BUGS programs for the Consensus Mean Problem

The problem of determining a consensus mean based on data from multiple labs or methods has been fully addressed at NIST from the classical point of view. The result is the DATAPLOT procedure Consensus Mean.

A Bayesian solution to this problem can be obtained through the application of hierarchical Bayes models via the Markov Chain Monte Carlo simulation software called BUGS.

Here we have BUGS code for two common hierarchical models and apply it to an example from SRM 1946, Lake Superior Fish Tissue.

Model 1.

This model assumes that the multiple labs data comes from Normal distributions which have different means and different variances, that is that Y_{ij} is distributed as

$Y_{ij} \sim N(\delta_i , \sigma_i)$,

and that the means δ_i, and the variances σ_i are related.

That is that

$\partial_i \sim N(\mu \ , \tau)$,

and

$\sigma_i \sim$ IGamma(a,b)

where

$$\mu \sim N(0,10000),$$

$$\tau \sim IGamma(0.0001, 0.0001),$$

$$a \sim \exp(1),$$

$$b \sim IGamma(0.0001, 0.0001).$$

This model allows for borrowing in the estimation of both the means ∂_i, and the variances σ_i.

This means that even when some of the labs have extremely small sample sizes (even n=1) the lab variances can be estimated through the pooling of the hierarchical model.

The prior distributions above are “vague” which is appropriate when real prior information in the form of expert opinion or prior data is not available.

Analysis performed using vague prior distributions can be considered objective and is generally preferred by classical statisticians.

When real prior information is available in the form of a mean and variance of μ it can be included by simply changing the mean and variance of the Normal distribution.

If more robustness is required a t-distribution with small degrees of freedom can be substituted for the Normal.

The following is the BUGS code which will carry out the MCMC simulation to estimate the consensus mean μ.

MODEL 1: two-stage Normal hierarchical model with vague priors and borrowing for both means and variances.

```
model model1;
const
N=13, k=7;
var
theta[k], sigma[k],mu,tau, Y[N],
lab[N], a, b , Y.p[N], p.smaller[N];
data lab, Y in "fat.dat";
inits in "fat.in";

{
mu    ~ dnorm(0, 1.0E-4);
 tau   ~ dgamma(1.0E-4, 1.0E-4);
 a     ~ dexp(1.0);
 b     ~ dgamma(1.0E-4, 1.0E-4);

 for(i in 1:k) {
   sigma[i]  ~dgamma(a,b);
   theta[i]  ~ dnorm(mu, tau);
   }
for (i in 1:N) {

       Y[i] ~ dnorm(
theta[lab[i]],sigma[lab[i]]);
       Y.p[i] ~ dnorm(
theta[lab[i]],sigma[lab[i]]);
p.smaller[i] <- step(Y[i]-Y.p[i]);}}
```

Model 2.

In some cases, the assumption that the variances are related may not be appropriate, in that case the following model should be used.

$$Y_{ij} \sim N(\delta_i, \sigma_i)$$

$$\partial_i \sim N(\mu \ , \tau),$$

and

$$\sigma_i \sim IGamma(0.0001, 0.0001)$$

where

$$\mu \sim N(0,10000),$$

$$\tau \sim IGamma(0.0001, 0.0001).$$

This model does not have the property of model 1 that allows for pooling of data in the estimation of the lab variances.

For that reason, when sample sizes are small for some of the labs, the precision of the posterior estimates of the means and variances will be smaller than for model 1.

The following is the BUGS code for this model:

```
model model2;
const
N= 13, k=7;
var
theta[k], sigma[k],mu,tau, Y[N],
lab[N], vw[k], vb;
data lab, Y in "fat.dat";
inits in "fatonest.in";

{
 mu    ~ dnorm(0, 1.0E-4);
 tau   ~ dgamma(1.0E-4, 1.0E-4);

 for(i in 1:k) {
   sigma[i]  ~dgamma(1.0E-4, 1.0E-4);
   theta[i]  ~ dnorm(mu, tau);
   vw[i] <- 1.0/(sigma[i]);  }

for (i in 1:N) {

       Y[i] ~ dnorm(
theta[lab[i]],sigma[lab[i]]);
      }
        }
```

Example.

Data from an experiment which weighed the amount of solids in a sample of dried fish tissue is given in the following table:

Table 1. Amount of solids measured per sample per lab.

Lab	solids
1	28.58
1	28.98
2	28.41
2	28.58
3	28.86
3	28.72
4	29.3
4	28.7
5	28.6
6	28.64
6	28.75
7	28.78
7	29.31
8	28.71
8	28.89
9	28.5
9	28.6

Both models were applied with the results given in the following table.

Table 2. Consensus means, posterior standard deviations, and 95% HPDs for the two models.

Model 1.

Consensus mean	sd	95% HPD
2.871E+1	6.116E-2	(2.861E+1,2.884E+1)

Model 2.

Consensus mean	sd	95% HPD
2.868E+1	7.077E-2	(2.858E+1,2.884E+1)

It is clear from the table that the estimates of the consensus means are very close and that there is a slight increase in the size of the posterior HPD due to the reduced pooling of Model 2.

It is interesting to note that lab 5 had only one observation. This causes most classical consensus mean methods to fail because they require at least two data points to estimate each labs variance.

The Bayesian models can handle this situation, even in the case of model 2 where there is no pooling for variance estimation.

This is due to the fact that a proper prior (probability distribution) is used and so the variance estimate is based on the prior together with any data that is available.

For a comparison, the DATAPLOT Consensus Mean procedure was run on this data and produced the following two estimates:

The Mean of Means

Consensus mean	95% CI
2.875E+1	(2.860E+1, 2.889E+1)

Grand Mean

Consensus mean	95% CI
2.875E+1	(2.863E+1, 2.888E+1)

Michelson's Determination of the Speed of Light (1879), a case study.

Data:

100 measurements made over 18 distinct days.

24 distinct sets were made corresponding to time-of-day and day. (see graph "Effect of Measurement Day)

5 runs of 20 measurements each, possibly adjustments were made to the apparatus between runs. (see graph "Measurement Runs")

Air temperature is given for each measurement. (see graph "Temperature Effect")

Data quality:
given as “good”, “intermediate”, and “poor”. (see graph “Data Quality)

Data Analysis:

Least squares fit to 4 different models in preparation for the Bayesian analysis.

Classical significance tests were used to pick a model. The most complex model, one which has terms for temperature, run and set and which allows different sample variances for different data quality was selected.

The Model:

$$Y_i = \alpha + \beta\left(T_i - \overline{T}\right) + r_i + s_i + e_i$$

Bayesian Hierarchical Model

α, β both have non-informative prior distributions.

r_i ~ normal(0, σ_r)
s_i ~ normal(0, σ_s)

all σ have noninformative prior distributions.

Results:

1. The error variances for the three different data quality classes appear to be different but do not fit the “good”, “medium”, “worst” categories of quality.
2. The standard deviations corresponding to the “run” and “set” variables are different. Note that the distribution of the “set” standard deviation is more diffuse (i.e., less info).
3. There is a shift in the mean value of the measurement due to “run”.
4. The effect of “set” on the mean measurement is minimal but there are some sets which indicate that something out of the ordinary happened.
5. The posterior distribution of   , that is the mean measurement not

adjusted by the other variables, shows a 95% posterior probability interval close to that of Michelson/s published value.

Goodness-of- fit testing

Posterior predictions were made for each of the 100 measurements. The plot shows 95% posterior predictive probability intervals.

The fit seems adequate.