Report on The Course "MARKOV CHAIN MONTE CARLO METHODS IN GENETICS": James Renie Bequest Committee

Victor Martinez

June 12, 2001

1 Introduction

The objective of the course was to deal with the theory behind the use of Markov Chain Monte Carlo Methods, specifically the Gibbs Sampler and also implementing these concepts on the study of the inheritance of Quantitative traits. This rapidly developing field in the area of applied statistics have much scope for the study of complex genetic models.

The objective of the course was fully fulfilled and it provided me with a very good introduction to the subject. Specifically, it allowed to have access to the basic background in order to understand many of the complexities of this method of estimation, when applied to very complex models of QTL detection.

2 Course Contents

The course was divided in various sub-topics that enable the attendants to understand the increase complexity of the subject. A summary of the course is given below :

PART I - Fundamentals
1.Random variables and probability distributions.
2.Functions of random variables.
3.An introduction to likelihood inference.
4.An introduction to Bayesian inference.
5.The EM algorithm.
6.An overview of Markov chain theory.
7.Markov chain Monte Carlo methods.
7.1. The Metropolis - Hastings algorithm.
7.2. The Gibbs sampler.
7.3. Reversible jump Markov chain Monte Carlo.
8.Model choice.

PART II - Applications of MCMC methods in Genetics 1.The single trait additive genetic model.
2.The mixed linear model with maternal effects.
3.Data augmentation.
4.Multiple trait Gaussian models.
5.Analysis of categorically distributed traits.
6.Joint analysis of Gaussian and categorical traits.
7.Segregation analysis models.
8.An introduction to QTL models.
9.General implementation issues and output analysis.

Initially, the course started with an outline of probability calculus, in which discrete and continuous variables were characterized in terms of their sampling distribution. This information was vital to introduce the study of

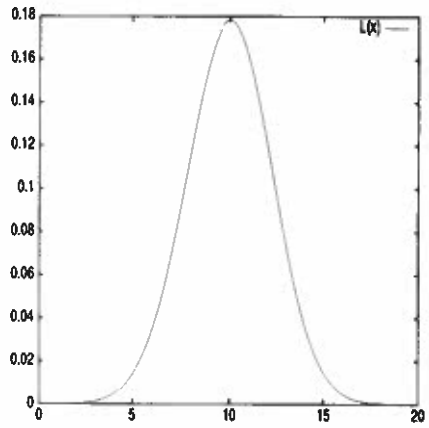
Markov Chains and both Bayesian and Frequentist, statistical approaches. In the Univariate case, binomial gama, beta, uniform and normal distribution were pursued, as well as for the Multivariate case the normal and truncated normal distributions which proved to be of much use when implementing infinitesimal and thresholds models during the course.

3 Likelihood inference

In this section likelihood inference will be sketched using as an example the estimation of the average of a variable y. For simplicity, we assume using as a data one single data point, for simplicity. If we consider a single data point (y=10) draw from a certain probability distribution (eg Normal) assuming that the variance is known (σ_y^2=25). The objective here is to draw inferences about the mean of the variable. The likelihood function can be expressed as:

$$L(\widehat{\mu}|\sigma_y^2 = 25, y = 10) = \frac{1}{(50\pi)^{-1/2}} exp[-\frac{(10 - \widehat{\mu}}{50}] \tag{1}$$

If we plot the likelihood, varying the value of the unknown μ then it is possible to obtain the maximum likelihood of μ ($\widehat{\mu}$) given the data. In this case the Maximum likelihood estimate is of course, equal to 10 (see figure 1).



Plot of the likelihood function (eq. 1) for various values of $\widehat{\mu}$.

The course also comprised the application of likelihood based techniques to gain understanding about different genetic models of use in genetics. For example, the additive genetic model was thoroughly outlined.This was especially appealing for me since the use of likelihood based techniques is much use in detecting genes using information form marker data. For example Profile likelihoods are used to map genes assuming different genetics models, and a profile likelihood similar to the one presented in the figure 1 is used to make inferences about the most likely position of the gene in the chromosome.

4 Bayesian Inference

Under Bayesian estimation, inferences about the parameters are drawn from probability distributions. ie. it is implied that parameters are random variables rather than fixed values (as in likelihood inference). The inference is carried out using prior information (information previously gathered by the researcher) which is then updated using information obtained from actual data available. The data is modeled as in the likelihood case in the form of a sampling model. This process is done to obtain the marginal posterior distribution of the variables (parameters) of interest.

The application of Bayesian statistics was hampered by the availability of tractable forms to obtain analytically the posterior distributions of the parameters of interest. Nowadays with the development of very fast computers it is possible to use a series of Monte-carlo simulation procedures to approximate the distributions of interest, i.e. the distributions of the parameters, which in some complex cases are intractable analytically. Basically Monte Carlo methods relies on the fact that increasing the number of datum generated using simulation, the greater the information and the closer the Monte Carlo estimate is from the true value. Using the information from the simulated data, it is possible to perform integration of very complex distributions without the need to have tractable forms to perform integration using analytical or numerical methods.

Using properties of Monte Carlo Markov Chains, ie a series of random variables which satisfy the condition that moving from one state (s) to the

next state (s+1) has probability associated which is the transition probability. This transition probability enable to obtain equilibrium distributions which satisfy the condition of " *ergodicity*" and "*reversibility*".

Different algorithms were cleverly developed using properties of such Markov chains. The Gibbs sampler, a especial case of the Metropolis Hastings Algorithm is derived using properties of the conditional posterior distribution of the parameters of interest which are often known. The basic steps of the Gibbs sampler can be summarized as follows :

1.- Compute the joint posterior distribution of the parameters given the data.
2.- Derive the conditional posterior distributions of all the parameters, in turn. Inspect the known distribution of the parameters, which have all close form.
3.- Simulate values of samples from distributions of close form.
4.- Update all the values of the parameters given the current values of the other samples.
5.- Run until convergence.

Following, I will give a simple example in which two parameters of a certain distribution are unknown. Here, the aim is to estimate the mean and the variance of a normally distributed variable with mean (μ) and variance (σ^2) unknown. The first step is to construct the "*posterior distribution*" of the parameters given the data, which is equal to :

$$p(\mu, \sigma^2|y) \propto p(\sigma^2)p(\mu)p(y|\mu, \sigma^2) \tag{2}$$

where $p(\mu)$ is the prior distribution of the mean, $p(\sigma^2)$ is the prior distribution of the mean and $p(y|\mu, \sigma^2)$ is the sampling model which is used to update the prior information $p(\mu)$ and $p(\sigma^2)$.

Using knowledge of probability distribution theory it is possible to deduce the which distribution follows the different parameters of interest. For the present case μ and σ^2 followed a normal and a inverse chi square distribution, respectively.

The next step is to use these distributions to generate samples using simulation, which are needed to compute the mean and the variance of the parameters. This statistics are analogous to the ML estimators of the mean and the variance, and its corresponding information, However conceptually are not comparable. Under Likelihood inference we produce ML estimators which are use to make inferences about unknown fixed values, ie the parameters. Under Bayesian Inference we focus our interest on the probability distribution of the parameters given the data. Parameters which are assumed to be random variables.

I simulate data and produce a simple piece of fortran code that implement the Gibbs sampler to obtain the probability distribution of the mean and the variance using the information generated from simulated data. The results are presented in the following table :

Table 1.- Summary statistics for μ and σ^2 over Monte-Carlo samples using the Gibbs sampler. C.I. = Confidence Interval

		μ	σ^2
Bayesian results	Mode	8.84	21.63
	Mean	8.85	21.74
	Variance	0.07	3.55
	C.I.(0.95)	8.4-9.3	18.8-25.1

Under Bayesian inference the posterior distribution can be obtained and is presented for the mean and the variance in figure 2a and 2b respectively. It may be neccesary to note that using maximum likelihood estimation would give a single point estimator, which in this case is equal to 8.9 for the simulated data and for having a measure of uncertainty, we would have to resort to asymptotic theory and derived the standard error of the mean (0.0721) (table 1). The same holds for the variance of the data.

This general example was used to outline basic principles, but the power of these methods is due to the analysis of very complex models of estimation with many effects and by which Maximum Likelihood Theory may proved to be very difficult to apply in practice.

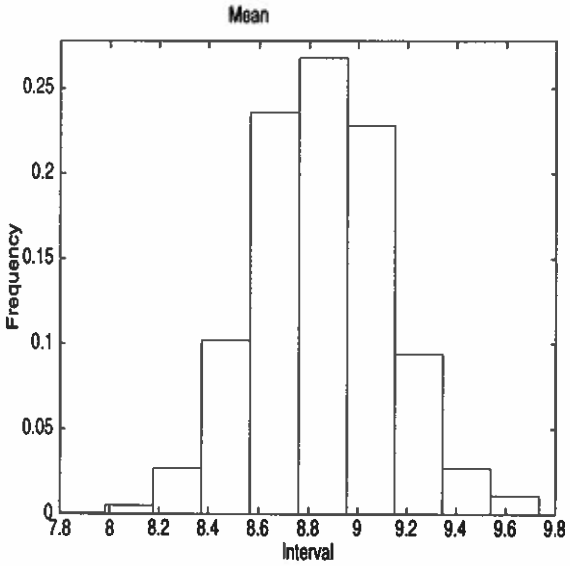
Different models were studied in detail. In the first instance, models assuming a very large number of loci affecting the trait were studied, in which the univariate mixed linear model of only additive effects was followed by maternal and multivariate normal models. Of special interest of joint analysis of normal and categorical traits was studied in much detail. The course comprise the use of Bayesian analysis of Quantitative trait loci models allowing for the number of QTL be a random variable. This was solved using

the reversible jump algorithm, which is the general case of the metropolis hastings algorithm, in which the vector of parameters is allow to changen of dimension. This is a very technical subject that was pursued in depth by the course tutor.

Conclusion

The Course provide with an excellent background to understand and use Monte Carlo Markov Chains and Bayesian inference to help the study of quantitative traits in human and animal species.

a



b

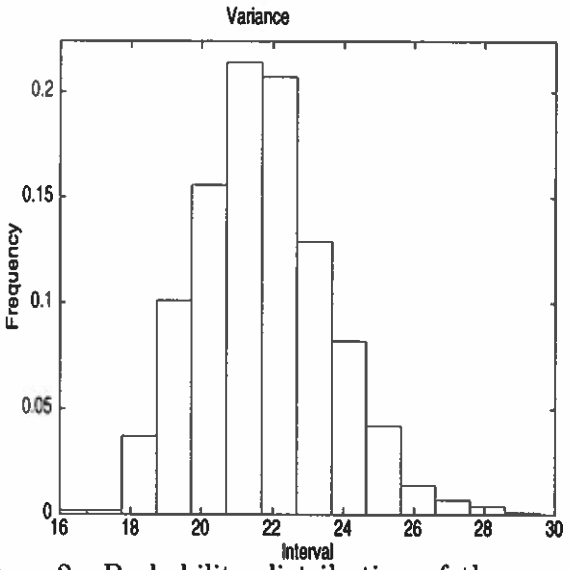


Figure 2.- Probability distribution of the μ (a) and σ^2(b) using the Gibbs sampler