Use of Bayesian Approach in Chemistry

Bayesci Yaklaşımın Kimyada Kullanımı

Research Article / Araştırma Makalesi

Turhan Menteş
Hacettepe University, Department of Statistics, Beytepe, Ankara, Turkey

ABSTRACT

Bayesian approach is a popular topic today in many fields of study in which statistics is used. The availability of stochastic simulation technique such as Markov Chain Monte Carlo makes exact Bayesian solution possible even in very complex and high dimensional models. The purpose of this short review paper is to emphasize the basic principles and to show the use of Markov Chain Monte Carlo technique for Chemistry data.

Key Words
Bayes theorem, Bayesian inference, Markov chain Monte Carlo, Gibbs sampling.

ÖZET

Günümüzde Bayesci yaklaşım istatistiğin kullanıldığı birçok alanda revaçta olan bir yaklaşımdır. Stokastik simülasyon tekniği olan Markov Zinciri Monte Carlo yönteminin varlığı, karmaşık ve yüksek boyutlu modellerde bile Bayesci çözümlemelerin elde edilmesine olanak sağlar. Bu kısa derlemenin amacı, Bayesci yaklaşımın temel ilkeleri üzerinde durmak ve Markov Zinciri Monte Carlo yönteminin kimya verileri için nasıl kullanılabileceğini göstermektir.

Anahtar Kelimeler
Bayes teorisi, Bayes çıkarımı, Markov zinciri Monte Carlo, Gibbs örneklemesi.

Article History: Received July 12, 2010; Revised November 12, 2010; Accepted November 17, 2010; Avaliable Online: February 16, 2011.

Correspondence to: T. Menteş, Hacettepe University, Faculty of Science, Department of Statistics, Beytepe, Ankara / Turkey

Tel: +90312 297 7900 Fax: +90312 297 7913 E-Mail: mentes@hacettepe.edu.tr

INTRODUCTION

Bayesian inference is the process of fitting a probability model to data and summarizing the results by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations [1]. Although Bayes' theorem can be traced back to 18th century by the work of Thomas Bayes, the modern use of Bayes theorem began 1950's. Until the 1990's, Bayesian methods found the little practical application because of the lack of computational techniques and softwares. Bayesian methods are now the tools of choice in many application areas such as Chemistry. Around two hundred literature from the areas of application, including general chemistry, chromatography, and mass spectrometry, spectroscopy, microbiology, and environmental chemistry are reported and reviewed by Hibbert and Armstrong [2].

This review is in two parts. The first is about the principles of Bayesian approach. The second is about a simulation technique called Markov Chain Monte Carlo, universally abbreviated MCMC.

Principles of Bayesian Approach

Generally, statistical inference is concerned with making decisions from the data about the unknown model parameter. Bayesian inference is made in terms of probability distributions. The main difference between Bayesian and classical approaches comes from the definition of the parameter. From the Bayesian view, the unknown quantity is a random variable and it should be represented by a probability distribution during the estimation process. Bayesian approach combines two sources of information [3]. One is the sampling information that comes from the data. The other is the prior information which reflects your experience, knowledge and expert belief about the parameter before observing any data. In Bayesian framework, posterior distribution is obtained by weighting the prior information and sampling information through Bayes' theorem. Let the prior distribution for the parameter θ be $f(\theta)$ and the likelihood function be $L(\theta;x)$. Then Bayes' theorem synthesizes the two sources of information by the simple process of multiplying. The result is the posterior distribution, which is denoted by $f(\theta/x)$. Thus,

$$f(\theta/x)\alpha f(\theta)L(\theta;x)$$

In words, the posterior is proportional to prior times likelihood. Bayesian process can be summarized by Figure 1.

There are two important tasks in Bayesian approach. One is to define the functional form of the prior distribution which conveys your belief for the parameter. For that purpose, informative

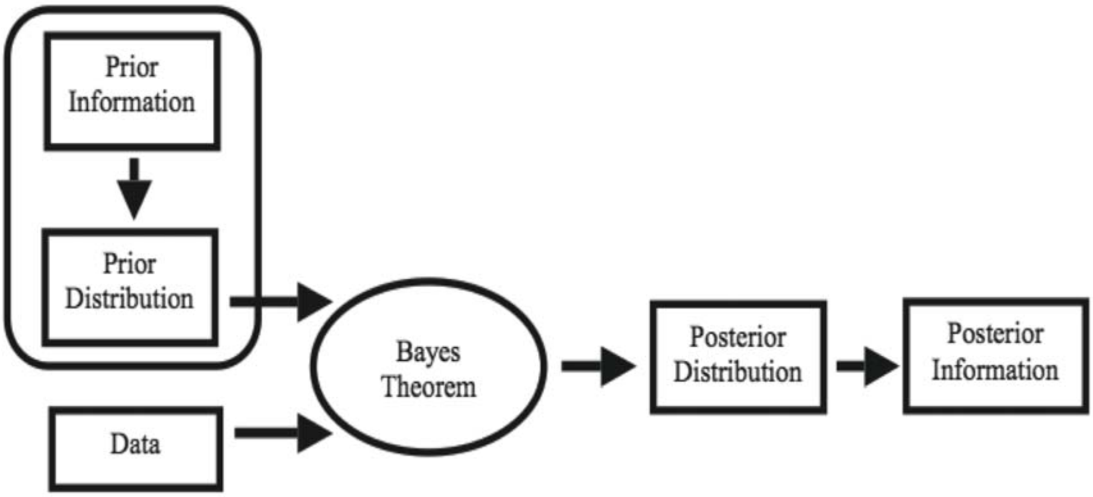

Figure 1. Bayesian framework for statistical inference.

and noninformative priors are available. While the prior information is weak, the distribution can be defined as a diffuse or a vague prior. Then the data will dominate the result. From that point, Bayesian approach uses more information than classical approach does. Bayesian inference uses both objective and subjective information. This approach allows all evidence to be taken into account in an explicit way. Beside, Bayesian techniques are particularly well suited for making decision. Bayesian statistics provides results in a more natural and intuitive form. It is an advantage of Bayesian approach. It usually reduces the sample size and, then, the cost of survey. Obtaining the posterior distribution is an important step but not the final one. One must be able to extract meaningful information from this distribution and translate it in terms of its impact on the study. This is mainly concerned with evaluation of mean, median or mode, or interval summaries given by the posterior probability intervals. For a simple and a low dimensional model, this summarization can be performed analytically. For instance, a point estimate of θ is the expectation of the posterior distribution which can be obtained analytically by taking simple integral. For detailed information about Bayesian framework, the reader is referred to O'Hagan [4], Bernardo and Smith [5], Wright and Ayton [6], Tesella [7]. In most cases, however, the complexity of the model prevents the simple solution. Nowadays, there are many problems that fall into the category of large dimensional models. Determination of posterior distributions comes down to the evaluation of complex, often high dimensional integrals. In addition, marginal posterior summarization often involves computing moments or quartiles, which leads to more integration. Many functions, equations and distributions cannot be integrated analytically. Markov Chain Monte Carlo provides an answer to difficult problems of simulation from the highly dimensional distribution of the unknown parameters that appear in complex model [8,9].

Markov Chain Monte Carlo Methods

Markov Chain simulation is a general method based on drawing values of θ from approximate distributions. The key to Markov Chain simulation is to create a markov process and run the simulation long enough that the distribution of the current draws is close enough to the stationary distribution [1,8,9]. Gibbs sampling and Metropolis algorithm work well for a wide range of problems. However, some other hybrid MCMC algorithms are also used to simulate the posterior distributions. In this short review, a brief information is given only Gibbs sampling.

Gibbs sampling is a MCMC scheme where the transition probability is formed by the full conditional distributions. Assume that the distribution of interest is $\pi(\theta)$ where $\theta=(\theta_1,\dots,\theta_d)$. Consider also that the full conditional distributions $\pi_i(\theta_i)=\pi(\theta_i/\theta_{-i})$, i=1,2,...,d are available. That means they are completely known and can be sampled from. Then the Gibbs sampling algorithm is described by the following steps [8].

Initialize the iteration counter of the chain j=1 and set the initial values for each parameters, $\theta^{(0)}=(\theta_1^{(0)},\dots,\theta_d^{(0)})$.

Obtain a new value of $\theta^{(j)}=(\theta_1^{(j)},\dots,\theta_d^{(j)})$ from θ^{j-1} through successive generation of values

$$\theta_1^{(j)} \sim \pi(\theta_1/\theta_2^{(j-1)},\dots,\theta_d^{(j-1)})$$
$$\theta_2^{(j)} \sim \pi(\theta_2/\theta_1^{(j)}, \theta_3^{(j-1)},\dots,\theta_d^{(j-1)})$$
$$.$$
$$.$$
$$\theta_d^{(j)} \sim \pi(\theta_d/\theta_1^{(j)},\dots,\theta_{d-1}^{(j)})$$

Change counter j to j+1 and return to step 2 until converge is reached.

This simulation strategy is used to obtain the values from a Markov chain. A sample from the ith component of θ is given by $\theta_{1i},\dots,\theta_{ni}$. Then point estimates of θ_i are calculated by Monte Carlo integration. Once the algorithm has been implemented, the convergence of the generated chain should be checked carefully. There are several tests for convergence such as the Geweke test, Geweke z-score, Gelman-Rubin test, Raftery-Lewis test. For the monitoring convergence, the reader is referred to Gelman and Rubin [10], Brooks and Giudici [11].

The availability of such a computational technique makes exact Bayesian inference possible even in very complex models. It is mentioned by O'Hagan [3] that generalized linear models can be analyzed exactly by the Bayesian method, whereas classical methods rely on approximations. As a result, numerous applications of MCMC have also appeared for Chemistry data and reviewed by Hibbert and Armstrong [2]. In parallel with the theoretical improvements in computational strategy, the software packages such as BUGS and R have allowed nonexperts in statistics to fit complex Bayesian models with minimal programming. Especially, WinBUGs is a powerful programme and can be freely downloaded from http://www.mrc-bsu.cam.ac.uk/bugs. However it is currently not very user friendly programme. Given the growing popularity of Bayesian methods, it is likely that more powerful, robust and user friendly software will emerge in the coming years.

CONCLUSION

Bayesian approach provides more intuitive and meaningful inferences. Stochastic simulation techniques make exact Bayesian inference possible even in very complex models. For a high dimensional space, sampling is a main step for estimating the model parameters exactly. Thus, MCMC technique is a useful tool for drawing conclusions from the posterior distributions. A key issue for a successful implementation of MCMC is the number of iteration until the chain approaches to stationary. To avoid the effect of poor/wrong choice of initial values for the parameters, the first 500-1000 draws should be thrown out from the sample. Finally, the test of convergence should be applied for more reliable and robust estimations.

REFERENCES

[1] A. Gelman, J.B. Carlin, S.H. Stern, D.B. Rubin, Bayesian data analysis, Chapman and Hall, 2004.

[2] D.B. Hibbert, N. Armstrong, An introduction to Bayesian methods for analyzing chemistry data: Part II: A review of applications of Bayesian methods in Chemistry, Chemometrics and intelligent laboratory systems, 97(2) (2009) 211.

[3] A. O'Hagan, Bayesian statistics: principles and benefits, Wageningen UR frontis series, Bayesian statistics and quality modeling in agro-food production chain, 3 (2004) 31.

[4] A. O'Hagan, Eliciting expert beliefs in substantial practical applications, The statistician, 47 (1998) 21.

[5] J.M. Bernardo, A.F.M. Smith, Bayesian theory, Wiley, 1994.

[6] G. Wright, P. Ayton, (eds.), Subjective probability, Wiley, 1994.

[7] The opportunities and advantages of using Bayesian statistics in clinical trials, http://www.tesella.com.

[8] D. Gamerman, Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference, Chapman and Hall, 1997.

[9] B.P. Carlin, T.A. Louis, Bayesian methods for data analysis, CRC Press, 2008.

[10] A. Gelman, D.B. Rubin, Inference from iterative simulation using multiple sequences (with discussion), Statistical Science, 7 (1992) 457.

[11] S. Brooks, P. Giudici, MCMC convergence assessment via two way ANOVA, J. Computational and Graphical Statistics, 9 (2000) 266.