Austrian Journal of Statistics
January 2019, Volume 48, 1–14.
http://www.ajs.or.at/
doi:10.17713/ajs.v48i3.814

Subjective Elicitation of Dirichlet Hyperparameters Using Past Data: A Study of Ovarian Cancer Patients

Akanksha Gupta
Banaras Hindu University, Varanasi

S. K. Upadhyay
Banaras Hindu University, Varanasi

Abstract

Elicitation of prior plays a very important role in Bayesian paradigm especially when dealing with rare disease problems in medical field. The reason being that we do not get enough data to draw valid inferences always. Since the subject of study is human population, one cannot do experiments with their health. The prior distribution supports the final results by some additional information gained from the experts. In any case if an appropriate expert is not available, we can use past data to get information about the prior and its hyperparameters. The present paper provides a technique of elicitation of prior hyperparameters based on a well known multinomial-Dirichlet model. Since the main focus is on medical data problems, the inferences on odds ratios and interaction parameters are also provided. Numerical illustration is based on a real dataset from Israel on patients having ovarian cancer. Although the details have been given in the context of ovarian cancer patients, the development in the paper is equally well applicable for any such disease.

Keywords: case-control study, Dirichlet prior, subjective elicitation, gene-environment interaction.

1. Introduction

Bayesian approaches to the analysis of epidemiological data represent a powerful tool for interpretation of study results and evaluation of hypotheses about exposure-disease relations. Definitely prior distributions have a major role in Bayesian inference and one has to be very careful while choosing them. A small error in the choice of prior may result into misleading inferences. There are various types of priors suggested in the literature, say, for example, Jefferys' prior, reference prior, conjugate prior, uniform prior, etc. In true sense, the spirit of Bayesian paradigm lies certainly in the subjective elicitation of prior that may be approached by the experimenter through her subjective assessment or expert opinion. The objective is to convert the assessment or the opinion into a probabilistic form. In most of the cases the expert is unaware of statistical terminologies. As a result, it becomes a difficult task to extract the information required to draw inferences on a particular hypothesis. Moreover, an elicitation is considered appropriate if the distribution that is derived accurately represents the expert's knowledge regardless of how good that knowledge is. The idea of subjective prior elicitation has been advocated by a number of researchers (see, for example, O'Hagan (2006)) but not given due emphasis by most of the applied Bayesian practitioners.

Elicitation or subjective elicitation generally involves two strategies. In one case it may involve elic-

iting the form of the prior distribution whereas in the other case it may be concerned with eliciting only the hyperparameters of an assumed form of prior. The situation may also be considered where one requires the elicitation of both prior distribution and its hyperparameters. This paper focusses on elicitation of hyperparameters assuming the multinomial-Dirichlet combination considered earlier by several authors. This may be equivalently considered as specifying the exact member of the Dirichlet family based on the expert opinion that may be appropriate for the considered situation. Since the prior distribution does play a crucial role in final inferences, a small ambiguity in expert's perception or a small misinterpretation of subjective opinion in converting into an appropriate *a priori* judgement may result into drastically poor inferences. Keeping this in mind, a number of strategies have been suggested in the literature for eliciting the hyperparameters of a chosen prior distribution, in general, and of a conjugate prior distribution, in particular. Staël von Holstein (1971) is perhaps the first reference where the author suggested a method for assessing the conjugate prior for a Bernoulli process. His technique for assessing a beta distribution made use of the estimates of median and first and third quartiles from the experts' subjective opinion. Kadane, Dickey, Winkler, Smith, and Peters (1980) presented a method for estimating conjugate priors for a linear regression model. Their method made use of the multivariate t predictive distribution, which involved the assessment of belief about measures of central tendency for the regression coefficients as well as an assessment of belief about variation or covariation and the appropriate value of a degrees-of-freedom parameter. Besides, we also have an important reference by Winkler, Smith, and Kulkarni (1978) where authors considered the elicitation of conjugate priors for the linear regression model by using predictive distributions.

Among other significant references, Chaloner and Duncan (1983) presented a method called predictive modal estimation in the context of assessing beta prior. In their method, the expert first provides an assessment of the mode of the distribution and then assesses the likelihood of other points along the distribution relative to the likelihood of the mode. Similarly, Garthwaite and Dickey (1988) examined the linear regression problem but introduced a technique that was based on the concept of points of constrained minimum variance. In this technique, certain values of the independent variables are given to the expert. The expert is then asked to select values for the remaining independent variables so that his or her uncertainty regarding the dependent variable is minimized. A few fractile assessments are also required from the expert in this approach to complete the elicitation of the subjective probability distributions.

Truly speaking, to determine the prior hyperparameters, expert judgement is usually sought on the quantities that may be easily assessed. The elicited quantities are then equated to their theoretical expressions in order to solve for the unknown prior hyperparameters. A few such quantities may include the central tendency measures such as the mean, median, mode, etc. Quantile estimates may also be sometimes used based on the expert judgement. The problem, however, encountered with most of these elicited characteristics is that they result in numerically challenging scenarios when equated with the corresponding theoretical expressions. Dorp and Mazzuchi (2003) is an important reference that makes use of the quantile estimates for specification of the parameters of the beta distribution and its multivariate analogue. They have given a detailed accountability of the inherent numerical intricacies and accordingly provided a few possible solutions based on certain iterative procedures.

This paper considers a problem from genetic epidemiology where it is supposed that disease in a human population occurs sometimes due to one's physical structure inherited by birth and sometimes due to outer environmental components that become part of individual's everyday life. Subtle differences in genetic factors also cause people to respond differently to the same environmental exposure. The complex interaction between an individual's genetic composition and environmental agents is the cause for almost all the diseases. Although the interaction of genes and environment are often discussed in the context of disease or negative traits, the impact of their interaction can be protective, neutral, or even harmful.

Mukherjee and Chatterjee (2008), Mukherjee, Ahn, Gruber, Ghosh, and Chatterjee (2010) and Gupta and Upadhyay (2014) are some recent references in this area where the authors have considered gene-environment association problems and obtained inferences on odds ratios and various interaction parameters. The current work provides a complete Bayes analysis of the same problem considering a multinomial-Dirichlet modelling assumption. It mainly focusses on the elicitation of Dirichlet prior

hyperparameters by the method given by Dorp and Mazzuchi (2003). As we do not have expert knowledge on the specific application discussed in this paper, we presume that this information is available from other sources. It is important to mention here that the objective of the present paper is to formalize the procedure presuming as if we have the expert opinion though in actual implementation we never got an expert who could be examined for her opinion. As such, we rely on an alternative strategy based on the past data or on a small subset of the given data to derive the expert opinion on the quartiles that are subsequently used in the elicitation of prior hyperparameters. There is no claim made here that our procedure is perfect. We believe, however, that our development can serve as the beginning of a new research where prior is given due consideration in the light of expert's opinion to get an appropriate conclusion.

The plan of the paper is as follows. Section 2 discusses the formulation of the model for the problem under consideration with a focus on our main objective, the subjective elicitation of prior. The complete method is described in subsection 2.1. Based on the method discussed in this subsection, the numerical illustration is provided in Section 3. Here we consider an ovarian cancer dataset on Israeli women and accordingly draw the relevant inferences. We also provide the same study on reduced dataset and the results are presented in subsection 3.1. Finally, a brief conclusion is given at the end.

2. Model formulation

At the outset let us consider a specific epidemiological problem that involves case-control scenario in the structure of $2 \times 2 \times (l+1)$. This structure is relevant where there are exactly three categorizations with each of the first two categorizations, F_1 and F_2, having two levels and the third categorization, F_3, having $(l+1)$ levels. We may refer, for instance, the two levels of F_2, that is, $F_2 = 0$ and $F_2 = 1$ as controls and cases, respectively. The complete structure consisting of $n(= n_0 + n_1)$ observations can be classified as in Table 1 in the form of different cell counts. We have used the notation r_{ijk} to represent the count of individual cell where both i and j correspond to F_1 and F_2, respectively, and k corresponds to the third categorization, F_3. Obviously, both i and j are binary variables taking values either 0 or 1 whereas the variable k is a polytomous variable that may take values $0, 1, ..., l$.

Table 1: Classification of a case-control scenario in $2 \times 2 \times (l+1)$ structure

	$F_1 = 0$				$F_1 = 1$				Total
	$F_3 = 0$	$F_3 = 1$	...	$F_3 = l$	$F_3 = 0$	$F_3 = 1$	...	$F_3 = l$	
$F_2 = 0$	r_{000}	r_{001}	...	r_{00l}	r_{100}	r_{101}	...	r_{10l}	n_0
$F_2 = 1$	r_{010}	r_{011}	...	r_{01l}	r_{110}	r_{111}	...	r_{11l}	n_1

The structure given in Table 1 can be very well represented by two multinomial distributions with $\mathbf{r}_0 \sim$ multinomial $(n_0, \mathbf{p}_0)$ and $\mathbf{r}_1 \sim$ multinomial $(n_1, \mathbf{p}_1)$ where $\mathbf{r}_0 = (r_{000}, ..., r_{00l}, r_{100}, ..., r_{10l})$, $\mathbf{p}_0 = (p_{000}, ..., p_{00l}, p_{100}, ..., p_{10l})$, $\mathbf{r}_1 = (r_{010}, ..., r_{01l}, r_{110}, ..., r_{11l})$ and $\mathbf{p}_1 = (p_{010}, ..., p_{01l}, p_{110}, ..., p_{11l})$. It may be noted that we have considered two multinomial distributions, one corresponding to controls ($F_2 = 0$) and the other corresponding to cases ($F_2 = 1$). Moreover, the components of $\mathbf{p}_0$ and $\mathbf{p}_1$ are the corresponding cell probabilities given by $p_{ijk} = r_{ijk}/n_j$ where $\Sigma r_{ijk} = n_j$, $i = 0, 1; j = 0, 1; k = 0, 1, ..., l$. The likelihood functions based on the above two configurations for controls and cases can be written as

$$L(\mathbf{p}_j) \propto {p_{0j0}}^{r_{0j0}} ... {p_{0jl}}^{r_{0jl}} {p_{1j0}}^{r_{1j0}} ... {p_{1jl}}^{r_{1jl}}, j = 0, 1. \tag{1}$$

Let us now consider separate Dirichlet priors for the cell probabilities $\mathbf{p}_j = (p_{0j0}, ..., p_{0jl}, p_{1j0}, ..., p_{1jl})$ for $j = 0, 1$ as

$$g_j(\mathbf{p}_j|\mathbf{a}_j) \propto {p_{0j0}}^{a_{0j0}-1} ... {p_{0jl}}^{a_{0jl}-1} {p_{1j0}}^{a_{1j0}-1} ... {p_{1jl}}^{a_{1jl}-1}, j = 0, 1, \tag{2}$$

where $\mathbf{a}_j = (a_{0j0}, ..., a_{0jl}, a_{1j0}..., a_{1jl}), j = 0, 1$, are the hyperparameters. The corresponding posteriors up to proportionality can be written as

$$P_j(\mathbf{p}_j|\mathbf{a}_j, \mathbf{r}_j) \propto p_{0j0}^{r_{0j0}+a_{0j0}-1} ... p_{0jl}^{r_{0jl}+a_{0jl}-1} p_{1j0}^{r_{1j0}+a_{1j0}-1} ... p_{1jl}^{r_{1jl}+a_{1jl}-1}, \tag{3}$$

where the posterior for the controls ($F_2 = 0$) corresponds to $j = 0$ and the posterior for the cases ($F_2 = 1$) corresponds to $j = 1$. Obviously, the two posteriors given in (3), when normalized, are again Dirichlet distributions with updated parameters $(\mathbf{r}_j + \mathbf{a}_j)$ where $\mathbf{r}_j$ and $\mathbf{a}_j$ for $j = 0, 1$ are already defined above.

In epidemiological studies, the objective generally lies in drawing inferences about the odds ratio, a ratio that measures the odds of exposure for cases against controls. These odds ratios can be calculated for any of the factors of interest, say, F_1, F_2, F_3 and also for the joint effect of these factors, say, for example, effect of F_2 and F_3 on different levels of F_1, etc. Let $OR_{F_{3_k}} = p_{000}p_{01k}/p_{00k}p_{010}$ denotes the odds ratio associated with the k^{th} level of F_3 for $F_1 = 0$ and $OR_{F_1} = p_{000}p_{110}/p_{010}p_{100}$ denotes the odds ratio associated with F_1 for $F_3 = 0$. There may be situations when any two of these three components might have a joint affect on the third one. To check whether such association exists, we define the measure of association between, say, F_1 and F_3 considering only its k^{th} level and keeping $F_2 = 0$. Thus

$$\theta_{F_1F_{3_k}} = log\left\{(p_{000}p_{01k})/(p_{00k}p_{010})\right\}, k = 1, ..., l, \tag{4}$$

where the subscript k with F_3 denotes its k^{th} level. If, however, this association comes out to be non-zero, one may go a step ahead for calculating the multiplicative interaction parameter between F_1 and F_3 at its k^{th} level (see also Mukherjee and Chatterjee (2008)). This interaction parameter can be given as

$$\psi_k = (p_{00k}p_{010}p_{100}p_{11k})/(p_{000}p_{01k}p_{10k}p_{110}), k = 1, ..., l. \tag{5}$$

2.1. Subjective elicitation of prior

To present the methodology in a simple way, we consider the same modelling formulation given earlier. We, however, use slightly different form in this subsection and the symbols used here may not be exactly related to what we have defined earlier. To begin with, let us consider $\mathbf{x} = (x_1, \ldots, x_\ell)$, $\mathbf{p} = (p_1, \ldots, p_\ell)$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_\ell)$ and define

- $\mathbf{x} \sim$ multinomial$(n, \mathbf{p}) \propto p_1^{x_1}p_2^{x_2}\cdots p_\ell^{x_\ell};\ \Sigma_{i=1}^{\ell}x_i = n,\ \Sigma_{i=1}^{\ell}p_i = 1,\ p_i \geq 0,\ \ i = 1, \ldots, \ell$
- $\mathbf{p} \sim$ Dirichlet$(\boldsymbol{\lambda}) \equiv g(\mathbf{p}|\boldsymbol{\lambda}) = \frac{\Gamma\left(\Sigma_{i=1}^{\ell}\lambda_i\right)}{\left(\prod_{i=1}^{\ell}\Gamma(\lambda_i)\right)}\prod_{i=1}^{\ell}p_i^{\lambda_i - 1}$
- $\mathbf{p}|\mathbf{x} \sim$ Dirichlet$(\mathbf{x} + \boldsymbol{\lambda})$

To describe briefly the elicitation procedure for the model under consideration, let us first reparameterize the Dirichlet $(\boldsymbol{\lambda})$ distribution by introducing new parameters $\beta = \sum_i \lambda_i$ and $\alpha_i = \lambda_i/\beta, i = 1, ..., \ell$, yielding the density,

$$g(\mathbf{p}|\boldsymbol{\alpha}, \beta) = \frac{\Gamma(\beta)}{\left(\prod_i \Gamma(\beta\alpha_i)\right)\Gamma\left(\beta\left(1 - \sum_i \alpha_i\right)\right)}\left(\prod_i p_i^{\beta\alpha_i - 1}\right)\left(1 - \sum_i p_i\right)^{\beta(1-\sum_i \alpha_i)-1}, \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_\ell)$, $p_i \geq 0$, $\sum_i p_i = 1$, $\alpha_i \geq 0$, $i = 1, 2, ..., \ell$, $\sum_i \alpha_i = 1$ and $\beta > 0$. Thus to obtain the estimates of $\lambda_i's$, we need to elicit both $\alpha_i's$ and β.

It is well known that beta distribution is a special case of Dirichlet distribution and all the marginal distributions of Dirichlet variables follow beta distributions. An important property of beta distributions is that for all $p \in [0, 1]$, $0 < m_1 < m_2 \Rightarrow g(p|m_1, n) > g(p|m_2, n)$ and $n_1 > n_2 > 0 \Rightarrow g(p|m, n_1) > g(p|m, n_2)$ where p denotes a univariate analogue of $\mathbf{p}$ and $g(p|m, n)$ denotes a beta distribution with parameters m and n given by

$$g(p|m, n) = \frac{\Gamma(m + n)}{\Gamma(m)\Gamma(n)}p^{m-1}(1 - p)^{n-1}, m, n > 0, 0 < p < 1. \tag{7}$$

If we re-parameterize the beta distribution in the same way as it was suggested for the Dirichlet distribution, the above property can be written as,

$$\alpha_2 > \alpha_1 > 0,\ \beta > 0 \Rightarrow Pr(P \leq p|\alpha_2,\ \beta) < Pr(P \leq p|\alpha_1,\ \beta), \tag{8}$$

where P is used to denote the random variable corresponding to p and $\alpha_{(.)}$ and β are the parameters of the re-parameterized beta distribution. Obviously, the marginals of Dirichlet distribution also follow this property. Finally, the quantile constraint concept given below in the form of a definition can be used as well (see, for example, Dorp and Mazzuchi (2003)).

Definition. Let $0 < p_q < 1$, $0 < q < 1$. A random variable P with support $[0, 1]$ satisfies quantile constraint (p_q, q) if and only if $Pr(P \leq p_q) = q$.

This property can be easily extended for the Dirichlet distribution at the level of each of its marginals. Thus to get the values of α_i, $i = 1, 2, ..., \ell$, and β, we need to solve $\boldsymbol{\alpha}$ and β for $\mathbf{p} \sim g(\mathbf{p}|\boldsymbol{\alpha}, \beta)$ given in (6) under the two quantile constraints (p^i_{qL}, q^i_L) and (p^i_{qU}, q^i_U) for any P_i, where $q^i_L < q^i_U$, and $(\ell - 1)$ single quantile constraint (p^j_q, q^j) for P_j, $j = 1, 2, ..., \ell$, $j \neq i$. As usual, $P_{(.)}$ is used to denote the random variable corresponding to the component $p_{(.)}$. Since this problem cannot be solved in a closed form, one has to resort to an appropriate numerical solution. A numerical solution is defined below (see also Dorp and Mazzuchi (2003)). A word of remark: although the method of Dorp and Mazzuchi (2003) is straightforward to deal with, it suffers from an important drawback in the sense that P_i's are not treated symmetrically. The method actually requires placing two quantile constraints on one of the probabilities and $(\ell - 1)$ single quantile constraint on others. Moreover, P_ℓ is treated quite differently than other probabilities (see, for example, Evans, Guttman, and Li (2017)). Alternative methods are, of course, suggested in the literature (see Zapata-Vázquez, O'Hagan, and Bastos (2014) and Evans *et al.* (2017)) to deal with the elicitation of Dirichlet parameters efficiently but we stick to the method of Dorp and Mazzuchi (2003) simply because of its inherent ease. Moreover, since our illustration in the next section considers a large data size, it is expected that a slight deviation in elicited prior will not affect our final inferences.

To clarify the details, let us suppose an initial interval (c_1, f_1) that contains β^* (an estimate of β) and let an initial value of β^* be $\beta^*_1 = \frac{c_1+f_1}{2}$. For this particular β^*_1, calculate α^*_{i1}, an initial value of α_i, by considering an interval $(d_1, e_1) = (0, 1)$, which is supposed to contain α^*_{i1} and take $\alpha^*_{i1} = \frac{d_1+e_1}{2}$. Next calculate $(q_U)_1 = Pr(P \leq p_{qU}|\alpha^*_{i1}, \beta^*_1)$, where p_{qU} is the upper quantile value suggested by the expert. If $(q_U)_1 \leq q_U$, where q_U is the area of upper quartile, then $(d_2, e_2) \equiv (d_1, \alpha^*_{i1})$, otherwise $(d_2, e_2) \equiv (\alpha^*_{i1}, e_1)$. This follows from the property of the beta distribution given by (8). Now take $\alpha^*_{i2} = \frac{d_2+e_2}{2}$ and again calculate $(q_U)_2 = Pr(P \leq p_{qU}|\alpha^*_{i2}, \beta^*_1)$ and repeat the same procedure described above until at the t^{th} stage $(q_U)_t \approx q_U$.

After this we shall go on for refining β^*_1. For this, we calculate $q_L = Pr(P \leq (p_{qL})_1|\alpha^*_{it}, \beta^*_1)$, where q_L is the area of lower quantile. If $(p_{qL})_1 < p_{qL}$, where p_{qL} is the lower quantile value given by the expert, then $(c_2, f_2) \equiv (\beta^*_1, f_1)$, otherwise $(c_2, f_2) \equiv (c_1, \beta^*_1)$ and $\beta^*_2 = \frac{c_2+f_2}{2}$. For this value β^*_2 of β again update the value of α_i by the same procedure. The value of β is updated until $(p_{qL})_t \approx p_{qL}$. In this way the final values of α_i and β are obtained as α^*_i and β^*, respectively. Using the final value of β^*, we calculate α^*_j, $j = 1, 2, ..., \ell$, $j \neq i$ through the same procedure by which α^*_i is calculated.

To determine the initial interval (c_1, f_1) containing β^*, first set $c_1 = 0$ as $\beta > 0$. To obtain the upper bound f_1 of this initial interval, set $\beta_{11} = 1$ which implies $f_1 = 2$. Now solve for α_{i1} satisfying $(q_U)_1 = Pr(P \leq p_{qU}|\alpha_{i1}, \beta_{11})$. Proceeding in a similar manner described above, we may go on for updating α_{i1}. Say, for α_{it} at t^{th} stage, we get $(q_U)_t \approx q_U$. We may then find out $(p_{qL})_{1,t}$ by solving $q_L = Pr(P \leq (p_{qL})_{1,t}|\alpha_{it}, \beta_{1,t})$. In case $(p_{qL})_{1,t} < p_{qL}$ then set $\beta_{1,t+1} = 2 * \beta_{1,t}$. We may repeat the above procedure for all t. Conversely, if $(p_{qL})_{1,t} > p_{qL}$, set $f_1 = \beta_{1,t}$. In this way the initial interval containing β^* may be determined.

3. Numerical illustration

To illustrate our procedure, we considered an ovarian cancer dataset based on the women of Israel. This is a partially real and a partially simulated dataset which has been analyzed earlier by a number of authors (see, for example, Modan, Hartge, Hirsh-Yechezkel, Chetrit, Lubin, Beller, Ben-Baruch, Fishman, Menczer, Struewing, Tucker, and Wacholder (2001), Chatterjee and Carroll (2005), Mukherjee and Chatterjee (2008), etc.). A recent analysis of this data has been done by Gupta and Upadhyay

(2014) but in an empirical Bayes way. The dataset is given in Table 2. The categories F_1, F_2 and F_3 are denoted by G, D and E representing genetic susceptibility (absent/ present), disease status (controls/ cases) and environmental exposure (absent/ present at level k, $k = 1, 2, 3$), respectively.

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	67	15	16	184	34	7	15	832
Parity									
	$G=0$				$G=1$				Total
	$E=0$	$E=1$	$E=2$	$E=3$	$E=0$	$E=1$	$E=2$	$E=3$	
$D=0$	42	506	155	32	1	8	2	1	747
$D=1$	68	373	116	35	20	188	30	2	832

The gene responsible for the occurrence of ovarian cancer is supposed to be $BRCA1$ and/or 2 mutation and the levels of environmental exposures, considered as oral contraceptive (OC) use and parity, are defined as follows:

- For OC use:
  $E = 0 \Rightarrow$ subjects who never used OC,
  $E = 1 \Rightarrow$ corresponds to those who used OC up to 3 years,
  $E = 2 \Rightarrow$ corresponds to those who used OC from 3 years to 6 years and
  $E = 3 \Rightarrow$ corresponds to those who used OC for more than 6 years.

- For parity:
  $E = 0 \Rightarrow$ corresponds to women who have no children,
  $E = 1 \Rightarrow$ corresponds to women having 1-3 children,
  $E = 2 \Rightarrow$ corresponds to women having 3-6 children and
  $E = 3 \Rightarrow$ corresponds to women having more than 6 children.

For convenience most of the previous authors have analyzed this problem considering that genes and environmental components are independent of each other. This dataset can be converted into a $2 \times 2 \times 2$ data by combining the non-zero cells of the category environmental exposure to $E = 1$. Such an attempt might be of general interest to those medical practitioners who simply want to study the impact of presence or absence of environmental components and are not interested to go into a detailed study at different levels.

Table 3: Classification of extracted data ($2 \times 2 \times 2$) of size 50 with respect to disease status, genetic susceptibility and environmental exposure

OC Use					
	$G=0$		$G=1$		Total
	$E=0$	$E=1$	$E=0$	$E=1$	
$D=0$	19	3	1	1	24
$D=1$	16	2	5	3	26
Parity					
	$G=0$		$G=1$		Total
	$E=0$	$E=1$	$E=0$	$E=1$	
$D=0$	2	20	1	1	24
$D=1$	2	15	2	7	26

We also considered a dataset of moderate sample size with $n = 50$ extracted from $2 \times 2 \times 2$ setup discussed above in such a way that each cell frequency remains at least equal to unity. The dataset is shown in Table 3 and it is taken exclusively for the purpose of comparison. It is to be noted that such datasets with small to very small sample sizes may not result in the desired inferences on

association and interaction parameters as the corresponding reported cell frequencies may not be the true representatives of the prevalence of gene and environmental components in the actual population.

In our analysis the primary objective is to estimate the odds ratio, measure of association between G and E and, in case this association is non-zero, the multiplicative interaction parameter between the latter two. It may be noted, however, that the $G-E$ association in the control population at k^{th} level of E (see (4)) and the multiplicative interaction parameter between G and $E=k$, $k=1,2,3$, (see (5)) based on the observations in Table 2 are not logically appealing as a number of cell frequencies are too small to draw any valid conclusion on the actual $G-E$ association or the multiplicative interaction parameter. We, therefore, propose to consider such measures for $2\times2\times2$ setup along with $2\times2\times4$ setup. In this case, we shall drop the subscript k from OR_{E_k}, θ_{GE_k}, and ψ_k keeping their interpretations the same.

Since we did not have any expert who can be contacted to get her subjective opinion, we divided the entire data into two parts, one having first 50 observations and other, the remaining 1529 observations. This division was done for both the $2\times2\times4$ and $2\times2\times2$ setups separately. The first part of the data was considered as past data and it was used for elicitation of hyperparameters whereas the second part of the data was used for the desired inferences. The two datasets are shown in Tables 4-5. The values in the parentheses correspond to $2\times2\times2$ setup.

Table 4: Past data with sample size 50 for elicitation of prior hyperparameters

OC Use									
	$G=0$				$G=1$				Total
	$E=0$	$E=1$	$E=2$	$E=3$	$E=0$	$E=1$	$E=2$	$E=3$	
$D=0$	18 (18)	3 (5)	1 –	1 –	1 (1)	0 (0)	0 –	0 –	24 (24)
$D=1$	15 (15)	2 (3)	1 –	1 –	5 (6)	1 (2)	0 –	1 –	26 (26)
Parity									
	$G=0$				$G=1$				Total
	$E=0$	$E=1$	$E=2$	$E=3$	$E=0$	$E=1$	$E=2$	$E=3$	
$D=0$	1 (2)	16 (22)	5 –	1 –	0 (0)	1 (0)	0 –	0 –	24 (24)
$D=1$	2 (2)	11 (16)	3 –	1 –	1 (1)	6 (7)	2 –	0 –	26 (26)

Table 5: Remaining data with sample size 1529 for the inferential developments

OC Use									
	$G=0$				$G=1$				Total
	$E=0$	$E=1$	$E=2$	$E=3$	$E=0$	$E=1$	$E=2$	$E=3$	
$D=0$	559 (559)	83 (153)	31 –	39 –	8 (8)	1 (3)	1 –	1 –	723 (723)
$D=1$	479 (479)	65 (95)	14 –	15 –	179 (178)	33 (54)	7 –	14 –	806 (806)
Parity									
	$G=0$				$G=1$				Total
	$E=0$	$E=1$	$E=2$	$E=3$	$E=0$	$E=1$	$E=2$	$E=3$	
$D=0$	41 (40)	490 (671)	150 –	31 –	1 (1)	7 (11)	2 –	1 –	723 (723)
$D=1$	66 (66)	362 (508)	113 –	34 –	19 (19)	182 (213)	28 –	2 –	806 (806)

In order to elicit the prior hyperparameters (see subsection 2.1), we first obtained the empirical Bayes (EB) estimates of Dirichlet hyperparameters $\mathbf{a_j}$, $j=0,1$, where $\mathbf{a_0}$ denotes the vector of hyperparameters corresponding to controls and $\mathbf{a_1}$ denotes the same for cases, based on the past data of size 50 (see also Gupta and Upadhyay (2014)). It is to be noted that some of the cell observations were zero in past data making the corresponding cell probabilities also zero. In order to find the EB estimates for these cases, we considered very small cell probabilities of the order 10^{-5} for these cells so that the

corresponding EB estimates can be worked out. Before we proceed further, let us describe briefly the EB procedure used in the present paper.

The EB approach traditionally uses a prior density with hyperparameters estimated on the basis of observed data usually by means of classical approach. If maximum likelihood (ML) estimates are easily obtainable, the same are often preferred. To provide a brief outline of used EB procedure, let us consider $\underline{\mathbf{x}} = (x_1, \ldots, x_\ell)$ as the set of observed multinomial data. The corresponding multinomial parameters $(p_1, \ldots, p_\ell)$ can be easily obtained as the estimates of different cell probabilities based on various cell counts. Let these estimates are denoted by $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_\ell)$ (see also Gupta and Upadhyay (2014)). Finally, the EB estimates of the parameters of the Dirichlet distribution, which is the prior for multinomial parameters, can be obtained by maximizing the corresponding log-likelihood function given by,

$$\begin{aligned} G(\boldsymbol{\lambda}) &= \log g(\hat{\mathbf{p}}|\boldsymbol{\lambda}) \\ &= \log \frac{\Gamma(\sum_{i=1}^{\ell} \lambda_i)}{\prod_{i=1}^{\ell} \Gamma(\lambda_i)} \prod_{i=1}^{\ell} \hat{p}_i^{\lambda_i - 1} \\ &= \log \Gamma\left(\sum_i \lambda_i\right) - \sum_i \log \Gamma(\lambda_i) + \sum_i (\lambda_i - 1) \log \hat{p}_i. \end{aligned} \tag{9}$$

There are a number of methods for numerically maximizing this objective function $G(\boldsymbol{\lambda})$ 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. (see also Gupta and Upadhyay (2014)). We, however, used the procedure given by Minka (2000). The results of EB estimates are shown in Table 6 where the bracketed values represent the EB estimates corresponding to $2 \times 2 \times 2$ setup. It may be noted here that we have described the EB procedure based on the model formulation given at the beginning of subsection 2.1 although the procedure was actually implemented on (2) for $j = 0, 1$. The description given on (9) is simply meant for notational convenience.

Table 6: EB estimates of the components of $\mathbf{a}_j (j = 0, 1)$ using past data

				OC Use				
	a_{000}	a_{001}	a_{002}	a_{003}	a_{100}	a_{101}	a_{102}	a_{103}
$D = 0$	125.59	24.52	8.33	3.37	8.21	0.42	0.55	0.63
	(126.65)	(35.28)	–	–	(9.49)	(0.19)	–	–
	a_{010}	a_{011}	a_{012}	a_{013}	a_{110}	a_{111}	a_{112}	a_{113}
$D = 1$	98.70	11.25	9.11	3.09	33.07	8.34	0.35	7.70
	(98.63)	(20.06)	–	–	(39.50)	(13.43)	–	–
				Parity				
	a_{000}	a_{001}	a_{002}	a_{003}	a_{100}	a_{101}	a_{102}	a_{103}
$D = 0$	8.00	118.60	30.86	5.78	0.41	6.98	0.43	0.56
	(14.68)	(156.42)	–	–	(0.25)	(0.27)	–	–
	a_{010}	a_{011}	a_{012}	a_{013}	a_{110}	a_{111}	a_{112}	a_{113}
$D = 1$	14.26	74.57	16.07	5.25	7.37	36.99	16.52	0.58
	(13.52)	(104.94)	–	–	(7.02)	(46.14)	–	–

No doubt, the EB criterion is close to classical paradigm because of the involvement of observed data in its evaluation. The method is often criticized by subjective Bayesians who consider prior information exogenous to observations. We, therefore, finally resort to subjective elicitation of hyperparameters based on the description given in subsection 2.1. Since the method is based on elicitation of hyperparameters using quartiles obtained from the experts, we actually work for these quartiles using EB estimates derived from the past data due to non-availability of experts in our case. To be explicit, because we do not have an expert available to specify the quartiles used in the elicitation, we estimated the hyperparameters of the Dirichlet distribution using the EB method just described and applied to a subset of the data. This Dirichlet gives the quartiles that are subsequently used in the elicitation procedure described in subsection 2.1. We do not claim that relying on EB estimates is the only possibility. One can, of course, use alternative method as well.

Table 7: Quartile values and estimates of hyperparameters corresponding to re-parameterized Dirichlet model for $2 \times 2 \times 4$ setup

Quartile constraints/Parameters	OC use		Parity	
	$D = 0$	$D = 1$	$D = 0$	$D = 1$
$(p_q,\ q)$	(0.75, 0.75)	(0.60, 0.75)	(0.05, 0.75)	(0.10, 0.75)
α_1	0.73	0.57	0.05	0.08
$(p_q,\ q)$	(0.16, 0.75)	(0.08, 0.75)	(0.71, 0.75)	(0.46, 0.75)
α_2	0.14	0.07	0.68	0.43
$(p_q,\ q)$	(0.06, 0.75)	(0.06, 0.75)	(0.20, 0.75)	(0.11, 0.75)
α_3	0.05	0.05	0.18	0.09
$(p_q,\ q)$	(0.02, 0.75)	(0.02, 0.75)	(0.04, 0.75)	(0.04, 0.75)
α_4	0.014	0.015	0.03	0.03
$(p_q,\ q)$	(0.06, 0.75)	(0.21, 0.75)	(0.003, 0.75)	(0.05, 0.75)
α_5	0.05	0.19	0.002	0.04
$(p_q,\ q)$	(0.003, 0.75)	(0.06, 0.75)	(0.05, 0.75)	(0.23, 0.75)
α_6	0.002	0.05	0.04	0.22
$(p_q,\ q)$	(0.004, 0.75)	(0.002, 0.75)	(0.003, 0.75)	(0.11, 0.75)
α_7	0.003	0.002	0.002	0.09
$(p_q,\ q)$	(0.71, 0.25)	(0.55, 0.25)	(0.03, 0.25)	(0.07, 0.25)
β	128	128	128	128

Table 8: Quartile values and estimates of hyperparameters corresponding to re-parameterized Dirichlet model for $2 \times 2 \times 2$ setup

Quartile constraints/Parameters	OC use		Parity	
	$D = 0$	$D = 1$	$D = 0$	$D = 1$
$(p_q,\ q)$	(0.76, 0.75)	(0.60, 0.75)	(0.10, 0.75)	(0.09, 0.75)
α_1	0.73	0.57	0.08	0.08
$(p_q,\ q)$	(0.22, 0.75)	(0.13, 0.75)	(0.93, 0.75)	(0.45, 0.75)
α_2	0.20	0.11	0.91	0.61
$(p_q,\ q)$	(0.06, 0.75)	(0.25, 0.75)	(0.001, 0.75)	(0.13, 0.75)
α_3	0.05	0.23	0.001	0.04
$(p_q,\ q)$	(0.71, 0.25)	(0.55, 0.25)	(0.07, 0.25)	(0.07, 0.25)
β	128	128	128	128

Table 9: Estimates of original Dirichlet hyperparameters

OC Use								
	a_{000}	a_{001}	a_{002}	a_{003}	a_{100}	a_{101}	a_{102}	a_{103}
$D = 0$	93.75	18.00	6.25	1.87	6.25	0.31	0.41	1.16
	(93.75)	(25.25)	–	–	(6.25)	(2.75)	–	–
	a_{010}	a_{011}	a_{012}	a_{013}	a_{110}	a_{111}	a_{112}	a_{113}
$D = 1$	73.00	8.50	6.25	1.87	24.00	6.25	0.25	7.87
	(73.00)	(14.50)	–	–	(29.00)	(11.50)	–	–
Parity								
	a_{000}	a_{001}	a_{002}	a_{003}	a_{100}	a_{101}	a_{102}	a_{103}
$D = 0$	5.75	87.25	22.75	4.00	0.31	5.12	0.31	2.50
	(10.75)	(116.75)	–	–	(0.17)	(0.33)	–	–
	a_{010}	a_{011}	a_{012}	a_{013}	a_{110}	a_{111}	a_{112}	a_{113}
$D = 1$	10.37	55.00	12.00	3.75	5.12	27.75	12.00	2.00
	(9.75)	(78.00)	–	–	(5.12)	(35.12)	–	–

Our next step, therefore, consisted of evaluating the quartiles by solving the incomplete beta functions (see subsection 2.1) after replacing the unknown Dirichlet parameters by their EB estimates given in Table 6 (see also (8) and the Definition given in subsection 2.1). Our final step consisted of estimating the parameters $\alpha_i,\ i = 1, ..., 8(4)$ and β by the method suggested in subsection 2.1. These estimated parameters are shown separately for OC use and parity in Table 7 and 8 where Table 7 corresponds to $2 \times 2 \times 4$ setup and Table 8 corresponds to the same for $2 \times 2 \times 2$ setup. These estimates then provide the estimates of original Dirichlet hyperparameters $\mathbf{a_j}, j = 0, 1$, used in our study, which are shown in Table 9. It can be seen that the elicited prior hyperparameters, in general, differ from the corresponding EB estimates (see Tables 6 and 9).

Table 10: Estimated sample based posterior means and the corresponding standard deviations of different cell probabilities for $2 \times 2 \times 4$ set up

Cell probabilities	OC use	Parity	Cell probabilities	OC use	Parity
p_{000}	0.7660	0.0543	p_{010}	0.5916	0.0821
	(0.0137)	(0.0075)		(0.0150)	(0.0084)
p_{001}	0.1192	0.6783	p_{011}	0.0796	0.4474
	(0.0102)	(0.0151)		(0.0084)	(0.0160)
p_{002}	0.0441	0.2035	p_{012}	0.0223	0.1333
	(0.0069)	(0.0132)		(0.0048)	(0.0113)
p_{003}	0.0480	0.0412	p_{013}	0.0182	0.0397
	((0.0076)	(0.0070)		(0.0043)	(0.0063)
p_{100}	0.0173	0.0014	p_{110}	0.2152	0.0257
	(0.0043)	(0.0012)		(0.0132)	(0.0050)
p_{101}	0.0016	0.0150	p_{111}	0.0418	0.2242
	(0.0014)	(0.0044)		(0.0062)	(0.0128)
p_{102}	0.0014	0.0024	p_{112}	0.0076	0.0436
	(0.0013)	(0.0016)		(0.0029)	(0.0066)
p_{103}	0.0024	0.0039	p_{113}	0.0237	0.0040
	(0.0016)	(0.0021)		(0.0048)	(0.0019)

Using these elicited parameters of Dirichlet prior, the final estimates of posterior cell probabilities, log odds ratios and interaction parameters were obtained for the dataset shown in Table 5. The estimated posterior means and the corresponding standard deviations of various cell probabilities are shown in Table 10. We also obtained the estimates of posterior cell probabilities for $2 \times 2 \times 2$ setup but we are not showing those results, instead we shall focus on the quantities which are of more interest to the practitioners. The corresponding estimates for log odds ratio and other interactive parameters are shown in Table 11.

Table 11: Posterior estimates of log odds ratio and other association parameters for $2 \times 2 \times 2$ setup

Parameters	OC use	Parity
θ_{GE}	0.3129	-0.0717
	(0.5328)	(1.1514)
$\log(OR_E)$	-0.3209	-0.6979
	(0.1416)	(0.1961)
$\log(OR_G)$	2.8701	3.0641
	(0.2823)	(1.1365)
$\log(\psi)$	0.1556	0.3804
	(0.5602)	(1.1793)

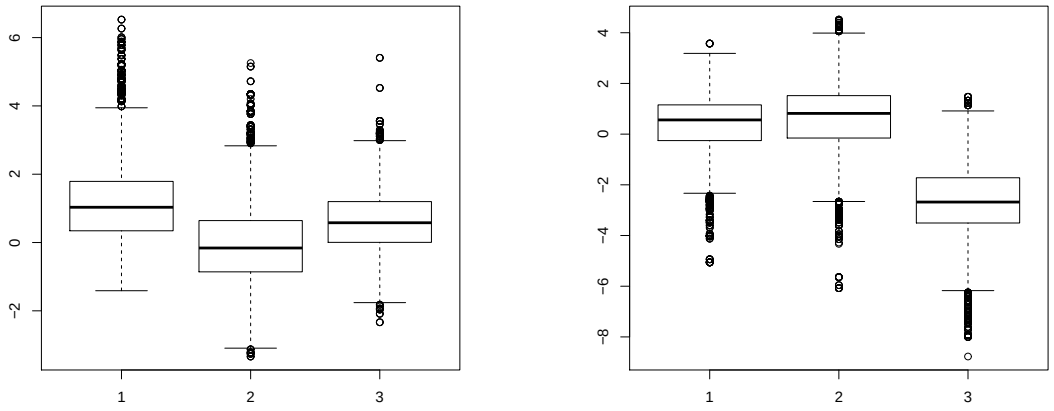
It can be seen from Table 11 that θ_{GE} is non-zero, which clearly conveys the message that genes and environment are associated with each other and cannot be considered independent. $\log(OR_E)$ is negative for both OC use and parity conveying that for non-susceptible subjects both the environmental components reduce the risk of ovarian cancer. Positive values of $\log(OR_G)$ suggest that patients having BRCA1/2 mutation are having high risk of disease in the control population. The estimated values of $\log(\psi)$ are showing positive association among genetic susceptibility and environmental exposure but in general these values are less than those obtained in the control population (refer the values of $\log(OR_G)$ in Table 11). Thus it can be concluded that both OC use and parity reduce the risk of ovarian cancer to a larger extent as compared to the categories of no-OC use and no-parity. It can be further seen that estimated variability along with the estimated posterior means convey that negative values for $\log(\psi)$ are also quite probable for both OC use and parity suggesting a further possibility of decrease in the values of $\log(\psi)$.

Table 12 provides the estimates of log odds ratios and interaction parameters for $2 \times 2 \times 4$ setup. $\theta_{GE_k}, k = 1, ..., 3$, is showing association for both OC use and parity discarding the assumption of independence between genetic susceptibility and environmental exposures. $log(OR_{E_k}), k = 1, ..., 3$, is 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 and/or parity.

Table 12: Posterior estimates of log odds ratio and other association parameters using elicited prior hyperparameters ($2 \times 2 \times 4$ setup)

Posterior estimates	OC use	Parity	Posterior estimates	OC use	Parity
θ_{GE_1}	-0.8273	0.1074	$\log(OR_{E_1})$	-0.1456	-0.8277
	(1.1272)	(1.0785)		(0.1565)	(0.1902)
θ_{GE_2}	0.0016	-0.6126	$\log(OR_{E_2})$	-0.4611	-0.8324
	(1.0690)	(1.2632)		(0.2841)	(0.2119)
θ_{GE_3}	0.6479	1.5276	$\log(OR_{E_3})$	-0.7598	-0.4450
	(0.8311)	(1.2423)		(0.2937)	(0.3035)

The estimated values corresponding to $\log(OR_G)$ are not shown in Table 12. These values were 2.8567(0.2883) and 2.8220(1.0933), respectively, for OC use and parity, where bracketed values correspond to the estimated standard deviations. Therefore, it can be concluded that the patients with BRCA1/2 mutation are having high risk of developing ovarian cancer in control population.



(a) OC use (b) Parity

Figure 1: Posterior density estimates of interaction parameter $\log(\psi_k), k = 1, 2, 3$ (from left to right) corresponding to OC use and parity.

The $log(\psi_k)$, $k = 1, 2, 3$, at different levels of environmental exposure are shown in the form of posterior density estimates using boxplot representations (see Figure 1). Although the density estimates at all the three levels of environmental exposures are showing more or less similar behaviour, they nicely convey an important message on the development of the disease with increasing OC and parity. It can be seen that as the OC use and parity increase, the risk of ovarian cancer decreases. One or two estimates appear exception probably because of the reasons explained earlier. 4

3.1. Results for an extracted data of reduced size

The results were also obtained for an extracted data of size 50 given in Table 3. This was done to see the effect of reduced sample size on our analysis. It is important to mention that if we have very small sample sizes, we may come across a situation where most of the cell frequencies are zero especially for large values of l. We, therefore, did not consider a value of n smaller than 50. The estimated cell probabilities along with their standard deviations (in parentheses) are given in Table 13. It is to be noted here that the prior hyperparameters are same that are presented in Table 9 for $2 \times 2 \times 2$ set up. Contrary to our expectation, we observed that the estimated cell probabilities were showing more or less similar trend that we observed earlier for $2 \times 2 \times 2$ setup.

We finally calculated the posterior estimates of log odds ratio and other interactive parameters, which

Table 13: Estimated sample based posterior means of different cell probabilities for extracted data of size 50

Cell probabilities	OC use	Parity
p_{000}	0.7427	0.0855
	(0.0368)	(0.0242)
p_{001}	0.1854	0.8979
	(0.0323)	(0.0259)
p_{100}	0.0476	0.0077
	(0.0170)	(0.0069)
p_{101}	0.0243	0.0089
	(0.0129)	(0.0075)
p_{010}	0.5788	0.0765
	(0.0377)	(0.0201)
p_{011}	0.1073	0.6035
	(0.0250)	(0.0387)
p_{110}	0.2195	0.0462
	(0.0328)	(0.0173)
p_{111}	0.0944	0.2738
	(0.0240)	(0.0365)

Table 14: Posterior estimates of log odds ratio and other association parameters for extracted data of size 50

Parameters	OC use	Parity
θ_{GE}	0.6503	-2.1705
	(0.7202)	(1.4753)
$\log(OR_E)$	-0.3083	-0.2936
	(0.3599)	(0.4382)
$\log(OR_G)$	1.8339	2.2885
	(0.4399)	(1.2423)
$\log(\psi)$	0.1948	1.9163
	(0.8352)	(1.5567)

are shown in Table 14. It can be seen that the results are certainly changed in terms of numerical values but the final messages are more or less same that we drew earlier based on $2 \times 2 \times 2$ setup for large sample size. That is, θ_{GE} is showing an association among genetic and environmental components for both OC use and parity indicating the need of going a step ahead for evaluating the multiplicative interaction parameter ψ. Similarly, for non-susceptible subjects both OC use and parity reduce the risk of ovarian cancer whereas the subjects unexposed to environmental factors but having BRCA1/2 mutation are having a higher risk of the disease. The estimated ψ is also showing a similar behaviour with values indicating that OC use and parity reduce the risk of ovarian cancer even in patients with BRCA1/2 mutation. This reduction is stronger for OC use, than for parity.

4. Conclusion

The paper provides a complete Bayesian analysis of a gene-environment problem where the main focus is on prior elicitation of hyperparameters. The approach based on subjective elicitation of prior hyperparameters and subsequently the use of such elicited prior in drawing the final posterior inferences can always be regarded in a true Bayesian spirit. It is, however, often the case that availability of expert is not always guaranteed and, as such, the appropriate specification of prior hyperparameters becomes difficult. The methodology given here recommends the use of past data or any small subset of the given data to achieve the process of elicitation. It is seen that the considered methodology is, in general, easy to implement and provides a systematic way to tackle an important issue of hyperparameter selection. To the best of our understanding, no such attempt was ever made on the specific application considered in the paper where it is perhaps always desired to refine the inferences. The future researches may consider defining such strategies in general where one has priors other than the

conjugate Dirichlet distribution. Elicitation of the complete functional form of the prior distribution is another important direction where future researchers may proceed.

Acknowledgements

The authors express their thankfulness to the editor and the referees for their valuable comments and suggestions that improved the earlier version of the manuscript.

References

Chaloner KM, Duncan GT (1983). "Assessment of a Beta Prior Sistribution: PM Elicitation." *Statistician*, **32**, 174–180.

Chatterjee N, Carroll RJ (2005). "Semiparametric Maximum Likelihood Estimation Exploiting Gene-environment Independence in Case-control Studies." *Biometrika*, **92**, 399–418.

Dorp JR, Mazzuchi TA (2003). "Parameter Specification of the Beta Distribution and Its Dirichlet Extensions Utilizing Quantiles." *Beta Distribution and Its Applications*, **29**(1), 1–37.

Evans M, Guttman I, Li P (2017). "Prior Elicitation, Assessment and Inference with a Dirichlet Prior." *Entropy*, **19**(10), 564.

Garthwaite PH, Dickey JM (1988). "Quantifying Expert Opinion in Linear Regression Problems." *Journal of the Royal Statistical Society Series B*, **50**, 462–474.

Gupta A, Upadhyay SK (2014). "An Empirical Bayes Study to Examine the Interaction between Genetic Susceptibility and Environmental Exposure." *Aligarh Journal of Statistics*, **34**, 105–119.

Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC (1980). "Interactive Elicitation of Opinion for a Normal Linear Model." *Journal of the American Statistical Association*, **75**, 845–854.

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

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 JP, Tucker MA, Wacholder S (2001). "Parity, Oral Contraceptives and the Risk of Ovarian Cancer among Carriers and Non-carriers of a BRCA1 or BRCA2 Mutation." *The New England Journal of Medicine*, **345**, 235–240.

Mukherjee B, Ahn J, Gruber SB, Ghosh M, Chatterjee N (2010). "Case-control Studies of Gene-environment Interaction: Bayesian Design and Analysis." *Biometrics*, **66**(3), 934–948.

Mukherjee B, 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.

O'Hagan A (2006). "Research in Elicitation." *Bayesian Statistics and its Applications*, pp. 375–382.

Staël von Holstein CAS (1971). "The Effect of Learning on the Assessment of Subjective Probability Distributions." *Organizational Behavior and Human Performance*, **6**, 304–315.

Winkler RL, Smith WS, Kulkarni RB (1978). "Adaptive Forecasting Models Based on Predictive Distributions." *Management Science*, **24**, 977–986.

Zapata-Vázquez R, O'Hagan A, Bastos L (2014). "Eliciting Expert Judgements about a Set of Proportions." *Journal of Applied Statistics*, **41**(9), 1919–1933.

Affiliation:

Akanksha Gupta
Department of Statistics &
DST Centre for Interdisciplinary Mathematical Sciences
Banaras Hindu University, Varanasi-221 005, India.
E-mail: akankshagupta1606@gmail.com

S. K. Upadhyay
Department of Statistics &
DST Centre for Interdisciplinary Mathematical Sciences
Banaras Hindu University, Varanasi-221 005, India.
E-mail: skupadhyay@gmail.com

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 48
January 2019

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2018-03-24
Accepted: 2018-10-28