BRAC University Journal, Vol. III, No. 1, 2006, pp. 65-69

USING BAYESIAN APPROACH TO OPTIMIZE SOME MICROBIOLOGY DATA

Ahmed Hossain
Department of Public Health Sciences
University of Toronto
Ontario, Canada
Email: ahmed.hossain@utoronto.ca

ABSTRACT

The main purpose of this paper is to use Bayesian technique for identifying the conditions under which the bacteria growth is optimal. Such kind of analysis is important in microbiology when bacteria need culture in an optimum manner, so they can be identified, before adequate antibiotics can be developed. The dataset was collected under three different conditions which can be found in Binnie (2004). Therefore our interest lies in investigating how these three covariates are affecting the measurements of Bacteria count.

Key words: 3-way ANOVA, Posterior distribution, MCMC, Convergence diagnostic.

I. INTRODUCTION

The paper mentions the purpose of the original study where the data set, generated by Copper (1999) is the identification of the conditions under which the bacteria growth is optimal. This is very important in microbiology and the analysis of bacteria, since during culturing scientists need to ensure that bacteria will grow as fast as possible if they are present in the sample. Here in this paper we will try to investigate how the different covariates are affecting the outcome (bacteria counts). Different models representing the relationship of the predictors with the outcome will be examined and the best options will be presented. All the proposed models will be tested under a Bayesian perspective, with the use of MCMC algorithms. The goal of this paper is to show an application of Bayesian technology in a design of experimental studies. The paper will be broken into 5 sections: understanding the data, some descriptive statistics of the data, model and prior specification for the data, results, analysis and diagnostics, and finally conclusion.

II. DATA AND VARIABLES

The dataset contains measurements of bacteria **counts** following the culturing of five strains of a bacterium called *Staphylococcus Aureus*. The measurements correspond to millions of colony forming units (CFU). In addition, values of 3 covariates are included. **Time** of incubation, which can be 24 or 48 hrs, **temperature,** which can be 27, 35 or 43 degrees, and **concentration** of tryptone (a nutrient), where possible percentage values are 0.6, 0.8, 1.0, 1.2, 1.4. One can notice that this constitutes a factorial design with no replicates.

III. DESCRIPTIVE STATISTICS

In order to identify any existing relationships between the covariates and the outcome, as well as other interesting patterns, some preliminary analysis was performed. The data were plotted in different ways and univariate statistics were calculated. As an example, one can see the box plots of the 3 covariates against the 5 different

counts in Figure 1. The box-plots show that in general, time a positive effect in the bacteria growth is expected. Another interesting result is the fact that in all 5 strains, middle level temperature (35 degrees) gave the highest counts. It is also interesting the fact that the variances show considerable differences among the different 5 strain counts. Also, for the concentration, again the effect is positive until it reaches 1.2%, except for the 5th strain, where 1.4% gives the most counts. One can notice again, the big differences in the variance among different concentration levels and the 5 strains.

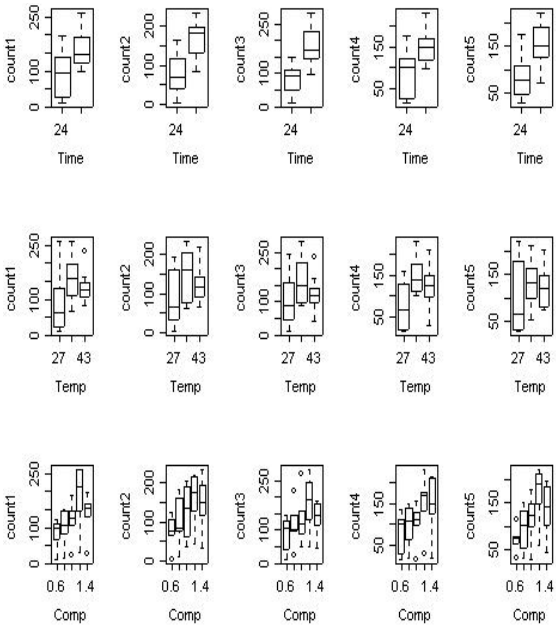

Figure 1. Box-plots of time, temperature and trypton concentration, against the growth (given in counts of millions of colony forming units) of 5 different strains of *Staphylococcus Aureus*

From these plots one can infer that the 3 covariates are actually important for the growth of the bacteria and they are affecting the outcome. Going one step further, we looked for evidence for interaction effects for these covariates and the outcome. Interaction plots provide some insight about this. Figure 2 shows the interaction plots of the 3 covariates against the counts of the 5 strains. Despite the irregularity of the plots, one can observe the points of evidence of interaction effects between the different covariates. It is also interesting that these effects are not uniform in all 5 strains.

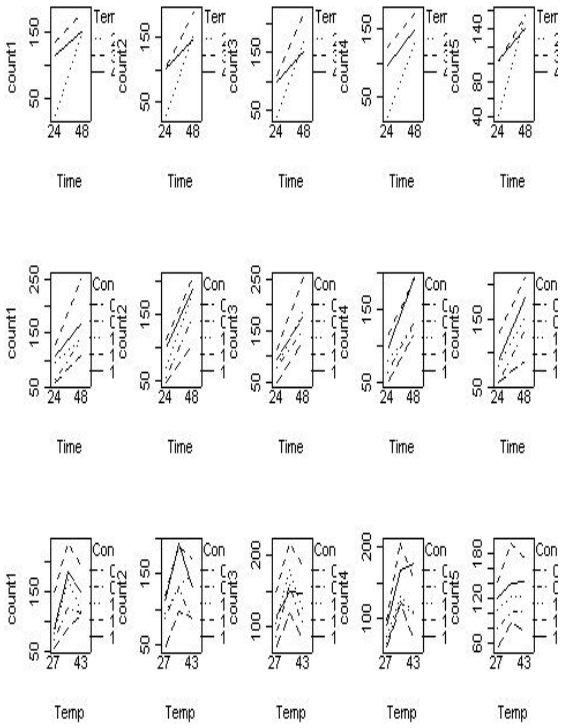

Figure 2. Interaction plots between the different covariates.

IV. MODEL AND PRIOR SPECIFICATION

Even though the original data contained data from 5 different bacteria strains and it was of interest how the results of the analysis differ for each individual strain, it is clear that going though the analysis 5 times does not add any significant scientific interest under the scope and purpose of this paper. Therefore it was decided to carry the analysis in 3-way ANOVA framework where strain is considered as one variable.

Taking into account the observations from the descriptive analysis as well as the factorial design of the data, a 3-way ANOVA model was selected to fit these data. Regarding the outcome, even though it is referred to as "counts", it is rather continuous with relative large values. Therefore, we assumed that the means of the measurements follow Normal distributions.

For the ANOVA model orthogonal parameterization was chosen for the different effects. Under these conditions the model for the mean of the outcome contains an intercept effect, three effects for the three covariates and three effects for the three interaction effects. For each

one of the interaction effects, the contrast matrix was the Kronecker product of the matrices of the individual effects. One of the questions that need to be answered is whether the interaction effects are adding anything significant to the model. In order to answer this question and check the different models we applied the Kuo and Mallick method of Bayesian Factors ([3]). Since we are only interested in how much each one of the three interaction terms contribute to the model, we have added the parameters del[1], del[2] and del[3] into the three interaction effects. So the rule that updates the mean of the outcome has become:

mu[i]<- conceff[i]+
del[1]*timeVtempeff[i]+del[2]*timeVconceff
[i]+del[3]*tempVconceff[i]……. (1)

In order to be able to compare the different terms in a fair way, the matrices have to be not only orthogonal but also orthonormal, meaning that the column vectors should have length equal to 1. These two conditions can be achieved when using the polynomial contrast matrices. Initially, standardization was not applied to the outcome of the model, but this generated unsatisfactory results, were all the models appear to have equal probability, and they were undistinguishable by this method. Therefore, additional centering was applied to the outcome, by subtracting off the mean and dividing it by the standard deviation.

Regarding the priors for the model, for the parameters for each one of the simple and interaction effects, weak priors were chosen, following Normal distributions with means equal to zero, while precision is given fixed value equal to 1/16, in accordance to Kuo and Mallick's suggestions. Here all the prior specifications have been taken by considering the conjugacy of the parameters involved in the model (1). Therefore the posterior distribution of the parameters will be tractable. Now using the numerical computational algorithms given in the software WINBUGS, Bayesian methods are now being applied in this hierarchical model.

V. RESULTS AND DISCUSSION

The model described above was executed for 39,000 iterations. 2000 iterations were used as burn-out. 3 chains were used for the model. Summary statistics results were calculated for "mod" and "del" parameters. The results can be found in Table 1. The results show that interaction term between time and temperature appears more than 95% of the time, while the rest are very close to zero. Similarly, the model that contains the interaction effect term between time and temperature, appears again 95.7% of the time, with all the rest being close to zero. The higher of the rest is the model where no interaction term is included. This model appeared 4.3%. So comparing the former (M1) with the latter model (M2), we have, P(M1/Data) = 0.957, P(M2/Data) = 0.043, so the Bayesian factor is B = (0.957)/(0.043) = 22.26 and so, according to Kass and Raftery ([4]), there is a strong evidence in selecting model M1, over M2. Therefore, for the final phase of the analysis, M1 (including only the interaction effect between time and temperature), was chosen.

In the final step the selected model was run again in WinBUGS. The model used for mu[i] was mu[i] <- b.0 + timeeff[i]+ tempeff[i] + conceff[i] + timeVtempeff[i]

The model was updated 15000 times using three different chains. The first 2000 iterations were discarded as burn out.

Table 1: Summary Statistics (Calculation of Bayes factor).

Node	Mean	SD	MC error	2.5%	Median	97.5%	Start	Sample
del[1]	0.9571	0.2027	0.009755	0	1	1	2001	39000
del[2]	1.53E-04	0.0124	1.09E-04	0	0	0	2001	39000
del[3]	0	0	2.92E-13	0	0	0	2001	39000
mod[1,1,1]	0	0	2.92E-13	0	0	0	2001	39000
mod[1,1,2]	1.28E-04	0.01132	1.06E-04	0	0	0	2001	39000
mod[1,2,1]	0	0	2.92E-13	0	0	0	2001	39000
mod[1,2,2]	0.9569	0.203	0.009754	0	1	1	2001	39000
mod[2,1,1]	0	0	2.92E-13	0	0	0	2001	39000
mod[2,1,2]	2.56E-05	0.00514	2.57E-05	0	0	0	2001	39000
mod[2,2,1]	0	0	2.92E-13	0	0	0	2001	39000
mod[2,2,2]	0.0429	0.2026	0.009748	0	0	1	2001	39000

Table 2: Summary Statistics (Final Model).

Node	Mean	SD	MC error	2.5%	Median	97.5%
b.0	7.44E-05	0.07005	3.48E-04	-0.1375	-2.00E-05	0.1393
b.conc[1]	0.9233	0.1724	8.21E-04	0.5815	0.923	1.267
b.conc[2]	-0.3339	0.1714	7.87E-04	-0.6735	-0.3334	0.004391
b.conc[3]	-0.5652	0.172	8.96E-04	-0.9048	-0.5656	-0.2259
b.conc[4]	-0.397	0.1716	8.34E-04	-0.7369	-0.3971	-0.05883
b.temp[1]	0.9013	0.2648	0.001356	0.3771	0.9011	1.424
b.temp[2]	0.2746	0.2688	0.001428	-0.2548	0.2745	0.8076
b.time[1]	-0.2951	0.6228	0.003201	-1.519	-0.292	0.9278
b.timeVtemp[1]	-1.383	0.3809	0.00203	-2.143	-1.38	-0.6314
b.timeVtemp[2]	1.927	1.126	0.005769	-0.2911	1.925	4.145

The most straightforward approach for assessing convergence is based on simply plotting and inspecting traces of the observed MCMC sample. If the trace of values for each of the parameters exhibits asymptotic behavior over the last few iterations, this may be satisfactory.Further, the results were processed in BOA where the convergence diagnostics were calculated. All the results are presented in Table 3. Three chains have been run each with 1000 MCMC samples to get the convergence of the estimates. Brooks, Gelman and Rubin convergence diagnostic based on a multi-chain comparison between-chain (B) and within-chain (W) dispersion of samples, in an analysis of variance approach. And as a default from the boa.menu() function from boa package for R software it takes 2nd half of the chain (5001-10000 samples) to calculate the estimates and quantiles. It is apparent from the scale reduction factors that each estimates are very close to 1. Also it is seen from the values of 97.5% quantiles that the values are approximately 1.0 which means that effective convergence may be diagnosed. This suggests that 5000 (half) iterations were sufficient to achieve convergence for the parameters.

Table 3: Corrected Scale Reduction Factors for Gelman, Brooks, and Rubin Convergence Diagnostics

Effects	Estimate	0.975
b_conc_1	0.999976	1.000047
b_conc_2	1.000002	1.000076
b_conc_3	1.000061	1.000395
b_conc_4	1.000062	1.000246
b_temp_1	1.000033	1.00009
b_temp_2	1.000554	1.001462
b_time	1.000129	1.000497
b_timeVtemp_1	1.00032	1.001116
b_timeVtemp_2	1.000105	1.000475
b0	1.000035	1.000312

Finally, the posterior distributions for the parameters were generated. The summary statistics are presented in Table 2. It is apparent from the table 2 that MCMC error for the mean estimates are vary low and the coverage probabilities are not wider for the parameter though the standard deviation is showing quite higher values. Therefore the Bayesian estimates can be improved by studying with prior values though we are getting pretty reasonable estimates with these vague priors.

VI. CONCLUSION

This paper suggests a guide for using Bayesian methods in Microbiology data which have been illustrated with an example of 3 way ANOVA model. Bayesian methods are now being applied in diverse and complex design of experiments by the advancement of numerical and computational algorithms. While the use of Bayesian methods is clearly increasing in quantity and quality, the authors would like to caution users in the areas of reference priors, sensitivity analysis, and inferences. For the reference priors, we would like to encourage more use of experts and more use of previously collected data to formulate prior distributions. Sensitivity analysis is needed to conclude that the results are robust to prior specification.

VI. REFERENCES

1. Binnie, N. (2004). “Using EDA, ANOVA and Regression ot Optimise some Microbiology Data”, *Journal of Statistics Education,* **12**(2).

2. Cooper, G. (1999). "The Efficiency of the Recovery of Methicillin Resistant *Staphylococcus aureus* from Salt Enrichment Broth," Bachelor of Applied Science project,

Department of Applied Science, Auckland University of Technology.

3. Kuo, L. and Mallick, B. (1998), "Variable Selection for Regression Models", *Sankhya B,* **60**, 65-81.

4. Kass, R.E. and Raftery, A.E. (1995). "Bayes Factors", *Journal of the American Statistical Association*, **90**, 773-795.

5. Smith B.J. (2004). "BOA Version 1.1 User's Manual", Department of Biostatistics, University of Iowa College of Public Health.