Int. Journal of Math. Analysis, Vol. 4, 2010, no. 41, 2043 - 2049

Application of Multinomial-Dirichlet Conjugate in MCMC Estimation : A Breast Cancer Study

Geetha Antony Pullen

Mary Matha Arts & Science College
Vemom P.O., Mananthavady - 670645, India
geethapullen@gmail.com

M. Kumaran

Department of Statistical Sciences
Kannur University, Kannur
mkdss@gmail.com

Abstract

Studies have been made to investigate the familial risk of breast cancer based on a large case-control study and conclude that a small number of affected cases were due to the presence of a rare autosomal dominant allele, where as a larger number of cases reported were non genetic [7]. The proportion of affected in the population in different age groups mentioned in [7] is modeled by a Dirichlet prior in the present paper. MCMC simulation from a Multinomial-Dirichlet conjugate using R program, calculates the estimates of these proportions. These estimates support the conclusion in [7], that the general population has a high probability of getting affected at an age of 40^+ and then at an age of 70^+.

Keywords: Multinomial, Dirichlet distributions, MCMC- Gibbs sampling, Rare autosomal dominant allele, R program

1 Introduction

Numerous studies have investigated the genetic transmission of breast cancer. In some studies, affected individuals have been defined solely by the presence of breast cancer, while in other studies, breast cancer cases have been divided into subgroups based on menopausal status [2,11], age at onset [5,6], bilaterality [12,13], time interval between first and second primary tumors for bilateral cases [13], and occurrence of other cancers [11]. The findings in Claus

et. al (1991)[7] lends support to the hypothesis that the distribution of breast cancer cases in the general population includes a small number of genetic cases and a larger number of non genetic cases. The non genetic reasons may be attributed to the occurrence of other cancers like ovarian cancer, many other environmental factors like alcohol, cigarette consumption, food adulteration and many other socio demographic parameters.

In the present study, we propose *beta distributions* to model the proportion θ_i of a person in the age group i to be affected. The study, based on a *Gibbs sampling* from *Multinomial-Dirichlet distributions* agrees with the conclusion in Claus et. al (1991)[7], that, there is higher probabilities for the age groups 40^+ and 70^+ to be the ages at onset of disease, for those getting affected.

1.1 Multinomial-Dirichlet distributions in Bayesian analysis

The variable $X = (X_1, X_2, \ldots, X_p)$ has the *Multinomial distribution* $X \sim Multinomial(N; \theta_1, \theta_2, \ldots, \theta_p)$, if $\sum \theta_i = 1$, $\Sigma X_i = N$ and

$$P(X \mid N; \theta_1, \theta_2, \ldots, \theta_p) = \begin{pmatrix} N \\ x_1 x_2 \ldots x_p \end{pmatrix} \theta_1^{x_1} \theta_2^{x_2} \ldots \theta_p^{x_p}.$$

The marginal distribution of each X_i being
$X_i \sim Binomial(N, \theta_i)$; $0 \leq \theta_i \leq 1$, $i = 1, 2, \ldots, p$.
The conjugate prior of the *Multinomial distribution* is the *Dirichlet distribution*, the multivariate generalization of beta distribution. Hence the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ has a prior distribution given by,
$\theta \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_p)$ whose density function is given by
$g(\theta \mid \alpha_1, \alpha_2, \ldots, \alpha_p) = \frac{\Gamma(\Sigma \alpha_i)}{\prod(\Gamma \alpha_i)} \theta_1^{(\alpha_1 - 1)} \theta_2^{(\alpha_2 - 1)} \ldots \theta_p^{(\alpha_p - 1)}$, $\alpha_i > 0$ and $0 \leq \theta_i \leq 1$, $\sum \theta_i = 1$.
Marginally, $\theta_i \sim Beta(\alpha_i, \Sigma_{k \neq i} \alpha_k)$, $i = 1, 2, \ldots, p$. The posterior distribution of θ given X is, $\theta \mid x \sim Dirichlet(x_1 + \alpha_1, x_2 + \alpha_2, \ldots, x_p + \alpha_p)$.

1.2 Gibbs sampling

Since the advent of MCMC methods in the early $1990's$, the Bayesian computations have been extended to a large and growing applications. Many authors discuss posterior simulation in detail [3,4,8,9,10]. Gibbs sampling is a popular MCMC algorithm commonly used when the conditional posterior distributions are easy to work with. It permits the analysis of any statistical model possessing a complicated multivariate distribution to be reduced to the

analysis of its much simpler and lower dimensional full conditional distributions. Hence, in the particular situation,
$\theta \mid x \sim Dirichlet(x_1 + \alpha_1, x_2 + \alpha_2, \ldots, x_p + \alpha_p)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$.
Gibbs sampling produces a sequence of samples, ${\theta_i}^{(r)}$ for $r = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, p$, with the property that $\frac{\sum_{r=1}^{m} \Phi({\theta_i}^{(r)})}{m}$ converges to $E[\Phi(\theta_i) \mid x]$ as m goes to infinity (provided $E[\Phi(\theta_i) \mid x]$ exists); where $\Phi(\theta_i)$ is a function of the model's parameters. Gibbs sampling produces this sequence by iteratively sampling from the posterior conditional distributions. Hence, at the $(t+1)^{th}$ iteration,
$\theta_1^{(t+1)} \sim g(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_p^{(t)})$
$\theta_2^{(t+1)} \sim g(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_p^{(t)})$
$\ldots$

$\theta_p^{(t+1)} \sim g(\theta_p \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)})$,

and forms, $\theta^{(t+1)} = ({\theta_1}^{(t+1)}, {\theta_2}^{(t+1)}, \ldots, {\theta_p}^{(t+1)})$.

2 Main Result

Segregation analysis and goodness of fit tests of genetic models provided evidence for the existence of a rare dominant allele (A) leading to increased susceptibility to breast cancer [7]. The effect of genotype on the risk of breast cancer is shown to be a function of a woman's age. The life time risk of breast cancer for carriers of the abnormal allele (A) was estimated to be nearly 100% [7]. Thus, persons with both the genotypes AA and Aa will be affected by the disease at some time during their life period. They estimated the proportion in the population affected by breast cancer at different age groups as given in table 1:

Age (years)	Probability
20 - 29	0.0167
30 - 39	0.1277
40 - 49	0.2314
50 - 59	0.1719
60 - 69	0.1266
70 - 79	0.2709
80 +	0.0548
Total	1.0000

Table 1: *Proportions*

Because of the extremely low occurrence of breast cancer in women before the age of 20 years, the probability of becoming affected with breast cancer before 20 years is assumed to be zero. In the present study, we considered the seven age groups as seven classes of a *multinomial distribution* (see table 1).

$$X_i^{(r)} = \begin{cases} 1 & \text{if } r^{th} \text{ person has breast cancer at an age belonging to age group } i \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, 2, \ldots, 7$; $r = 1, 2, \ldots, N$.
$\theta_i = P(X_i^{(r)} = 1 \mid X_{i-1}^{(r)} = 0); i = 1, 2, \ldots, 7$, with $X_0^{(r)} = 0$.
Different age groups are the ages at onset of the disease and age group 80^+ can be thought of as being not affected and hence, $\theta_7 = 1 - \sum_{i=1}^{6} \theta_i$.
Also, each $X_i^{(r)} \sim Bernoulli(\theta_i); i = 1, 2, \ldots, 7$, and,
$\theta_i = P(\text{occurrence of breast cancer in the age group i }) = E(X_i^{(r)})$.
Also, $X_i = \sum_r X_i^{(r)} \sim Binomial(N, \theta_i); i = 1, 2, \ldots, 7$.
Hence, $X = (X_1, X_2, \ldots, X_7) \sim Multinomial(N, \theta_1, \theta_2, \ldots, \theta_7)$.

A lot of non genetic parameters may lead to the occurrence of the disease to a non carrier (a person with two normal allele of the gene - aa); for example, the occurrence of ovarian cancer, late marriage, infertility etc. It is well known that the beta (dirichlet) is the conjugate prior of the binomial (multinomial). One of the many applications of the beta distribution in Statistics is in the context of bioassay. The most common use of the beta in bioassay is in modeling dispersion of a binomial parameter. A typical application is in quantal bioassay, where 'success' may constitute detection of a tumor of a certain type in a certain organ. In more general settings, such as multinomial response vector, the multivariate generalization of the beta distribution, the Dirichlet, is often used as a model for the response vector; in this case the beta will appear as a model for the marginal distributions.
Hence, we let, $\theta_i \sim Beta(\alpha_i, \beta_i); i = 1, 2, \ldots, 7$, where $\beta_i = \sum_{k \neq i} \alpha_k$, and

$\theta = (\theta_1, \theta_2, \ldots, \theta_7) \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_7)$. Hence the prior distribution of θ is
$g(\theta) = Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_7)$.
The posterior distribution of $\theta \mid x$ is *Dirichlet* $(\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_7 + x_7)$.
We begin Gibbs sampling by assigning to $\theta = (\theta_1, \theta_2, \ldots, \theta_7)$, the prior probabilities obtained from Table 1. A random sample is drawn from *Multinomial* $(N; \theta_1, \theta_2, \ldots, \theta_7)$, with $N = 100$. These values are taken as the first parameter values of beta distributions for the initial values of the iterations. At the end of the iterations, we have a sample of size m from the posterior distribution,

arrayed as $\begin{pmatrix} \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_7^{(1)} \\ \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_7^{(2)} \\ \dots & \dots & \dots & \dots \\ \theta_1^{(m)} & \theta_2^{(m)} & \dots & \theta_7^{(m)} \end{pmatrix}$.

Each row is a sample from the joint posterior distribution and columns are samples from the marginal distributions. First column gives samples from $g(\theta_1 \mid x)$ and second column gives samples from $g(\theta_2 \mid x)$ and so on. Now, discarding the initial values and averaging over the sample size, we get the improved estimates of the parameters. The R program is being used for the estimation.

2.1 Output Analysis

Table 2 depicts the output analysis of the study for the seven groups for four runs of the program.

Age (years)	$\widehat{\theta}$	Run1	Run2	Run3	Run4
20 - 29	$\widehat{\theta_1}$	0.01007200	0.01033389	0.01965272	0.01007097
30 - 39	$\widehat{\theta_2}$	0.14942026	0.13028229	0.09966381	0.09102453
40 - 49	$\widehat{\theta_3}$	0.25005136	0.24093310	0.27035829	0.26979353
50 - 59	$\widehat{\theta_4}$	0.15055422	0.18986951	0.19958647	0.22040184
60 - 69	$\widehat{\theta_5}$	0.14980235	0.11028538	0.08959871	0.12051873
70 - 79	$\widehat{\theta_6}$	0.25036396	0.24954914	0.26031505	0.23963357
80 +	$\widehat{\theta_7}$	0.03999377	0.07052424	0.06031410	0.04930833
Total	-	1.00025792	1.001778	0.9994892	1.000751

Table 2: *Proportions*

Invoking Table 2, the value of $\widehat{\theta_3}$ is larger, compared to the values of all other parameters, leading to the conclusion that, the age group 40^+ is the highest probable age at onset of breast cancer; the next highest value of proportions is $\widehat{\theta_6}$ corresponding to age group 70^+.

As is evident from Table 3, the extremely small variances of the iterated values illustrate the suitability of Gibbs sampling in the situation. Also, as we discard the initial values of iteration, we may use improper priors like $Uniform(0, 1)$ to get the initial values, and the output of various runs is given in Table 4.

Variances	Run1	Run2	Run3	Run4
θ_1	0.000106198	0.0001073495	0.0001842645	0.0001019053
θ_2	0.0012585498	0.001106535	0.0008381735	0.0008451778
θ_3	0.0019129253	0.0017651658	0.0019923381	0.0019049895
θ_4	0.0012695066	0.0015165971	0.0015674542	0.0017302377
θ_5	0.0012813251	0.0009974279	0.0008056178	0.001045555
θ_6	0.0018793944	0.0018086309	0.0019176001	0.0018403571
θ_7	0.0003776579	0.0006548418	0.0005499078	0.0004634782

Table 3: $Variances$

$\widehat{\theta}$	Run1	Run2	Run3	Run4
$\widehat{\theta_1}$	0.01009671	0.01981130	0.01014169	0.0099534367
$\widehat{\theta_2}$	0.12982770	0.14016432	0.07934393	0.159175893
$\widehat{\theta_3}$	0.26999674	0.27926111	0.25974321	0.229998541
$\widehat{\theta_4}$	0.12967862	0.18090535	0.22011336	0.188592595
$\widehat{\theta_5}$	0.11975033	0.08004237	0.11018176	0.099838067
$\widehat{\theta_6}$	0.26946581	0.2497542	0.27010255	0.250795979
$\widehat{\theta_7}$	0.06970986	0.04947005	0.04989858	0.0060248363
Total	0.99862209	0.9994087	0.9995251	0.998603

Table 4: $Proportionsusinginitialvaluesfrom Uniform$

3 Conclusion

When we compare the pattern in which the proportion estimates appear, with prior proportion values, the Multinomial - Dirichlet conjugate seems a good model in modeling the relationship between the age at onset and proportion of women affected of breast cancer.

References

[1] Berger J O., Statistical Decision Theory, Springer Verlag,N. York (1980).

[2] Bishop D.T., Cannon - Albright L., McLellan T., Gardner E. J., Skolnick M. H., Segregation and Linkage analysis of nine Utah breast cancer pedigrees ,Genet. Epidemiol.,5(1988),151 - 169.

[3] Carlin B P., Louis T A.,Bayes and Empirical Bayes methods for data analysis, Chapman and Hall , CRC Press, Boca Raton, (2000).

[4] Chib S., Markov Chain Monte Carlo Methods : Computation and Inference, Handbook of Econometrics, 5 (eds. J J Heckerman and E Leamer), Elsevier, Amsterdam, (2001),3569 - 3649.

[5] Claus E B., Risch N., Thompson W D., Age of onset as an indicator of familial risk of breast cancer, Am. J. Epidemiol.,131(1990a),961 - 972.

[6] Claus E B., Risch N., Thompson W D., Using age of onset to distinguish between sub forms of breast cancer , Ann. Hum. Genet.,54(1990b),169 - 177.

[7] Claus E B., Risch N., Thompson W D.,Genetic analysis of breast cancer in the cancer and steroid hormone study , Am. J. Hum. Genet.,48(1991),232 - 242.

[8] Gelman A J., Carlin B., Stern H., Rubin D.,Bayesian Data Analysis, Chapman and Hall, London, 2^{nd} ed, (2004).

[9] Geweke J., Using simulation methods for Bayesian econometric models : Inference, development, and communication , Econometric Reviews,18,(1999),1 - 126.

[10] Geweke J.,Contemporary Bayesian Econometrics and Statistics, Wiley, NJ,(2005).

[11] Go RCP, King M C., Bailey - Wilson J, Elston R C., Lynch H T., Genetic epidemiology of breast cancer and associated cancers in high - risk families . I. Segregation analysis, J. Natl. Cancer Institute., 71(1983), 455 - 461.

[12] Golstein A M., Haile RWC., Hodge S E., Paganini - Hill A., Spence M A.,Possible heterogeneity in the segregation pattern of breast cancer in families with bilateral breast cancer, Genet. Epidemiol., 5(1988), 121 - 133.

[13] Golstein A M., Haile RWC., Marazita M L., Paganini - Hill A., A genetic epidemiologic investigation of breast cancer in families with bilateral breast cancer. I. Segregation analysis, J. Natl. Cancer Institute.,78(1987),911 -918.

[14] Greenberg E.,Introduction to Bayesian Econometrics,Cambridge University Press,(2003).

[15] Heckerman D., Geiger D., Chickering D M., Learning Bayesian Networks : The combination of knowledge and statistical data, Machine Learning(1995).

Received: April, 2010