1040642,
-0.7417224, -2.1873614, 1.1508363, 0.1306749, -1.1931158, 0.2093445, -0.1040642),
.Dim=c(3,7)))
```

13.3.4 External data and cross validation

The above methods simply used the *same data* for model fit assessment as well as a basis for the posterior distribution. A more challenging model assessment could be based on using *different data* that was not used when the posterior was computed.

• Model criticism based on external data: if we have a large data set, we can use part of it for bayesian inference, and the rest for model criticism. For example, use 1st part for computing posterior predictive distribution, then compare predictions with observables taken from the 2nd part. This could be extended for prior specification too: use 1st part of data to construct prior, 2nd part for computing posterior, 3rd part for model assessment. But this is feasible only if we have reasonably large data set which can be divided into meaningful parts. If we only have 10 data points in total, it makes no sense.

• Model criticism based on cross validation: take one data point out and use all the rest to predict it. Repeat this for every data point. Could be done with blocks of data as well.

In BUGS, these (especially cross validation) are more difficult to implement since we need to compute the posterior for every separate data. But using R, this could be done easily in a loop which calls BUGS for every given data set. Then it really makes sense to use BUGS from R. And we could even wrap this inside another loop which runs through different prior specifications or different models. The whole sensitivity analysis could then be run from R in nested loops. However, the analysis of all these results could be more laborious unless we have a clear criterion against which the selections could be ranked. We could also construct hypothetical data to test how the model performs under that data. But this may not be meaningful unless some meaningful criterion exists for selecting such test data. There are infinitely many data sets which have not occurred. Why should they all be equally meaningful for the current problem?

Bayesian inference is always *internally coherent*, but this does not guarantee that the model would be well *calibrated*. That is, if the model is used to predict oil prices each month, and 95% posterior predictive intervals are produced, then the model is not very good if the actual price is within the interval only 50% of the time. Calibration is obviously a very frequentist concept. It can only be applied to frequently occurring problems that are considered enough 'similar' repetitions. It would be very difficult to calibrate a model for predicting whether there is life on other planets. Either there is or there isn't. Repetitions of the universe are harder to observe, so what is a good model in that case?

There are many different approaches to model assessment, but none of them can truly validate a model (because all models are wrong). Therefore, instead of model validation, bayesian approach emphasizes **model criticism** in which the model performance is judged. Does it predict badly some important features, or only those features that are unimportant for us? This calls for declaring what aspects are important in a given problem. The most important message learned from modeling may not be the

numerical estimate of some quantity as such, but the overall insight we get in the cross light of several models.

Gelman et al:

> *It is difficult to include in a probability distribution all of one's knowledge about a problem, and so it is wise to investigate what aspects of reality are not captured by the model.*

13.4 DIC in model assessment: comparison of models

This has been used in several examples earlier, and is easy to check in BUGS, whenever DIC can be applied to the problem. Other approaches might be to consider bayesian model averaging (BMA) or bayesian nonparametrics (BNP) which avoid choosing a single parametric model. Also, reversible jump MCMC methods might be used for BMA and to get probabilities for different competing models.