

## Statistical Research

The Department of Statistics in collaboration with Center for Statistical Science, Brown University, USA is currently working on development and evaluation of “alternative markers for HIV staging in resource-limited settings: evaluation of TLC as a surrogate for CD4 counts”. The Department of Statistics also in collaboration with Institute for Research in Medical Statistics, ICMR, New Delhi, undertaken a study entitled “Usage and Acceptability of ISM&H” in CGHS and private hospitals, dispensaries at Chennai city and its suburban areas. The study also covers both government (CGHS) and private allopathic hospitals. Separate schedules were used for hospitals, dispensaries, inpatients and outpatients.

### **Studies completed:**

#### **Mixed effects modeling of paraplegia data using Markov chain Monte Carlo method**

Markov Chain Monte Carlo (MCMC) is a powerful technique for performing integration by simulation. In recent years MCMC has revolutionized the application of Bayesian statistics. Many high dimensional complex models, which were formally intractable, can now be handled routinely. MCMC has also been used in specialized non-Bayesian problems. The use of MCMC was first introduced in statistical mechanics to study the equation of the state of a

two-dimensional rigid sphere system. The choice made by Metropolis was one of many other possibilities. Now this method has become a miraculous tool of Bayesian analysis and the flag of what has been called as the model liberation moment. Bayesian calculations not analytically tractable can be performed once a likelihood and prior are given. For non-Bayesian applications MCMC is considered as a very powerful numerical device in likelihood analysis or decision theory. MCMC methods have been successfully used to overcome problems caused by missing data when using small networks for conventional statistics. All MCMC methods are ways to produce a stochastic process, which has a desired distribution as its stationary distribution. The theory of stochastic processes tells as that the empirical average of a function of the stochastic process will converge to the expectation of that function under the desired distribution. MCMC is the idea of using simulations of the Markov Chain to approximate expectations by sample averages from the equilibrium distribution also called invariant distribution, stationary distribution or ergodic limit of the Markov Chain.

The two most commonly used MCMC algorithms that are applicable to both discrete and continuous systems are: (i) Gibbs Sampler and (ii) Metropolis Algorithm. The Gibbs sampler also known as the heatbath algorithm, is conceptually the simplest of the Markov chain sampling methods, but has come into prominence only recently. It is widely applicable to problems where the variables take on values from a small finite set,

or have conditional distributions of a parametric form that can easily be sampled from.

The basic operation used in the Gibbs sampling algorithm is the generation of a random value for some component of the state,  $X_i$ , from its conditional distribution given the current values of all the other components,  $X_j$ , for  $j \neq i$ . The speed of the algorithm depends crucially on whether this operation can be done quickly. For discrete components that take on values from a small set, the usual approach is to simply calculate the joint probabilities of all the states in which  $X_i$  takes on its various possible values, while the other  $X_j$  remain fixed at their current values. The conditional distribution for  $X_i$  is then found by normalizing these probabilities so they sum to one, and a new value for  $X_i$  is picked from this distribution. For complex and multimodal problems, the Metropolis algorithm is more appropriate. It has proved to be a flexible tool that is applicable to a wide range of problems. However when the required conditional distributions can be sampled from easily, the Gibbs sampler may be preferred, and when it is difficult to decompose the state into local components that are not too dependent, the dynamical algorithms may be more attractive.

Many problems such as generalized linear models and hierarchical models direct simulation is not possible even with two or more steps. Until recently approximating the desired distribution by normal or transformed normal distributions from which direct simulation can be drawn has attacked these problems. In recent years iterative simulation methods such as MCMC have been developed to draw from general distributions without any direct need for normal approximation. The advantage of these iterative methods is that

they can be setup with virtually any model that can be setup in statistics. The main limitation is that they currently require extensive programming and debugging. In Bayesian posterior distribution, the goal of iterative simulation is the inference about the target distribution and not merely some moments of the target distribution. So it is desirable to choose starting points that widely dispersed in the target distribution over dispersed starting points are an important design feature of MCMC for two major reasons:

1. Starting far apart can make lack of convergence apparent.
2. Starting over dispersed can ensure that all major reasons of the target distributions are represented in the simulation.

The chain of the class of models where MCMC is easy to use, assessing the convergence, good guidelines for starting values. Many authors extensively discuss methods assessing the behavior of the chain and useful software.

### **An application to medical data**

We illustrate the application of MCMC with a mixed model to data obtained from patients with Pott's paraplegia. This application is complicated with much observation on few patients that Markov Chain simulation methods are the most effective tool for exploring the posterior distribution. In this study 33 patients were admitted – 8 patients received only chemotherapy (streptomycin, rifampicin, isoniazid, pyrazinamide and ethambutol) and 21 received chemotherapy with surgery. Out of the 21 patients who received surgery, 5 received costotransversectomy and the remaining received modified Hong Kong surgery. Complete neurological assessments were done on admission, daily for 3 days and there after on

alternate days till 2 weeks, weekly till 3 months, monthly till 9 months and 3 monthly thereafter. We have considered a total of 32 measurements on each patient up to 24 months.

### Finite Mixture Likelihood and Hierarchical Population Models

A brief review of the basic statistical approach is described in the following section. To address the problem of modeling the neurological responses, the following basic model was fit. The neurological score is described by random effect model in which the responses  $Y_{ij}$  ( $i = 1, 2, \dots, 32$ ) of chemotherapy regimen patients  $j$  ( $j = 1, 2, \dots, 8$ ) are normally distributed with distinct mean  $\alpha_j$  and common variance  $\sigma_y^2$ . To reflect the response of surgery regimen ( $j = 9$  to  $29$ ), the scores are modeled as a two compartment mixture with probability  $(1-\lambda)$  for costotransversectomy patients which are assumed to be normally distributed with mean  $\alpha_j$  and variance  $\sigma_y^2$  and with probability  $\lambda$  for modified Hong Kong surgery patients with mean  $\alpha_j + \tau$  and the same variance  $\sigma_y^2$ .

The comparison of the components of  $\alpha = (\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{29})$  for chemotherapy patients verses surgery patients addresses the magnitude of

decrease in recovery period and increase the response score. A hierarchical parameter  $\beta$  measuring the activity was included. Specifically variation among the individuals is modeled by having the means  $\alpha_j$  follow a normal distribution with mean  $\mu$  for chemotherapy and  $\mu + \beta$  for surgery patients with each distribution having variance  $\sigma_\alpha^2$ . i.e. the mean of  $\alpha_j$  in the population distribution is  $\mu + \beta S_j$  where  $S_j$  is an indicator variable with 1 if the person  $j$  is surgery and 0 otherwise. The Bayesian model with an improper uniform prior distribution on the hyper parameters  $\phi = (\sigma_y^2, \sigma_\alpha^2, \lambda, \mu, \beta, \tau)$  as given by Gelman and Rubin were followed.

A sample of 100 points were drawn at random from the distribution and used as a starting point for the Expectation Conditional Maximization (ECM) algorithm to search for modes as given by Gelman and Rubin. The posterior distribution was approximated by a multivariate t distribution centered at the major mode of ECM with covariance matrix as the inverse of negative of the second derivative matrix of the log posterior density. Another 1000 independent samples were drawn and importance resample subset of 10 was used as a starting point for independent Gibbs samplers.

**Table XI : Posterior quantiles and estimated potential scale reduction factors for parameters.**

Parameter	After 10 iterations				After 100 iterations			
	2.5%	50%	97.5%	√R	2.5%	50%	97.5%	√R
$\lambda$	0.10	0.29	0.58	2.4	0.21	0.22	0.31	1.01
$\tau$	0.52	0.86	1.35	2.1	0.72	0.90	1.10	1.02
$\beta$	0.27	0.51	0.65	1.7	0.36	0.44	0.62	1.02

Table XI displays the posterior inferences and potential scale reduction factor for selected parameters after 10 iterations and 100 iterations. Only three-parameter estimate values are presented. After 100 iterations, the potential scale reduction factors were approximately 1 for all parameters in the model. The other hyper parameters were also estimated.

The MCMC methods provide a powerful statistical tool and have revolutionized statistical inference specifically Bayesian inference over the past few years. The ability to fit complicated models with little programming effort is in fact a key advantage of MCMC methods. The MCMC simulation should be undertaken after the problem

has been approximated and explore using simple methods. There are a variety of methods of constructing efficient MCMC algorithms. However the implementation of many of these methods requires some expertise. Even though the recent works focus on construction of samples without this problem using exact samples, considerable work is still needed on the implementation issues to high dimensional databases. Comparison of different approaches for choosing initial values and convergence criteria needs further work.

**(Contact person: Dr.P. Venkatesan, e-mail: [venkatesanp@icmr.org.in](mailto:venkatesanp@icmr.org.in))**

