Aligarh Journal of Statistics
Vol. 34 (2014), 105-119

AN EMPIRICAL BAYES STUDY TO EXAMINE THE INTERACTION BETWEEN GENETIC SUSCEPTIBILITY AND ENVIRONMENTAL EXPOSURE

Akanksha Gupta and S. K. Upadhyay

ABSTRACT

Genes and environmental components lead to the development of several complex diseases and the analyses of these factors play a major role in the prevention and/or control of the diseases in human population. The analysis is, however, not easy because of the fact that genes and environmental factors interplay while causing diseases. If this interaction is ignored at the time of estimation of the contributions of these factors, it may provide incorrect estimates of the proportions of the disease caused by these factors. The case-control study, where sampling is conditional on the presence or absence of the disease, is a powerful epidemiological tool for studying such problems and the Bayesian framework offers increased level of flexibility for the possible modelling. A well known and user-friendly approach to analyze such data is multinomial-Dirichlet modelling assumption. The present paper analyzes a retrospective data on ovarian cancer based on the assumption of multinomial-Dirichlet model. Empirical Bayes method is used to select the prior hyperparameters. Results are found to be logical and appealing.

1. INTRODUCTION

Medical data may be made available in a variety of ways. One such possibility includes data in the form of counts, say, for example, cases and controls where cases are diseased individuals and controls are those who are disease free at the time of taking observations but exposed to the risk of the same. The main objective in any such study is to examine the causes for the development of a disease so that the appropriate remedial measures and/or the treatments can be suggested accordingly.

Case-control studies are generally conducted either retrospectively or prospectively (see, for example, Prentice and Pyke (1979), Abramson and Abramson (2001), Makkar (2009), etc.). In both the cases, the objective generally lies in drawing inferences about the odds ratio, a ratio that measures the odds of exposure for cases against controls. In fact, odds ratio is a relative measure of risk, which attempts to convey how much likely it is that someone exposed to the factor under study will develop the outcome as compared to someone who is not exposed (see, for example, Breslow (1996)).

Odds ratio when calculated retrospectively appears to be more appropriate as people who have developed a disease may be more likely to remember the

causes than those without the disease although this may sometimes introduce a recall bias, a tendency of individuals to report the events in a manner that is different among the two groups. Thus it is more pertinent to devise means for reducing recall bias for the successful implementation of case-control study and this can be done by using some biological and non-biological markers to assess the amount of exposure in both the categories. Schlesselman (1982) and Rothman (1986), etc. are a few important references that provide the relevant details on these aspects.

In genetic epidemiology, these markers are naturally inherent in the process in the form of genetic susceptibility and environmental exposure components. It is generally assumed that these factors are independent of each other at an individual level but these may be associated at the population level (see, for example, Modan *et al.* (2001)) because of their dependence on other factors such as age, ethnicity, past history of disease, etc. A review of the literature suggests that both prospective and retrospective studies are carried out with several merits or demerits of each such analysis (see, for example, Chatterjee and Carroll (2005), Mukherjee and Chatterjee (2008), etc.).

Important early reference on covariate information on genetic epidemiology may include Cornfield (1956) where the author showed that the prospective odds ratio of a disease given a covariate is related to the retrospective odds ratio of a covariate given a disease in some or other way and, therefore, the former can be estimated from the latter with appropriately chosen case-control design. Among some important classical references, Modan *et al.* (2001) proposed a simple case-control analysis for drawing inferences on odds ratio in the study of ovarian cancer patients. Lee *et al.* (2010) is another important recent classical reference where the authors provided an easy-to-implement approach for analyzing case-control and case-only study designs under the assumption of gene-exposure independence and Hardy-Weinberg equilibrium assumption.

On Bayesian front the work appears to be relatively meager although the new researches are going on very rapidly. Among the earlier important Bayesian references, Ashby *et al.* (1993) considered data on cases and controls and assumed beta-binomial modelling assumption to draw the desired inferences on odds ratio. Arora *et al.* (2011) provided a multivariate extension of the work done by Ashby *et al.* (1993). They used multinomial-Dirichlet modelling assumption and studied instead the effect of covariates in the bile acids of gallbladder diseases. An important problem with the Bayesian analysis given in the above references is the exact assessment of prior hyperparameters. The authors mostly used some logically adopted *ad hoc* mechanisms based on sensitivity analysis to recommend for the appropriate choices of prior hyperparameters.

Our proposed plan attempts to use multinomial-Dirichlet modelling assumption and provides an empirical Bayes (EB) analysis with a view to obtain the complete inferences on relevant odds ratios and the gene-environment

interaction parameter. EB is an approach to inference in which the observations are used to specify the exact prior hyperparameters usually via the marginal distribution. Once the prior is specified, the inference proceeds in a standard Bayesian framework. This approach cannot be considered strictly in accordance with the Bayesian logic as the prior specification is usually done in the spirit of classical paradigm. We, however, advocate its use simply because it is easy and offers at least a systematic way of assessing the prior hyperparameters.

The early references for the EB method include Robbins (1956), Casella (1985) among others. A detailed accountability of the method can be had from Carlin and Louis (2000), an important text on both Bayes and EB methods. The literature also includes references on the use of EB methods in the study of gene-environment problems. Chatterjee and Carroll (2005) is perhaps the first reference where the authors used a semi-parametric approach and worked out for maximum likelihood estimates in a case-control design involving both genetic and environmental components but treating the two to be independent. The approach is semi-parametric in the sense that it considers the distribution of environmental factors to be completely non-parametric. The method given by Chatterjee and Carroll (2005) is definitely advantageous but complicated, in general, to implement. Xu *et al.* (2013) is another reference where author worked parallel to Chatterjee and Carroll (2005). Mukherjee and Chatterjee (2008) extended the work of Chatterjee and Carroll (2005) by allowing dependence structure between genetic and environmental components with a trade off between bias and efficiency. The authors considered a weighted estimator of interaction parameter which was motivated by the expression for the posterior mean obtained in a conjugate analysis under a normal-normal modelling assumption.

The plan of the paper is as follows. The next section provides the necessary modelling formulation and other related inferential details for the proposed Bayesian implementation based on multinomial-Dirichlet modelling assumption. A separate subsection details the selection of relevant prior hyperparameters using EB method. The section also provides a separate subsection on a simple case-control design involving $2\times 2\times (l+1)$ setup that fits the Dirichlet-multinomial formulation given earlier in the Section. This subsection finally mentions a few important characteristics relevant to $2\times 2\times (l+1)$ setup that may be required for the latter inferences. Section 3 provides the description of a data set based on case-control study of ovarian cancer patients in Israel. This data set reports observations on genetic and environmental components besides a few usual characteristics of the patients. Due to confidentiality, we however considered a redesigned version of the data set initially reported by Modan *et al.* (2001) (see, for example, Chatterjee and Carroll (2005)). The Section also provides the numerical illustration based on the discussion given in Section 2. Finally, a brief conclusion is given at the end.

2. MODELLING FORMULATION AND THE RELEVANT INFERENTIAL DETAILS

To begin with let us consider to obtain inferences from data $\mathbf{x} = (x_1,....,x_k)$ categorized in k categories with unknown probability for the category i as $p_i, i = 1,...,k$. If $\Sigma x_i = n$, this can be represented by a multinomial probability law $x \sim$ multinomial (n, p), $\mathrm{p} = (p_1,..., p_k)$, with probability function given by,

$$f(\mathrm{x} \mid \mathrm{p}) \propto p_1^{x_1}.p_2^{x_2}...p_k^{x_k}, p_i \geq 0, i = 1,...,k. \tag{1}$$

It is well known that for multinomial parameters, a conjugate prior is Dirichlet distribution with probability density function given by,

$$g(\mathbf{p} \mid \boldsymbol{\lambda}) \propto \prod_{i=1}^{k} p_i^{\lambda_i - 1}, p_i \geq 0, \Sigma_{i=1}^{k} p_i = 1, \tag{2}$$

where $\lambda_1,...,\lambda_k$ are the positive parameters known as the strength of the distribution. The constant of proportionality for this model can be written as $1/B(\boldsymbol{\lambda})$ where

$$B(\boldsymbol{\lambda}) = \frac{\left(\prod_{i=1}^{k} \Gamma(\lambda_i)\right)}{\Gamma\left(\Sigma_{i=1}^{k} \lambda_i\right)}.$$

Dirichlet prior offers a conjugate family for multinomial likelihood and, as such, the family is quite flexible and rich and results in computationally easy inferences. However, the most crucial thing with the Dirichlet prior is the choice of its hyperparameters because different choices may result in different shapes of the model and hence different *a priori* information. This may, in turn, lead to different inferential developments. A systematic discussion on the choice of hyperparameters is beyond the scope of this paper. The interested readers may refer to Arora *et al.* (2011) for a simple strategy based on systematic choices (see also Gupta and Upadhyay (2013)). A strategy based on subjective elicitation of prior hyperparameters using quantile estimates with the help of past data can be had from Gupta and Upadhyay (2013). We shall not go into the details of various strategies rather focus on the present problem using a simple EB formulation.

Combining (1) with (2) via Bayes' theorem yields the posterior distribution that can be specified up to proportionality as

$$h(\mathbf{p} \mid \mathbf{x}, \boldsymbol{\lambda}) \propto \prod_{i=1}^{k} p_i^{x_i + \lambda_i - 1}, p_i \geq 0, \Sigma_{i=1}^{k} p_i = 1. \tag{3}$$

It can be seen that (3) again turns out to be the Dirichlet distribution with updated parameters $(x_1 + \lambda_1, ..., x_k + \lambda_k)$. The advantage with the Dirichlet model is that all the marginal posteriors can be reduced to beta distribution giving routine implementation of various inferential developments at least for some standard loss functions.

2.1 Empirical Bayes Technique for Selection of Prior Hyperparameters

The EB approach traditionally uses the prior density that maximizes the marginal probability of the observed data, integrating out with respect to the prior distribution of the parameters. For prior hyperparameter specification, one can, however, proceed parallel to the classical spirit of maximizing the likelihood function but with reference to the prior distribution. Thus, given a set of observed multinomial data $D = (x_1, ..., x_k)$, the parameters of the multinomial distribution can be easily obtained as the estimates of probabilities of cell counts $(\hat{p}_1, ..., \hat{p}_k)$. Now the parameters of the Dirichlet distribution, which is the prior for multinomial parameters, can be obtained by maximizing the logarithm of $g(\lambda)$ where $g(\lambda)$ is the form (2) obtained after replacing p_i with $\hat{p}_i$, $i = 1, 2, ..., k$. The logarithm of $g(\lambda)$ can be given as

$$G(\lambda) = \log g(\lambda)$$

$$= \log \frac{\Gamma(\sum_{i=1}^{k} \lambda_i)}{\prod_{i=1}^{k} \Gamma(\lambda_i)} \prod_{i=1}^{k} \hat{p}_i^{\lambda_i - 1}$$

$$= \log \Gamma\left(\sum_i \lambda_i\right) - \sum_i \log \Gamma(\lambda_k) + \sum_i (\lambda_i - 1) \log \hat{p}_i. \quad (4)$$

There are a number of methods for numerically maximizing this objective function $G(\lambda)$ as there is no closed form solution for the same. A detailed survey for various methods can be found in Lwin and Maritz (1989), Minka (2000), etc. Moreover, it was noted that the Dirichlet distribution is a special case of a larger class of distribution called the exponential family and, therefore, the log-likelihood function of data drawn from this distribution is convex in λ. This, in turn, guarantees a unique optimum (see Minka (2000)). The component wise gradient of G can be given as

$$(\nabla G)_k = \frac{\partial G}{\partial \lambda_k} = \xi(\sum_i \lambda_i) - \xi(\lambda_k) + \log \hat{p}_k, \quad (5)$$

where $\xi(.) = d \log \Gamma(.) / d(.)$ is the *digamma* function.

Of the various methods discussed in the literature, perhaps the one given by Minka (2000) is straightforward. The author provides a convergent fixed point

iteration technique for estimating the parameters. The idea behind this is to guess an initial value of $\boldsymbol{\lambda}$, find a function that bounds $G(\lambda)$ from below which is tight at $\boldsymbol{\lambda}$, and then to optimize this function to arrive at a new guess at $\boldsymbol{\lambda}$.

There are many inequalities associated to the ratio $\frac{\Gamma(\lambda_{t+1})}{\Gamma(\lambda_t)}$, where $\lambda_{t+1} \geq \lambda_t$ and t denotes the iteration number, which have been extensively studied by many mathematicians (see, for example, Guo and Qi (1976), Dragomir *et al.* (1999), etc.). A commonly cited one is,

$$\Gamma(\lambda_{t+1}) \geq \Gamma(\lambda_t)\exp((\lambda_{t+1}-\lambda_t)\xi(\lambda_t)),$$

which leads to a lower bound on the log likelihood, $G(\lambda)$, as

$$G(\lambda) \geq \left(\sum_i \lambda_i\right)\xi\left(\sum_i (\lambda_i)_t\right) - \sum_i \log\Gamma(\lambda_i) + \sum_i \lambda_i \log \hat{p}_i + C,$$

where C is a constant with respect to $\boldsymbol{\lambda}$. Now this expression can be maximized by setting the gradient (5) to zero and solving for $\boldsymbol{\lambda}$. The updating step is given by,

$$(\lambda_k)_{t+1} = \xi^{-1}\left(\xi\left(\sum_i (\lambda_i)_t\right) + \log \hat{p}_k\right),$$

where the digamma function ξ can be inverted efficiently by using a Newton-Raphson updating procedure to solve $\xi(.) = y$.

2.2 A Simple Case-Control Structure and A Few Associated Characteristics

To begin with let us consider a structure involving n (= n_0 + n_1) individuals as reported in Table 1 in the form of counts. These counts are represented as r_{DGE}, where D and G are binary variables taking values either 0 or 1 and E is a polytomous variable taking values $0,1,...,l$ for each combination of D and G.

Table 1: Classification of a case-control structure with respect to disease status, gene status and environmental exposure

	$G=0$				$G=1$				Total
	$E=0$	$E=1$	…	$E=l$	$E=0$	$E=1$	…	$E=l$	
$D=0$	r_{000}	r_{001}	…	r_{00l}	r_{010}	r_{011}	...	r_{01l}	n_0
$D=1$	r_{100}	r_{101}	...	r_{10l}	r_{110}	r_{111}	...	r_{11l}	n_1

Obviously, the structure can be very well represented by a multinomial distribution with $\mathbf{r_0} \sim$ multinomial $(n_0, \mathbf{p_0})$ where $\mathbf{r_0} = (r_{000},...,r_{00l}, r_{010},...,r_{01l})$ and $\mathbf{p_0} = (p_{000},...,p_{00l}, p_{010},...,p_{01l})$. Similarly, $\mathbf{r_1} \sim$ multinomial $(n_1, \mathbf{p_1})$ where $\mathbf{r_1} = (r_{100},...,r_{10l}, r_{110},...,r_{11l})$ and $\mathbf{p_1} = (p_{100},...,p_{10l}, p_{110},...,p_{11l})$. The components of $\mathbf{p_0}$ and $\mathbf{p_1}$ are the corresponding cell probabilities, that is,

$$p_{0GE} = r_{0GE} / n_0, \text{ for } G = 0,1; E = 0,1,...,l,$$

$$p_{1GE} = r_{1GE} / n_1, \text{ for } G = 0,1; E = 0,1,...,l,$$

where $\Sigma r_{0GE} = n_0$ and $\Sigma r_{1GE} = n_1$. We consider Dirichlet priors $g(\mathbf{p_0} \mid \mathbf{a})$ and $g(\mathbf{p_1} \mid \mathbf{b})$ for the corresponding cell probabilities $\mathbf{p_0} = (p_{000},...,p_{00l}, p_{010},...,p_{01l})$ and $\mathbf{p_1} = (p_{100},...,p_{10l}, p_{110},...,p_{11l})$, respectively, where $\mathbf{a} = (a_0, a_1,...,a_{2l+1})$ and $\mathbf{b} = (b_0, b_1,...,b_{2l+1})$ are the hyperparameters. Combining likelihoods and priors via Bayes theorem, we get the Dirichlet posteriors $h(\mathbf{p_0} \mid \mathbf{r_0}, \mathbf{a})$ and $h(\mathbf{p_1} \mid \mathbf{r_1}, \mathbf{b})$ with updated parameters $(\mathbf{r_0} + \mathbf{a})$ and $(\mathbf{r_1} + \mathbf{b})$, respectively, and because of being available in closed forms, these can be easily generated by a number of techniques (see, for example, Devroye (1986)). An important and easy procedure for generating from Dirichlet distribution can be managed through generation from gamma variates.

Let $OR_{10_i} = p_{000}p_{10i} / p_{00i}p_{100}$ denotes the odds ratio associated with $E = i$ for non-susceptible subjects $(G = 0)$, $OR_{01} = p_{000}p_{110} / p_{010}p_{100}$ denotes the odds ratio associated with G for unexposed individuals $(E = 0)$ and the odds ratio associated with $G = 1$ and $E = i$ compared to $G = 0$ and $E = 0$ are denoted by $OR_{11_i} = p_{000}p_{11i} / p_{01i}p_{100}$, $i = 1,...,l$. The measure of $G - E$ association in the control population at $i - th$ level of E is, therefore, given by,

$$\theta_{GE_i} = log(p_{000}p_{01i}) / (p_{00i}p_{010}), i = 1,...,l. \tag{6}$$

An easy procedure for getting the sample based estimates of θ_{GE_i} and other log odds ratios is straightforward. We simply need to simulate the values of $\mathbf{p_0}$ and $\mathbf{p_1}$ from the posterior distributions and after substituting the generated v and $\mathbf{p_1}$ in the concerned relationships, we may obtain the corresponding samples from the posterior distributions of the same. These samples may be used to estimate the entire posterior distribution of the corresponding variate, say, using kernel density estimate or histogram, etc. One may also draw other desired features of interest in a routine manner. Suppose, for example, one is interested in the point estimate of θ_{GE_i} and finds the estimated value close to zero. This means that G and $E = i$ are not dependent on each other in the control population. Contrary to

that if the estimated θ_{GE_i} is non-zero, one may go a step ahead and calculate the multiplicative interaction parameter between G and $E = i$ (see also Mukherjee and Chatterjee (2008)) given by,

$$\psi_i = (p_{00i}p_{010}p_{100}p_{11i}) / (p_{000}p_{01i}p_{10i}p_{110}), i = 1,...,l. \quad (7)$$

(7) provides a measure of association between the gene and environmental component $E = i, i = 1,...,l,$ and this may be the parameter where the interest often centers. Moreover, the measures of association and interaction as given in (6)-(7) are calculated fixing the base environmental exposure $E = 0$ although such measures can be obtained among any pairs of environmental exposure $E = i$ and $E = j$, $i \neq j \neq 0,$ provided the interest centres among different pairs and the cell frequencies are large enough to support such evaluations.

A word of remark: the measures given in (6) - (7) at each level of E may be mathematically correct but logically not always appealing. One may not be interested in evaluating the $G-E$ association or the multiplicative interaction parameter at each level of E rather may prefer to obtain the overall $G-E$ association or the interaction parameter. These may be obtained by combining all the categories of exposed individuals ($E = i, i = 1,...,l$) in to a single exposed category and then defining the corresponding parameters based on a $2\times2\times2$ setup. This situation is a particular case of $2\times2\times(l+1)$ design discussed above with $l = 1$ and results when the environmental component is also treated as binary. Obviously, the odds ratios can be redefined as $OR_{10} = p_{000}p_{101} / p_{001}p_{100}$, $OR_{01} = p_{000}p_{110} / p_{010}p_{100}$, $\theta_{GE} = log(p_{000}p_{011}) / (p_{001}p_{010})$ and the multiplicative interaction parameter as $\psi = (p_{001}p_{010}p_{100}p_{111}) / (p_{000}p_{011}p_{101}p_{110})$ for $2\times2\times2$ setup (see also Mukherjee *et al.* (2010)).

3. DATA DESCRIPTION AND NUMERICAL ILLUSTRATION

We consider a partially real and partially simulated data set related to ovarian cancer study among Jewish women in Israel which was taken from the web page of Chatterjee and Carroll (2005). The necessary details can be had from http://dceg.cancer.gov/people/ Chatterjee Nilanjan.html under the software link (see also Gupta and Upadhyay (2013)). This data set consists of real data values on disease status, D, and non-genetic co-factors, Y. For reasons of privacy, however, the real genetic data are not made available publicly. Instead, the data consist of simulated genetic data, G, generated using the conditional distribution of $[G \mid D, Y]$ as specified by the parameter estimates obtained from the real data (see Chatterjee and Carroll (2005) for details).

In spite of the fact that multiparity and use of oral contraceptive (OC) reduce the risk of ovarian cancer in women (see, for example, Modan *et al.* (2001)), we propose to study the affect of these factors on women with BRCA 1 and/or BRCA 2 mutation as well. The data set contains 747 controls and 832 cases of

women who underwent mutation analysis. The data set, referred to as the case-control data, is given in Table 2 for a ready reference.

For OC use, $E=0$ corresponds to those subjects who never used OC. $E=i$ corresponds to those who used OC up to 3 years, from 3 years to 6 years and for more than 6 years depending on whether $i=1,2,$ and 3, respectively. Similarly, for parity, $E=0$ corresponds to women who have no children, $E=i$ corresponds to 1 - 3 children, 3 - 6 children, and more than 6 children accordingly as $i=1,2,$ and 3, respectively. Obviously, the data set given in Table 2 represents a $2\times2\times4$ case-control design and our primary focus includes the analysis of the same data set to get the desired inferences.

Table 2: Classification of case-control data with respect to disease status, genetic susceptibility and environmental exposure

	OC Use								
	$G=0$				$G=1$				Total
	$E=0$	$E=1$	$E=2$	$E=3$	$E=0$	$E=1$	$E=2$	$E=3$	
$D=0$	577	86	32	40	9	1	1	1	747
$D=1$	494	7	15	16	184	34	7	15	832
	Parity								
$D=0$	42	506	155	32	1	8	2	1	747
$D=1$	68	373	116	35	20	188	30	2	832

The data set given in Table 2 can be easily converted in the form of a $2\times2\times2$ design by combining cells with non-zero E in to a single cell, say, $E=1$. Therefore, in this case, when OC use is considered as the environmental exposure, there are 586 unexposed individuals among controls. That is, among 747 controls, 586 are those who are not using OC whereas 161 individuals are using the same. While among cases, 678 are unexposed to environment and 154 are exposed. Similarly, when parity is considered as the environmental exposure, only 43 individuals are unexposed among controls while 704 are exposed. Among cases the number of unexposed individuals is 88 while that of exposed is 744. It is to be noted that the $2\times2\times2$ design has an advantage in the sense that various cell frequencies become appreciably large and, as such, the conclusions can be more readily relied upon.

To begin with the analysis of the data given in Table 2, we first obtained the maximum likelihood estimates of Dirichlet hyperparameters for $2\times2\times4$ setup by the method given in subsection 2.1. These estimates are $\tilde{\mathbf{a}}=(139.69,13.45,5.02,8.75,2.67,0.77,0.51,0.52)$ and $\tilde{\mathbf{b}}=(99.84,11.88,4.36,4.02,38.21,6.44,1.93,4.91)$ when OC use is considered as the environmental exposure. Similarly, the maximum likelihood estimates

corresponding to parity as the environmental exposure are obtained as $\tilde{\mathbf{a}} = (10.78, 116.63, 30.01, 8.75, 0.26, 2.86, 1.29, 1.01)$ and $\tilde{\mathbf{b}} = (9.01, 96.34, 27.09, 5.43, 3.34, 19.27, 6.34, 4.79)$. Based on these estimates of hyperparameters, the estimated posterior means of various cell probabilities are evaluated and presented in Table 3. The estimates are based on 5000 posterior samples. The bracketed values in the table show the corresponding estimated posterior standard deviations.

Table 3: Estimated sample based posterior means and the corresponding standard deviations of different cell probabilities based on EB procedure

Cell probabilities	*OC* use	Parity	Cell probabilities	*OC* use	Parity
p_{000}	0.7812	0.0572	p_{100}	0.5928	0.0757
	(0.0171)	(0.0048)		(0.0172)	(0.0035)
p_{001}	0.1081	0.6786	p_{101}	0.0784	0.4649
	(0.0052)	(0.0151)		(0.0041)	(0.0144)
p_{002}	0.0402	0.2013	p_{102}	0.0191	0.1411
	(0.0046)	(0.0041)		(0.0033)	(0.0025)
p_{003}	0.0528	0.0442	p_{103}	0.0198	0.0439
	((0.0048)	(0.0046)		(0.0033)	(0.0037)
p_{010}	0.0126	0.0013	p_{110}	0.2214	0.0228
	(0.0032)	(0.0012)		(0.0034)	(0.0034)
p_{011}	0.0019	0.0117	p_{111}	0.0401	0.2045
	(0.0014)	(0.0031)		(0.0039)	(0.0029)
p_{012}	0.0016	0.0035	p_{112}	0.0088	0.0356
	(0.0013)	(0.0019)		(0.0025)	(0.0036)
p_{013}	0.0016	0.0022	p_{113}	0.0196	0.0115
	(0.0013)	(0.0015)		(0.0033)	(0.0028)

We also obtained the same results for $2\times2\times2$ setup. The maximum likelihood estimates corresponding to *OC* use are $\tilde{\mathbf{a}} = (131.68, 36.26, 2.54, 1.125)$ and $\tilde{\mathbf{b}} = (103.62, 24.51, 34.76, 8.72)$ and same for the parity are $\tilde{\mathbf{a}} = (10.11, 157.94, 0.59, 2.98)$ and $\tilde{\mathbf{b}} = (14.26, 107.12, 5.58, 44.64)$. The posterior probabilities are also obtained using these sets of hyperparameters but we are not showing the results due to the fact that they are not conveying any additional

messages. Instead we present the results for log odds ratios and interaction parameters, the quantities of interest to epidemiologists, corresponding to $2\times2\times2$ design. These quantities have already been defined in subsection 2.2. The estimates of log odds ratios and other interactive parameters for $2\times2\times2$ and $2\times2\times4$ setup using the EB estimates of various cell probabilities are shown in Table 4 and 5 respectively. The bracketed values once again represent corresponding standard deviations.

Table 4: Posterior estimates of log odds ratio and other association parameters for $2\times2\times2$ setup

Parameters	*OC* use	Parity
θ_{GE}	0.1543	-0.3576
	(0.6303)	(0.9599)
$\log(OR_{10})$	-0.2921	-0.7585
	(0.1350)	(0.1918)
$\log(OR_{01})$	3.1474	2.5699
	(0.3098)	(0.9572)
$\log(\psi)$	0.2136	0.7263
	(0.6506)	(0.9979)

It can be seen from the results that the value of θ_{GE} clearly conveys the message that the genetic and environmental components are associated with each other. Odds ratios for non-susceptible subjects, that is, when $G=0$, are negative which means that disease and *OC* use/parity are inversely proportional to each other. The risk of ovarian cancer decreases with longer duration of *OC* use and increasing parity. The positive values of odds ratios $\log(OR_{01})$ take us towards the conclusion that the genes BRCA1/2 increase the risk of ovarian cancer. Both the interaction parameters are again positive representing that the interaction of genes and environmental components have a positive role in the occurrence of ovarian cancer although the risk is reduced in comparison to the situation when the environmental components are absent.

Observing Table 5 we can easily conclude that $\theta_{GE_i}, i=1,...,3,$ is again showing association for both *OC* use and parity discarding the assumption of independence between genetic susceptibility and environmental exposures, an assumption that earlier authors used to take. $log(OR_{10_i}), i=1,...,3,$ is as usual coming out to be negative suggesting that for non-susceptible subjects both *OC* use and parity reduce the risk of ovarian cancer and this risk mostly decreases

for higher levels of OC use and/or parity. The values of $\log(OR_{01})$ again help us to reach the conclusion that the patients with BRCA1/2 mutation are having high risk of developing ovarian cancer in control population.

Table 5: Posterior estimates of log odds ratio and other association parameters for $2\times2\times4$ setup

Posterior estimates	OC use	Parity	Posterior estimates	OC use	Parity
θ_{GE_1}	-0.0624	0.0845	$\log(OR_{10_3})$	-0.7398	-0.4152
	(0.9158)	(1.0868)		(0.2689)	(0.2881)
θ_{GE_2}	0.4299	-0.0460	$\log(OR_{01})$	3.1768	2.9454
	(1.0113)	(1.1887)		(0.3147)	(1.0945)
θ_{GE_3}	0.3520	0.8655	$\log(\psi_1)$	0.3659	0.3247
	(0.9851)	(1.3718)		(0.9330)	(1.1298)
$\log(OR_{10_1})$	-0.0389	-0.6715	$\log(\psi_2)$	-0.2447	-0.1033
	(0.1532)	(0.1833)		(1.1088)	(1.2561)
$\log(OR_{10_2})$	-0.4801	-0.6518	$\log(\psi_3)$	0.6475	-1.4905
	(0.2888)	(0.2027)		(1.0516)	(1.4973)

The values of interaction parameters are conveying a very important message. The interaction between BRCA 1/2 and OC use definitely increases the risk of ovarian cancer, however, this risk is reduced as compared to the situation when BRCA1/2 was acting alone. The negative value of $\log(\psi_2)$ is somewhat odd perhaps due to the small cell frequencies in $2\times2\times4$ setup. Coming to the interaction between BRCA1/2 and parity we can see that the risk decreases as the parity increases.

4. CONCLUSION

The paper is a successful attempt to study the interaction between genetic susceptibility and environmental exposure components as important causes for the development of any disease based on the assumption of multinomial-Dirichlet modelling. Other parameters involved in the modelling process are also estimated and their estimated standard deviations are obtained. Throughout we have used sample based approaches because of their apparent advantages in

the sense that every desired inference can be routinely obtained. This is otherwise difficult especially when one is concerned with functions of the parameters like odds ratio, multiplicative interaction, etc.

The second important finding relates to assessing the values of hyperparameters of the Dirichlet prior. The methodology given in this paper is straightforward and use data itself to estimate the prior hyperparameters. This approach is certainly less troublesome and perhaps may be enjoyed by those applied statisticians who do not want to invite the complications in selecting the hyperparameters.

REFERENCES

Abramson, J. H. and Abramson, Z. H. (2001): *Making Sense of Data: A Self-Instruction Manual on the Interpretation of Epidemiological Data*. 3rd edn., Oxford University Press.

Arora, P., Upadhyay, S. K. and Shukla, V. K. (2011): A Bayesian case-control study to study the effect of covariates in gall bladder carcinoma. *STM Journals, Research & Reviews: A Journal of Medicine*, **1**(**1**), 1–24.

Ashby, D., Hutton, J. L. and McGee, M. A. (1993): Simple Bayesian analyses for case-control studies in cancer epidemiology. *The Statistician*, **42**, 385–397.

Breslow, N. E. (1996): The case-control study. *J. Amer. Statist. Assoc.*, **91**, 14–28.

Carlin, B. P. and Louis, T. A. (2000): *Bayes and Empirical Bayes Methods for Data Analysis (2nd ed.)*. Chapman & Hall: London.

Casella, G. (1985): An Introduction to Empirical Bayes Data Analysis. *J. Amer. Statist. Assoc.*, **39**(**2**), pp. 83–87.

Chatterjee, N. and Carroll, R. J. (2005): Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, **92**, 399–418.

Cornfield, J. (1956): A statistical problem arising from retrospective studies. Proceedings of the third Berkeley symposium on mathematical statistics. Berkeley, CA: Berkeley University Press, **4**, 135–48.

Devroye, L. (1986): *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

Dragomir, S. S., Agarwal, R. P. and Barnett, N. (2000): Inequalities for beta and gamma functions via some classical and new integral inequalities. *Journal of Inequalities and Applications*, vol. 5, pp. 103–165.

Guo, B. N. and Qi, F. (1976): Inequalities and monotonicity for the ratio of gamma functions. *Taiwanese Journal of Mathematics*, vol. **19**(**7**), 407–409.

Gupta, A. and Upadhyay, S.K. (2013): A Bayes Study to Examine the Interaction between Genetic Susceptibility and Environmental Exposure: A Study Based on Ovarian Cancer. *Communicated*.

Gupta, A. and Upadhyay, S.K. (2013): Subjective Elicitation of Dirichlet Hyperparameters using Past Data − A Study of Ovarian Cancer Patients. *Communicated*.

Lee, W. C., Wang, L. Y. and Cheng, K. F. (2010): An Easy-to-implement Approach for Analyzing Case-control and Case-only Studies Assuming Gene-environment Independence and Hardy-Weinberg Equilibrium. *Statistics in Medicine*, **29**, 2557–2567.

Lwin, T. and Maritz, J. S. (1989): *Empirical Bayes Methods*. London: Chapman and Hall.

Makkar, P. (2009): *Bayesian Solutions of Some Medical Data Problems*. Unpublished Ph. D. Thesis, Banaras Hindu University, India.

Minka, T. (2000): *Estimating a Dirichlet Distribution*. Technical report, MIT,

Modan, B., Hartge, P., Hirsh-Yechezkel, G., Chetrit, A., Lubin, F., Beller, U., Ben-Baruch, G., Fishman, A., Menczer, J., Struewing, J. P., Tucker, M. A. and Wacholder, S. (2001): Parity, oral contraceptives and the risk of ovarian cancer among carriers and non-carriers of a BRCA1 or BRCA2 mutation. *New England Jr. of Medicine*, **345**, 235–240.

Mukherjee, B. and Chatterjee, N. (2008): Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade off between bias and efficiency. *Biometrics*, **64,** 685–694.

Mukherjee, B., Ahn, J., Gruber, S. B., Ghosh, M. and Chatterjee, N. (2010): Case-control studies of gene-environment interaction: Bayesian design and analysis. *Biometrics*, **66**(**3**), 934–948.

Prentice, R. L. and Pyke, R. (1979): Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403–411.

Robbins, H. (1956): An Empirical Bayes Approach to Statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics*, **1**, 157–163.

Rothman, K. J. (1986). *Modern Epidemiology*. Little Brown and Company: Boston.

Schlesselman, J. J. (1982): *Case-Control studies, Design, Conduct, Analysis*. New York: Oxford University Press.

Xu, J., Zheng, G. and Yuan A. (2013): Case-control genome-wide joint association study using semiparametric empirical model and approximate Bayes factor. *J. Stat. Comput. Simul.*, **83**(**7**), 1191–1209.

Received : 21.11.2013
Revised : 21.04.2014

Akanksha Gupta and S. K. Upadhyay

Department of Statistics
DST Centre for Interdisciplinary
Mathematical Sciences
Banaras Hindu University
Varanasi, India
email: akankshagupta1606@gmail.com