12 Model assessment

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 hard to see.

There are many different approaches to model assessment, but none of them can truly validate a model (because 'all models are wrong, but some of them are useful', as statistician George Box has said). 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 does it fail with those features that are unimportant for us? This calls for declaring what aspects are important in a given problem. One approach is to define discrepancy functions which are judged against predictive distributions.

Model criticism based on substance knowledge: are the results plausible, considering what experts know about the substance of the application?

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. (Use 1st part for computing posterior predictive distribution, compare predictions with observables taken from the 2nd part. Or use 1/3 of data for constructing prior, 1/3 for computing posterior, and 1/3 for model assessment).

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 and summarize the model performance.

Model criticism based on residuals.

Model criticism based on deviation index.

Model criticism based on Bayes factors.

Bayesian model averaging: compute "probabilities for each model".

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.*

12.1 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. In hierarchical models, we can then consider model improvements in several levels from hyper priors down to conditional distributions of data. For example, in hierarchical models, 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 [4].

12.1.1 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: 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
```

12.2 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 and 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.

12.3 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.

12.3.1 Example: normal model

Consider the model $X_i \sim \text{N}(\mu, \sigma^2)$ with unknown μ and assumed (fixed) σ^2. After the data $X_1, \ldots, X_n$ has been observed, the posterior of μ is $\text{N}(\bar{X}, \sigma^2/n)$, and the predictive distribution of a new X^* is $\text{N}(\bar{X}, \sigma^2 + \sigma^2/n)$. 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)))}
```

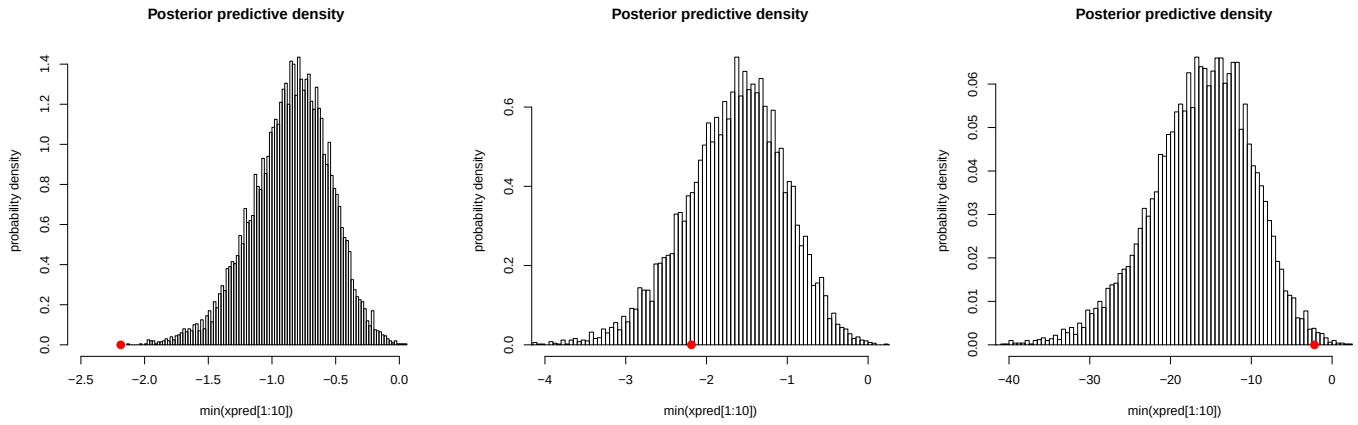


Figure 1: Posterior predictive distribution of X^*_{smallest}, based on $\sigma = 0.5$ (left), $\sigma = 1$ (middle), $\sigma = 10$ (right). Minimum data point (-2.2) shown as red bullet.

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

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

12.3.2 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:

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

12.3.3 Example: schools

The hierarchical model was based on some assumptions: (1) the normality of the Y_j, given θ_j and σ_j (the latter assumed to be known); (2) the exchangeability of the prior distribution of θ_j's; (3) the normality of the prior of each θ_j, given μ and τ; (4) the uniformity of the hyper prior distribution of μ, τ. Each of these assumptions could be assessed critically. Assumptions (3) and (4) are harder to justify *a priori* than (1) and (2). Why normal distribution of θ_j? And why should the location and scale parameters of this distribution be uniformly distributed? *Sensitivity to other choices could be checked.* Also, based on substantive knowledge of SAT scores, the inferences (posterior summaries) could be checked: is the result realistic with respect to general knowledge of SAT scores? And finally, consider posterior predictive distributions of SAT scores for 8 similar schools: are the predictions reasonable? How about predicting some test quantities:

```
model{
for(j in 1:J){
y[j] ~ dnorm(theta[j],tau.y[j])
ypred[j] ~ dnorm(theta[j],tau.y[j]) # prediction
theta[j] ~ dnorm(mu.theta,tau.theta)
tau.y[j] <- pow(sigma.y[j],-2)
}
maxy <- ranked(ypred[],8) # compare this to 28
miny <- ranked(ypred[],1)  # compare this to -3
meany <- mean(ypred[])  # compare this to 8.75
sdy <- sd(ypred[])    # compare this to 10.44
mu.theta ~ dnorm(0,1.0E-6)
tau.theta <- pow(sigma.theta,-2)
sigma.theta ~ dunif(0,1000)
}
list(J=8,y=c(28,8,-3,7,-1,1,18,12),sigma.y=c(15,10,16,11,9,11,10,18))
```

12.3.4 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 is the same as $\text{N}(\bar{X}, \sigma^2/7 + \sigma^2)$. (It might also be simulated by sequentially drawing values of μ from $\text{N}(\bar{X}, \sigma^2/7)$, and then values of X^{pred} from $\text{N}(\mu, \sigma^2)$ if we are interested in obtaining the posterior of μ as well). 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)))
```

12.4 Deviance Information Criterion, DIC

Deviance is defined as: $D(y, \theta) = -2\log(\pi(y \mid \theta))$. For example: in the case of normal density, with fixed variance, this is (up to a constant term) the same as the following test quantity:

$$T(y, \theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathrm{E}(y_i \mid \theta))^2$$

Deviance is a generalization of this more familiar mean squared error criterion. The posterior mean of deviance can be obtained approximately as:

$$\bar{D}(y) = E(D(y, \theta) \mid y) \approx \frac{1}{k}\sum_{k=1}^{K} D(y, \theta^{(k)}),$$

where k is the index of MCMC iterations. The deviance could also be calculated using the point estimate $\hat{\theta}$ (e.g. posterior mean):

$$\hat{D}(y) = D(y, \hat{\theta}).$$

$\bar{D}(y)$ is a better summary of the 'model error' than $\hat{D}(y)$ because the fit of the model is always better when computed with a fitted point estimate, than if computed with some other parameter values; and $\bar{D}(y)$ presents the 'average error' over all plausible values of θ.

The *effective number of parameters* in a model can be described by:

$$p_D = \bar{D}(y) - \hat{D}(y),$$

because this represents the effect that can be achieved by estimating θ by $\hat{\theta}$. Then, p_D can be interpreted as the number of 'unconstrained parameters' in the model. So, how many effective parameters we have in a model? Every parameter counts as one if it is estimated without constraints or constraining prior knowledge, but as zero if it is completely constrained of if all information about it comes from the prior. For example:

$$y \sim \mathrm{N}(\theta, 1) \quad , \quad \theta \sim \mathrm{U}(0, \infty).$$

In this model, θ is constrained to be positive, but otherwise the prior is uninformative. Then how many effective parameters there are? This depends on data y. If y happens to be near zero, then about half of the information in posterior density is coming from data and half from the prior constraint of positivity. So, there are about 0.5 effective parameters. But if y is very large, the constraint has no effect at all, and the effective number of parameters is one.

In hierarchical models the number of effective parameters depends on the hyper prior: a tight prior forces all group level parameters (θ_j) towards the global mean, so that the effective number of parameters is small. But a diffuse prior leads to the situation where group level parameters become separately estimated from group level data, with no shrinkage to the global mean. Then the effective number of parameters is about the same as the number of groups.

Also, p_D is the 'expected improvement of model fit' that could be achieved by estimating the model parameters. Deviance Information criterio (DIC) can be found in WinBUGS, and it can be used for some models when assessing the model fit together with the effective number of parameters. See the

presentation by Spiegelhalter in the ICEBUGS workshop for more details. (Link from OpenBUGS website). A better model has smaller DIC:

$$\mathrm{E}(D(y^{\mathrm{rep}}, \hat{\theta}(y))) \approx 2\bar{D}(y) - \hat{D}(y) = \mathrm{DIC}$$

12.5 Model comparison

12.5.1 Bayes factors

If we have a discrete set of models, say M_1 and M_2, these could be compared:

$$\frac{P(M_2 \mid X)}{P(M_1 \mid X)} = \frac{P(M_2)}{P(M_1)} \times B(M_2, M_1)$$

where

$$B(M_2, M_1) = \frac{\pi(X \mid M_2)}{\pi(X \mid M_1)} = \frac{\int \pi(\theta_2 \mid M_2)\pi(X \mid \theta_2, M_2)\mathbf{d}\theta_2}{\int \pi(\theta_1 \mid M_1)\pi(X \mid \theta_1, M_1)\mathbf{d}\theta_1}$$

12.5.2 Bayesian Model Averaging, BMA

By specifying probabilities for models $P(M_i)$, so that $\sum P(M_i) = 1$, it is possible to compute also posterior probabilities $P(M_i \mid \text{data})$, and to make predictions of observables X by averaging over the possible models

$$\pi(X \mid \text{data}) = \sum \pi(X \mid M_i) P(M_i \mid \text{data})$$

Each of the models can have different number of parameters, which makes the MCMC even more challenging (RJMCMC=Reversible Jump MCMC).

12.5.3 Bayesian NonParametrics, BNP

For real valued variables $Y_i \in \mathbb{R}$ de Finetti says:

$$P(Y_1, \ldots, Y_n) = \int_{\mathcal{F}} \prod_{i=1}^{n} P(Y_i \mid F)\ \mathbf{d}P(F),$$

where F denotes any arbitrary cumulative probability function, $F \in \mathcal{F}$. The role of the prior $P(F)$ is to give weighting to different functions so that it is a probability measure in the space of functions $\mathcal{F}$. Any single parametric function F_θ is just a special case of all functions in the set $\mathcal{F}$. This set has an infinite number of functions, whereas in BMA we had a finite number of different models (or at least countably infinite number).

12.5.4 Dirichlet-process prior and mixture distributions

Application: unknown number of components in a mixture, with unknown parameters. Observables $x_1, \ldots, x_n$ are modeled with a mixture distribution with K components, so that $\sum_{k=1}^{\infty} P(K = k) = 1$. The parameter dimension is thus changing as K changes. Changing dimensions are difficult to describe in WinBUGS, but we could choose the maximum value for K so that $K = 1, \ldots, K_{\max}$.

Dirichlet-process is a 'distribution of distributions' where random variable G is a discrete distribution:

$$G(\eta) = \sum_{i=1}^{\infty} p_i 1_{\eta_i}(\eta),$$

In other words G is made of point probabilities p_i located at points η_i on real axis. And there are infinite number of these: $\sum_{i=1}^{\infty} p_i = 1$. Every random variable G from a dirichlet-process is one of such discrete distributions:

$$G \sim \text{Dirichlet Process}(\alpha, G_0),$$

where G_0 is a baseline -distribution and parameter α determines how 'close' the discrete distribution G is to the continuous density G_0. With a large α, G would resemble the continuous density G_0, although it would still always be a discrete distribution. With a small α, G would be a discrete distribution with a large probability at some point η_i.

Probabilities p_i can be generated stepwise:

$$\begin{aligned} p_1 &= r_1 \\ p_2 &= r_2(1-r_1) \\ p_3 &= r_3(1-r_2)(1-r_1) \\ &\vdots \end{aligned}$$

where every $r_j \sim \text{Beta}(1, \alpha)$. The probabilities get smaller and smaller: interval [0,1] is first broken to take r_1, then the remainder $(1-r_1)$ is broken to take r_2, after which the remainder $((1-r_1)(1-r_2))$ is broken to take r_3, etc. This is the 'stick-breaking'-algorithm. When every point probability is attached with the random number $\eta_i \sim G_0$, we obtain the above mentioned discrete distribution with $p_i = P(\eta_i)$.

For any partition of the set S ($\eta \in S$): $B_1, \ldots, B_n$, where $B_u \cap B_v = \emptyset$ when $u \neq v$, and $\cup B_u = S$ we have the result:

$$(G(B_1), \ldots, G(B_n)) \sim \text{Dir}(\alpha G_0(B_1), \ldots, \alpha G_0(B_n)),$$

where $G(B_u)$ is a sum of all point probabilities $p_1, p_2, ...$ that are attached to the points in set B_u. Moreover, $G_0(B_u)$ is the probability, with respect to the continuous density G_0, $\int_{B_u} g_0(s)\mathbf{d}s$.

Stick breaking-algorithm is a way to produce random draws from a Dirichlet process. In WinBUGS, we need to set a finite end-point to the algorithm, so that the finite collection of probabilities is re-normalized to make this a proper distribution: $\sum_{k=1}^{K\text{max}} P(K = k) = 1$. The number of mixture components should not exceed the number of data points n. A probable number of components could be much smaller. Parameters of each component distribution are defined (and priors assigned) regardless of whether they become 'chosen' for some data point or not. (Baseline-distribution G_0). The data points x_i become allocated by using the indicator z_i as before:

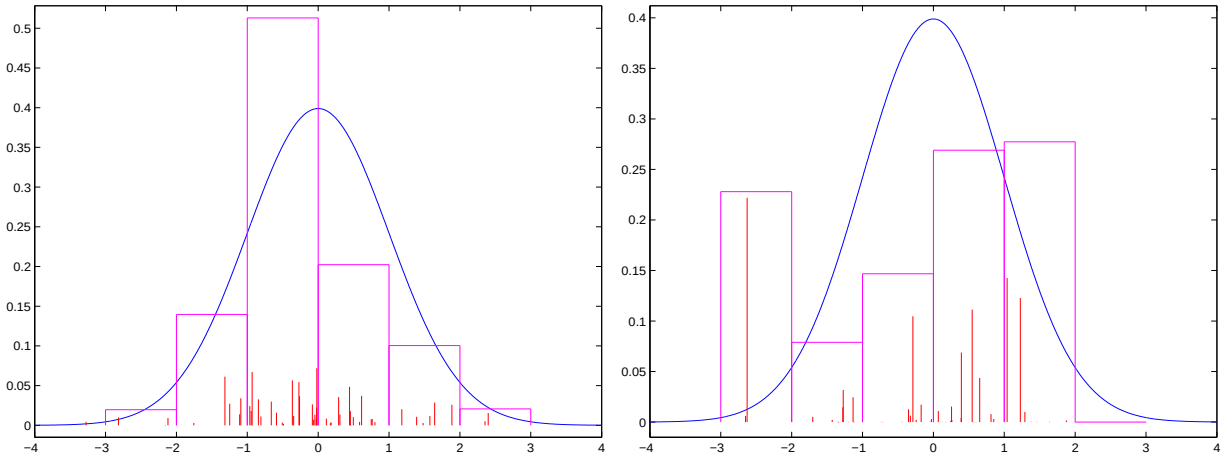
$$x_i \sim \text{Some.Distribution}(\eta_{z_i}),$$

$$z_i \sim \text{Cat}(p_1, \ldots, p_{K_{\max}}),$$

where:

$$G \sim \text{Dir process}(\alpha, G_0),$$

$$G(\eta) = \sum_{i=1}^{\infty} p_i 1_{\eta_i}(\eta).$$



Random draws from Dir-process (red vertical lines, 50 in both figures), $\alpha = 100$ (left), $\alpha = 5$ (right). A histogram was added to describe the probability that is the sum of the point probabilities under each bar of the histogram. Blue curve denotes G_0 which was here chosen as $G_0 = \text{N}(0, 1)$.

Example:

```
model{
for(i in 1:N){
pmy[i] ~ dpois(mu[S[i]])
epmy[i] <- mu[S[i]]
S[i] ~ dcat(pi[])
for(cc in 1:C){  SC[i,cc] <- equals(cc,S[i]) }
}
alpha <-1
    # Constructive DPP:
    p[1] <- r[1]
    for (j in 2:C) {p[j] <- r[j]*(1-r[j-1])*p[j-1]/r[j-1]}
    for (k in 1:C){
    r[k] ~ dbeta(1,alpha)
    mu[k] ~ dgamma(0.01,0.01)
# scaling to ensure sum to 1:
    pi[k] <- p[k]/sum(p[])
# whether a cluster is 'in use' or not:
    cluster.is.there[k] <- step(sum(SC[,k])-1)
    }
# total number of clusters:
    NC <- sum(cluster.is.there[])
# prediction:
```

```
    ind ~ dcat(pi[])
    ppmy ~ dpois(mu[ind])
 }
list(N=26,C=26,pmy=c(0,0,0,0,0,0,5,10,13,3,0,3,28,14,5,211,5,0,0,3,18,0,0,5,8,0))
```

References

[1] Berger J: The Case for Objective Bayesian Analysis. Bayesian Analysis, 2006, Vol 1, 3, 385-402.

[2] Goldstein M: Subjective Bayesian Analysis: Principles and Practice. Bayesian Analysis, 2006, Vol 1, 3, 403-420.

[3] Gelman A, Carlin J B, Stern H S, Rubin D B: Bayesian data analysis, 2nd edition. Chapman & Hall/CRC. 2004.

[4] Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, No 3, pp. 515Ű533. 2006.

[5] Jaynes E T: Probability theory: the logic of science. Cambridge university press. 2003.

[6] Sivia D S: Data Analysis, a Bayesian tutorial, 2nd edition. Oxford university press. 2006.

[7] Robert C P, Casella G: Monte Carlo Statistical Methods. Springer 1999.

[8] Congdon P: Bayesian Statistical Modelling. John Wiley & Sons, Ltd. 2001.

[9] Congdon P: Applied Bayesian Modelling. John Wiley & Sons, Ltd. 2003.

[10] Bernardo J M, Smith A F M: Bayesian Theory. John Wiley & Sons, Ltd. 2000.

[11] Aarnisalo et al. Use of results of microbiological analyses for risk-based control of Listeria monocytogenes in marinated broiler legs. *International Journal of Food Microbiology 121 (2008) 275 - 284.*