Brazilian Journal of Probability and Statistics
2020, Vol. 34, No. 1, 71–89
https://doi.org/10.1214/18-BJPS413

Effects of gene–environment and gene–gene interactions in case-control studies: A novel Bayesian semiparametric approach

Durba Bhattacharya and Sourabh Bhattacharya

Indian Statistical Institute

Abstract. Present day bio-medical research is pointing towards the fact that cognizance of gene–environment interactions along with genetic interactions may help prevent or detain the onset of many complex diseases like cardiovascular disease, cancer, type2 diabetes, autism or asthma by adjustments to lifestyle.

In this regard, we propose a Bayesian semiparametric model to detect not only the roles of genes and their interactions, but also the possible influence of environmental variables on the genes in case-control studies. Our model also accounts for the unknown number of genetic sub-populations via finite mixtures composed of Dirichlet processes. An effective parallel computing methodology, developed by us harnesses the power of parallel processing technology to increase the efficiencies of our conditionally independent Gibbs sampling and Transformation based MCMC (TMCMC) methods.

Applications of our model and methods to simulation studies with biologically realistic genotype datasets and a real, case-control based genotype dataset on early onset of myocardial infarction (MI) have yielded quite interesting results beside providing some insights into the differential effect of gender on MI.

1 Introduction

Although many people tend to classify the cause of a disease as either genetic or environmental, only a few diseases like Huntington's Disease (HD) or GM2 gangliosidosis have so far been identified as purely genetic disorders. As indicated by many epidemiological studies, a different effect of a genotype is often observed on disease risk in persons with different environmental exposures (see Mapp (2003), Khouri (2005)). Also there may be multiple genes which interact with each other to cause a disease only when an environmental factor passes a given threshold, implying thereby that presence of a risk allele may not be exposing all individuals to the same risk.

Hunter (2005) and Mather and Caligary (1976), point out that estimation of only the separate contributions of genes and environment to a disease, ignoring their interactions, will lead to incorrect estimation of the proportion of the disease (the "population attributable fraction") that is explained by the genes, the environment, and their joint effect.

Study of gene–environment interaction is important to the field of pharmacogenetics also, since the efficacy and side-effects of some medications can vary depending on an individual's genotype (see Scott (2011)). Hence, extensive study of gene–environment interactions through sophisticated statistical modelling is necessary to devise new methods of disease prevention, detection and intervention.

Gene–environment interaction is often conceptualized as genetic control of sensitivity to different environments (Purcell (2002)). According to Mather and Caligary (1976) (see

Key words and phrases. Case-control study, Dirichlet process, gene–gene and gene–environment interaction, matrix normal, parallel processing, transformation based MCMC.

Received August 2017; accepted August 2018.

also Ottman (2010)) gene–environment interaction is defined as "a different effect of an environmental exposure on disease risk in persons with different genotypes". As genes are the fundamental units of change in an environmental response system, in order to model the gene–environment interaction effectively, it is important to understand the mechanism through which genes and environment interact together to bring about a physiological change in an individual. An environmental exposure could trigger a physiological change in a number of ways. Exposure to certain environmental stimuli may directly or indirectly alter the epigenome of an individual, which is a network of chemical compounds surrounding DNA that modify the genome without altering the DNA sequences and have a role in determining which genes are active in a particular cell. Exposure to mutagens like high doses of x-ray or nuclear radiation, smoking etc. can enter into the body through tissues and directly interfere with the DNA sequence or replication mechanism. Some environmental stimuli may affect DNA indirectly by altering transcription factors and hence changing the expressions of certain genes. Many gene–gene interactions have been shown to be started by some environmental exposure. For example, excessive alcohol intake has been shown to suppress TACE gene, which then activates less MTHFR, resulting in reduced folate metabolism, causing depression.

Although the study of gene–environment interaction has become essential to the understanding of the aetiology of almost every disease, very little success has so far been achieved in this field. This want of success may be attributed to many causes like inadequacy of models incorporating the complex mechanism through which genes and environment may affect a disease risk (Wang, Elston and Zhu (2010)). Indeed, given the complexity involved in the gene–environment interactions, no simple linear or additive relationship alone can model the relationship effectively. According to Wright, Carothers and Campbell (2002) and Wang, Elston and Zhu (2010), although statistical definition of gene environment interaction may lack clear biological interpretations, quantification of biological interaction should be based on statistical concepts of interaction. Furthermore, inadequacy of data regarding environmental exposure of individuals and stratified population structure are also important factors impeding success of the existing methods in this field. Association tests based on a pooled set of genetically diverse sub-populations (i.e., having differences in allele frequencies across subpopulations) may result in extremely inflated rates of false positives (see Bhattacharjee et al. (2010)).

The above discussion points towards the fact that the widely-used log-linear models (see, for example, Mukherjee et al. (2008, 2010, 2012), Mukherjee and Chatterjee (2008), Sanchez, Kang and Mukherjee (2012), Ahn et al. (2013), Ko et al. (2013)) are perhaps not quite adequate for modeling complex gene–gene and gene–environment interactions. Moreover, such models consider quite restrictive and ad-hoc association structures for simplifying computation and only attempt to test whether or not the interaction is present without being able to quantify the strength of the interaction. Uncertainty regarding unknown number of subpopulations are also not generally accounted for in the existing interaction models.

Our Bayesian hierarchical mixture model framework is aimed at incorporating all the aforementioned desirable mechanisms through which gene–environment interaction, along with the isolated effects of genes and their interactions may affect an individual's risk of being affected by a disease, taking into account the fact that the underlying population may be stratified in nature. Since the number of sub-populations is not usually known, one must coherently and carefully account for the uncertainty associated with the unknown number of sub-populations. An additional feature of our model is learning about the number of underlying genetic sub-populations.

Because of dependence on environmental variables, our Bayesian semiparametric model comprises Dirichlet process (henceforth, DP) finite mixture models even at the individual

subject level, in addition to genetic and case-control status. The mixtures share a complex dependence structure between themselves through suitable hierarchical matrix-normal distributions, suitably taking account of the dependence induced by the environmental variable. To detect the roles of genes, environment, gene–gene and gene–environment interactions, we extend the gene–gene interaction model and the associated Bayesian hypotheses testing methods of Bhattacharya and Bhattacharya (2016) (henceforth, BB), and for the purpose of computation we develop a powerful parallel Markov chain Monte Carlo (MCMC) algorithm which exploits the conditional independence structures inherent in our Bayesian model, and combines the efficiencies of our Gibbs sampling method associated with the mixtures and Transformation based MCMC (TMCMC) of Dutta and Bhattacharya (2014). It is to be noted that parallel computation in statistics is not very rare nowadays. In contexts different from ours, promising parallelisable MCMC algorithms are making their appearances in the recent times; see Martino, Elvira and Camps-Valls (2018), Martino et al. (2016), Chen et al. (2016), Jacob, Robert and Smith (2011), Calderhead (2014) and Brockwell (2006). These works aim to improve the performance, save computational cost of Gibbs samplers and to parallelize the Metropolis-Hastings technique under various setups.

The rest of our paper is structured as follows. We introduce our proposed Bayesian semiparametric gene–environment interaction model in Section 2. In Section 3, we extend the Bayesian hypothesis testing procedures proposed in BB to learn about the roles of genes, environmental variables and their interactions in case-control studies. In Section 4, we demonstrate the validity of our model and methods with successful applications to five biologically realistic simulated data sets associated with five different set-ups. We also analysed a case-control type myocardial infarction data set obtained from dbGap with our model and methods, the results of which we report and discuss in detail in Section 5. As we point out, our results broadly agree with and in some cases contrast the existing results on this data set. Finally, we summarize our work with concluding remarks in Section 6. Further details are provided in the supplement (Bhattacharya and Bhattacharya (2020)), whose sections and figures have the prefix "S-" when referred to in this paper. The main notations, abbreviations and their explanations associated with our work are summarized in Table 1.

2 A new Bayesian semiparametric model for gene–gene and gene–environment interactions

2.1 Case-control genotype data

We first notify the statistical reader that each cell of the human body consists of 23 pairs of chromosomes; one chromosome from each pair is inherited from the mother and the other from the father. Now, for $s = 1, 2$ denoting the two chromosomes inherited from mother and father, let $x^s_{ijkr} = 1/0$ indicate respectively, the presence and absence of the minor allele at rth locus of the jth gene for the ith individual belonging to the kth group of case/control, where $k = 0, 1$, with $k = 1$ denoting case; $i = 1, \ldots, N_k$; $r = 1, \ldots, L_j$ and $j = 1, \ldots, J$; let $N = N_0 + N_1$. Here minor allele refers to the second most common allele occurring in a given population.

Let $\boldsymbol{E}_i$ denote a set of environmental variables associated with the ith individual. In what follows, we model this case-control genotype data, along with the information on the environmental variables using our Bayesian semiparametric model, described in the next few sections.

Table 1 *Notations and their explanations*

Notation	Explanation
x^s_{ijkr}	Indicates the presence/absence of minor allele for ith individual, jth gene, kth case/control status at the rth locus, where $s = 1, 2$ indicates the two chromosomes; $i = 1, \ldots, N_k$ indicates the individuals; $j = 1, \ldots, J$ denotes the J genes; $k = 1/0$ stands for case/control; $r = 1, \ldots, L_j$ denotes the L_j loci corresponding to the jth gene.
$\mathbf{x}_{ijkr}$	$= (x^1_{ijkr}, x^2_{ijkr})$ represents the genotype of the ith individual, jth gene, belonging to the kth group of case/control at the rth locus.
$\boldsymbol{X}_{ijk}$	$= (\mathbf{x}_{ijk1}, \mathbf{x}_{ijk2}, \ldots, \mathbf{x}_{ijkL_j})$ denotes the genotype information at all the L_j loci of the ith individual's jth gene of the kth case/control group.
$\boldsymbol{E}_i$	Set of environmental variables associated with the ith individual.
p_{mijkr}	Denotes the minor allele frequency related to the mth subpopulation, ith individual, jth gene, kth case/control group at the rth locus.
$\boldsymbol{p}_{mijk}$	$= (p_{mijk1}, p_{mijk2}, \ldots, p_{mijkL_j})$ denotes the vector of minor allele frequencies corresponding to the mth subpopulation at the L_j loci of the ith individual's jth gene belonging to the kth group of case/control.
π_{mijk}	The unknown probability of the mth mixture component, where $m = 1, \ldots, M$; M is the maximum possible number of subpopulations.
z_{ijk}	Allocation variables such that $P[z_{ijk} = m] = \pi_{mijk}$.
$DP(\alpha_{ijk}\boldsymbol{G}_{0,ijk})$	Stands for Dirichlet process with expected probability measure $\boldsymbol{G}_{0,ijk}$ having precision parameter α_{ijk}.
$\mathcal{B}(\nu_{1ijkr}, \nu_{2ijkr})$	Stands for Beta distribution with non-negative parameters ν_{1ijkr}, ν_{2ijkr}.
$\mathcal{D}(\alpha_{1ijk}, \ldots, \alpha_{M_0,ijk})$	Dirichlet distribution with non-negative parameters $\alpha_{1ijk}, \ldots, \alpha_{M_0,ijk}$.
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Stands for p-dimensional ($p > 1$) multivariate normal disributtion with mean vector $\boldsymbol{\mu}$ and variance—covariance matrix $\boldsymbol{\Sigma}$.
$N(\mu, \sigma^2)$	Stands for univariate normal distribution with mean μ and variance σ^2.
$\mathcal{G}(\alpha, \beta)$	Stands for gamma distribution with non-negative parameters α and β.

2.2 Modeling genotypic sub-populations with mixture models driven by Dirichlet processes

Let $\mathbf{x}_{ijkr} = (x^1_{ijkr}, x^2_{ijkr})$ represent the genotype of ith individual, jth gene, belonging to the kth group of case/control at the rth locus, and let $\boldsymbol{X}_{ijk} = (\mathbf{x}_{ijk1}, \mathbf{x}_{ijk2}, \ldots, \mathbf{x}_{ijkL_j})$ denote the genotype information at all the L_j loci of ith individual's jth gene of the kth group. Also, let p_{mijkr} stand for the minor allele frequency related to the mth subpopulation, ith individual, jth gene, kth case/control group at the rth locus. Note that minor allele frequency is the frequency at which the second most common allele occurs in a given population.

We assume that for every triplet (i, j, k), $\boldsymbol{X}_{ijk}$ have the mixture distribution

$$[\boldsymbol{X}_{ijk}] = \sum_{m=1}^{M} \pi_{mijk} \prod_{r=1}^{L_j} f(\mathbf{x}_{ijkr} | p_{mijkr}), \tag{2.1}$$

where $f(\cdot|p_{mijkr})$ is the Bernoulli mass function given by

$$f(\mathbf{x}_{ijkr}|p_{mijkr}) = \{p_{mijkr}\}^{x^1_{ijkr}+x^2_{ijkr}} \{1 - p_{mijkr}\}^{2-(x^1_{ijkr}+x^2_{ijkr})}, \tag{2.2}$$

and M denotes the *maximum* number of mixture components possible, with π_{mijk} being the (unknown) probability of the mth mixture component.

Allocation variables z_{ijk}, with probability distribution

$$[z_{ijk} = m] = \pi_{mijk}, \tag{2.3}$$

for $i = 1, \ldots, N_k$ and $m = 1, \ldots, M$, allow representation of (2.1) as

$$[\boldsymbol{X}_{ijk}|z_{ijk}] = \prod_{r=1}^{L_j} f(\mathbf{x}_{ijkr}|p_{z_{ijk}ijkr}). \tag{2.4}$$

Following Majumdar et al. (2013), BB, we set $\pi_{mijk} = 1/M$, for $m = 1, \ldots, M$, and for all (i, j, k).

Letting $\boldsymbol{p}_{mijk} = (p_{mijk1}, p_{mijk2}, \ldots, p_{mijkL_j})$ denote the vector of minor allele frequencies corresponding to the mth subpopulation at the L_j loci of the ith individual's jth gene belonging to the kth group of case/control. We next assume that

$$\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk} \overset{\text{iid}}{\sim} \boldsymbol{G}_{ijk}; \tag{2.5}$$

$$\boldsymbol{G}_{ijk} \sim \text{DP}(\alpha_{ijk}\boldsymbol{G}_{0,ijk}), \tag{2.6}$$

where $\text{DP}(\alpha_{ijk}\boldsymbol{G}_{0,ijk})$ stands for Dirichlet process with expected probability measure $\boldsymbol{G}_{0,ijk}$ having precision parameter α_{ijk}. We specify the base probability measure $\boldsymbol{G}_{0,ijk}$ as follows: for $m = 1, \ldots, M$ and $r = 1, \ldots, L_j$,

$$p_{mijkr} \overset{\text{iid}}{\sim} \mathcal{B}(\nu_{1ijkr}, \nu_{2ijkr}), \tag{2.7}$$

under $\boldsymbol{G}_{0,ijk}$. Coincidences among $\boldsymbol{P}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk}\}$, which occur with positive probability, is the property of the DP based mixture models that we exploit to learn about the actual number of mixture components.

The associated Polya urn distribution of $\boldsymbol{P}_{Mijk}$ can be derived by marginalizing over $\boldsymbol{G}_{ijk}$:

$$\begin{aligned}\left[\boldsymbol{p}_{mijk}|\boldsymbol{P}_{Mijk} \setminus \{\boldsymbol{p}_{mijk}\}\right] \sim{}& \frac{\alpha_{ijk}}{\alpha_{ijk} + M - 1}\boldsymbol{G}_{0,ijk}(\boldsymbol{p}_{mijk}) \\ &+ \frac{1}{\alpha_{ijk} + M - 1}\sum_{m' \neq m=1}^{M} \delta_{\boldsymbol{p}_{m'ijk}}(\boldsymbol{p}_{mijk}),\end{aligned} \tag{2.8}$$

where $\delta_{\boldsymbol{p}_{m'ijk}}(\cdot)$ denotes point mass at $\boldsymbol{p}_{m'ijk}$. This scheme is useful for constructing an efficient Gibbs sampling strategy for simulating the mixtures conditional on the other parameters, embedded in a parallel MCMC strategy that we devise, bypassing the infinite-dimensional random measure $\boldsymbol{G}_{ijk}$.

Coincidences among the mixture components associate the triplets (i, j, k) to different mixtures with varying number of components. Indeed, the genotype distributions of any two individuals i and i' arising from a given sub-population with the same gene indexed by j but with different case-control status, are likely to be different, so that $(i, j, k = 0)$ and $(i', j, k = 1)$ may correspond to different mixtures. Also, for any two genes indexed by j and j', (i, j, k) and (i, j', k) may correspond to different mixtures because of differences in the distribution of genotypes of genes j and j' for the ith individual. Furthermore, for any two individuals indexed by i and i', (i, j, k) and (i', j, k) are likely to be associated with different mixtures because the genotype distribution of the jth gene may be affected by different environmental exposures $\boldsymbol{E}_i$ and $\boldsymbol{E}_{i'}$. Thus, it seems that the DP based mixtures realistically take account of the various genotypic sub-populations and the number of such sub-populations the data arise from.

The above ideas are similar in essence to those in BB, but note that in their case, since the environmental effect $\boldsymbol{E}_i$ is not considered, the mixtures were with respect to (j, k) only, not with respect to (i, j, k) as in our current scenario influenced by $\boldsymbol{E}_i$.

Following BB, we set M, the maximum possible number of sub-populations to be 30 and $\alpha_{ijk} = 10$ in our applications. These choices are not affected by the presence of environmental variables, and performed adequately in our Bayesian analyses.

We remark that when the population structure is accurately known, then the situation is rendered a special case of our DP formulation. To clarify, first let us suppose that the true number of mixture components, say, $M_{0,ijk}$, are known. In that case, we shall set $M = M_{0,ijk}$ and let $\alpha_{ijk} \to \infty$, so that the DP tends to point mass on $\boldsymbol{G}_{0,ijk}$. In practice, we shall simply assume that $\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{M_0,ijk} \overset{\text{iid}}{\sim} \boldsymbol{G}_{0,ijk}$. Thus, the representation (2.1) reduces to an $M_{0,ijk}$-component parametric mixture, where the parameters independently follow the above $\boldsymbol{G}_{0,ijk}$ distribution. Here, rather than setting $\pi_{mijk} = 1/M$, we shall assume that $(\pi_{1ijk}, \ldots, \pi_{M_0,ijk}) \sim \mathcal{D}(\alpha_{1ijk}, \ldots, \alpha_{M_0,ijk})$, a Dirichlet distribution with non-negative parameters $\alpha_{1ijk}, \ldots, \alpha_{M_0,ijk}$. As is usual, we may consider a non-informative Dirichlet prior by setting $\alpha_{mijk} = 1$, for $m = 1, \ldots, M_{0,ijk}$. The rest of our model remains the same as before. Since gene–gene and gene–environment interaction effects are modeled via the parameters of $\boldsymbol{G}_{0,ijk}$, reduction of the Bayesian nonparametric mixture model to Bayesian parametric mixture model does not compromise with the interaction effects. The same is true when not only the number of components, but even the population stratification is explicitly known. Indeed, in such case the allocation variables z_{ijk} are known, so that the probability π_{mijk} on the right-hand side of (2.3) is either 0 or 1. In other words, the mixture representation (2.1) reduces to a single-component distribution with the associated parameter vector $\boldsymbol{p}_{z_{ijk}ijk} \sim \boldsymbol{G}_{0,ijk}$. The above scenarios clearly leads to great reduction in computational complexity. Unfortunately, in practice usually the population structure is unknown and hence we recommend the DP based mixture model in general.

We conclude this section on DP based mixture modeling by drawing attention to a different style of learning population structure which proceeds by modeling the allocation variables by DP with discrete base measure; see De Iorio, Favaro and Teh (2015) and the references therein for the details.

2.3 Modeling the complex gene–gene and gene–environment dependence structure with appropriate modeling of the parameters of $\boldsymbol{G_{0,ijk}}$

We specify the dependence structure between the genes and the environment by primarily seeing to it that the environment may act upon gene–gene interaction without affecting the marginal distributions of the genotypes of the individual genes. However, we also take into account the fact that in some cases the environmental variables may cause changes in the distributions of the genotypes. Modelling the parameters of the expected probability measure $\boldsymbol{G}_{0,ijk}$ through a relevant hierarchical matrix-normal prior helps us incorporate the complex G × E, G × G and also the SNP × SNP effects appropriately.

2.3.1 *Modeling the parameters of* $\boldsymbol{G}_{0,ijk}$. We model ν_{1ijkr} and ν_{2ijkr}, for each loci $r = 1, \ldots, L_j$, in jth gene, of every individual i, having case or control status k, that is for every (i, j, k), as the following:

$$\nu_{1ijkr} = \exp(u_{jr} + \lambda_{ijk} + \mu_{jk} + \boldsymbol{\beta}'_{jk}\boldsymbol{E}_i); \tag{2.9}$$

$$\nu_{2ijkr} = \exp(v_{jr} + \lambda_{ijk} + \mu_{jk} + \boldsymbol{\beta}'_{jk}\boldsymbol{E}_i). \tag{2.10}$$

In the above, for fixed k, $u_{jr} + \lambda_{ijk} + \mu_{jk}$ can be interpreted as the total additive effect of the rth SNP of the jth gene of the ith individual, where u_{jr} is the effect of the rth locus of the jth gene, λ_{ijk} is the effect of the jth gene of the ith individual and μ_{jk} is the effect of the jth gene. Allowing u_{jr} to be different from v_{jr} ensures that the mean of p_{mijkr} under $\boldsymbol{G}_{0,ijk}$ depends upon the rth SNP of the jth gene. The complex dependence structure that may exist between the SNPs within a gene and between the genes has been incorporated in our model by the parameters u_{jr}, v_{jr} and λ_{ijk}, μ_{jk} respectively (see BB for details). Here $\boldsymbol{E}_i$ is the d-dimensional vector of continuous environmental variables for the ith individual.

The model can be easily extended to include categorical environmental variables along with the continuous ones.

Note that, non-null $\boldsymbol{\beta}_{jk}$ indicates significant marginal effect of the environmental variable $\boldsymbol{E}$ on the jth gene. In Section 2.3.2, we introduce a modeling strategy that accounts for the complex phenomenon through which gene–gene interaction gets modified under the environmental effect, even though the marginal effects of the genes remain unchanged.

2.3.2 *Matrix normal prior for λ_{ijk}'s.* Let $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_J)$, where $\boldsymbol{\lambda}_j = (\lambda_{1j0}, \ldots, \lambda_{n_0 j0}, \lambda_{1j1}, \ldots, \lambda_{n_1 j1})$, for $j = 1, \ldots, J$. Note that λ_{ijk} is shared by every locus of the jth gene of the individual indexed by (i, k).

We consider the following model for $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} \sim \mathcal{N}_J(\boldsymbol{\xi}, \boldsymbol{A} \otimes \tilde{\boldsymbol{\Sigma}}), \tag{2.11}$$

where $\boldsymbol{A}$ is the $J \times J$ left covariance matrix, indicating gene–gene interaction in the absence of environmental effect, and $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \phi\mathcal{E}$ is the right covariance matrix under the effect of the environmental variable E. Here $\phi \geq 0$, $\boldsymbol{\Sigma}$ is some positive definite matrix, and the (i, j)th element of the positive definite matrix $\mathcal{E}$, associated with the environmental variable $\boldsymbol{E}$, is given by

$$\mathcal{E}_{ij} = \exp\bigl(-b\|\boldsymbol{E}_i - \boldsymbol{E}_j\|^2\bigr), \tag{2.12}$$

where $b > 0$ is a smoothness parameter.

Note that $\phi = 0$ indicates absence of environmental effects on gene–gene interaction. It is quite important to observe that, because of the above Gaussian assumption, even for non-zero ϕ, which points towards indirect effect of environmental factors on the epigenome, triggering genetic interactions, the marginal genotypic distributions associated with the J genes of our model remain unaffected by $\boldsymbol{E}$.

For convenience, we represent the JN-dimensional vector $\boldsymbol{\lambda}$ as a $J \times N$ matrix $\boldsymbol{\Lambda}$, which has the following probability density function:

$$\pi(\boldsymbol{\Lambda}) = \frac{\exp[-\operatorname{tr}\{\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\Lambda} - \boldsymbol{\xi})^T \boldsymbol{A}^{-1}(\boldsymbol{\Lambda} - \boldsymbol{\xi})\}]}{(2\pi)^J |\boldsymbol{A}|^N |\boldsymbol{\Lambda}|^J}. \tag{2.13}$$

It follows that

$$\boldsymbol{\Lambda}^{\mathrm{col},k} \sim \mathcal{N}_J\bigl(\boldsymbol{\xi}^{\mathrm{col},k}, \tilde{\sigma}_{kk}\boldsymbol{A}\bigr), \tag{2.14}$$

where $\boldsymbol{\Lambda}^{\mathrm{col},k}$ and $\boldsymbol{\xi}^{\mathrm{col},k}$ are the kth columns of $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}$, respectively. The covariance matrix between $\boldsymbol{\Lambda}^{\mathrm{col},k_1}$ and $\boldsymbol{\Lambda}^{\mathrm{col},k_2}$ is given by

$$\operatorname{cov}\bigl(\boldsymbol{\Lambda}^{\mathrm{col},k_1}, \boldsymbol{\Lambda}^{\mathrm{col},k_2}\bigr) = \tilde{\sigma}_{k_1 k_2}\boldsymbol{A}, \tag{2.15}$$

where $\tilde{\sigma}_{k_1 k_2}$ denotes the (k_1, k_2)th element of $\tilde{\boldsymbol{\Sigma}}$. Also,

$$\boldsymbol{\Lambda}^{\mathrm{row},j} \sim \mathcal{N}_N\bigl(\boldsymbol{\xi}^{\mathrm{row},j}, a_{jj}\tilde{\boldsymbol{\Sigma}}\bigr), \tag{2.16}$$

where $\boldsymbol{\Lambda}^{\mathrm{row},j}$ and $\boldsymbol{\xi}^{\mathrm{row},j}$ are the jth rows of $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}$, respectively. Further,

$$\operatorname{cov}\bigl(\boldsymbol{\Lambda}^{\mathrm{row},j_1}, \boldsymbol{\Lambda}^{\mathrm{row},j_2}\bigr) = a_{j_1 j_2}\tilde{\boldsymbol{\Sigma}}. \tag{2.17}$$

In our applications, following BB, we choose $\boldsymbol{\xi} = \boldsymbol{0}$.

To summarize, the matrix-normal prior imposes a dependence structure between the genes through the gene–gene interaction matrix $\boldsymbol{A}$, and $\tilde{\boldsymbol{\Sigma}}$ features the direct or indirect effect of the environmental factors, on the epigenome of the individuals. The randomness associated with the matrix-normal prior on $\boldsymbol{\Lambda}$ incorporates dependence between the SNPs within a gene.

Further discussion regarding the effect of environmental variables on gene–gene interaction is provided in Section S-1 of the supplement.

2.3.3 *Priors for u_{jr} and v_{jr}.* We follow BB in setting, for $j = 1, \ldots, J$, $u_{jr'} = u_{r'}$ and $v_{jr'} = v_{r'}$ for $r' = 1, \ldots, L$, where $L = \max\{L_j; j = 1, \ldots, J\}$, and assuming for $r' = 1, \ldots, L$,

$$u_{r'} \overset{\text{iid}}{\sim} N(0, 1); \tag{2.18}$$

$$v_{r'} \overset{\text{iid}}{\sim} N(0, 1). \tag{2.19}$$

See BB for the details regarding the choice of u_{jr} and v_{jr}.

2.3.4 *Priors on μ_{jk}, $\boldsymbol{\beta}_{jk}$, $\boldsymbol{A}$, $\boldsymbol{\Sigma}$, b and ϕ.* We put the following hierarchical priors on $\boldsymbol{\mu} = (\mu_{jk}; j = 1, \ldots, J; k = 0, 1)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_\ell; \ell = 1, \ldots, D)$, where $\boldsymbol{\beta}_\ell = (\beta_{\ell jk}; j = 1, \ldots, J; k = 0, 1)$:

$$\boldsymbol{\mu} \sim \mathcal{N}_{2J}(\mathbf{0}, \boldsymbol{A}_\alpha \otimes \boldsymbol{\Sigma}_\alpha); \tag{2.20}$$

$$\boldsymbol{\beta}_\ell \overset{\text{iid}}{\sim} \mathcal{N}_{2J}(\mathbf{0}, \boldsymbol{A}_\beta \otimes \boldsymbol{\Sigma}_\beta); \qquad \ell = 1, \ldots, D. \tag{2.21}$$

For priors on $\boldsymbol{A}_\alpha$, $\boldsymbol{A}_\beta$, $\boldsymbol{\Sigma}_\alpha$ and $\boldsymbol{\Sigma}_\beta$, we first consider their respective Cholesky decompositions: $\boldsymbol{A}_\alpha = \boldsymbol{C}_\alpha \boldsymbol{C}'_\alpha$, $\boldsymbol{A}_\beta = \boldsymbol{C}_\beta \boldsymbol{C}'_\beta$, $\boldsymbol{\Sigma}_\alpha = \boldsymbol{D}_\beta \boldsymbol{D}'_\beta$ and $\boldsymbol{\Sigma}_\beta = \boldsymbol{D}_\beta \boldsymbol{D}'_\beta$. We assume that the diagonal elements of the above Cholesky factors are identically and independently distributed as $\mathcal{G}(0.01, 0.01)$, that is, gamma distribution with mean 1 and variance 100. We assume the non-zero off-diagonal elements of the Cholesky factors to be identically and independently distributed as $N(0, 10^2)$.

Using the same Cholesky decomposition idea, we assume that the off-diagonal elements of the Cholesky factors of $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ to be identically and independently distributed as $N(0, 10^2)$, and the diagonal elements to be identically and independently distributed as $\mathcal{G}(0.01, 0.01)$.

We put log-normal priors on b and ϕ, so that both $\log(b)$ and $\log(\phi)$ are normally distributed with mean zero and variance 100.

Recall that the mixtures associated with gene $j \in \{1, \ldots, J\}$, and individual $i \in \{1, \ldots, N_k\}$ and case-control status $k \in \{0, 1\}$, are conditionally independent of each other, given the interaction parameters. This allows us to update the mixture components in separate parallel processors, conditionally on the interaction parameters. Once the mixture components are updated, we update the interaction parameters using a specialized form of TMCMC, in a single processor. A schematic representation of our model and the parallel processing algorithm is provided in Figure 1. Details of our parallel processing algorithm are provided in Section S-2 of the supplement.

3 Detection of the roles of environment, genes and their interactions in case-control studies

3.1 Formulation of appropriate Bayesian hypothesis testing procedures

In order to investigate if genes have any effect on case-control, we first define

$$h_{0j}(\cdot) = \frac{1}{M} \sum_{m=1}^{M} \prod_{r=1}^{L_j} f\big(\cdot | p^r_{mi_0 jk=0}\big); \tag{3.1}$$

$$h_{1j}(\cdot) = \frac{1}{M} \sum_{m=1}^{M} \prod_{r=1}^{L_j} f\big(\cdot | p^r_{mi_1 jk=1}\big), \tag{3.2}$$

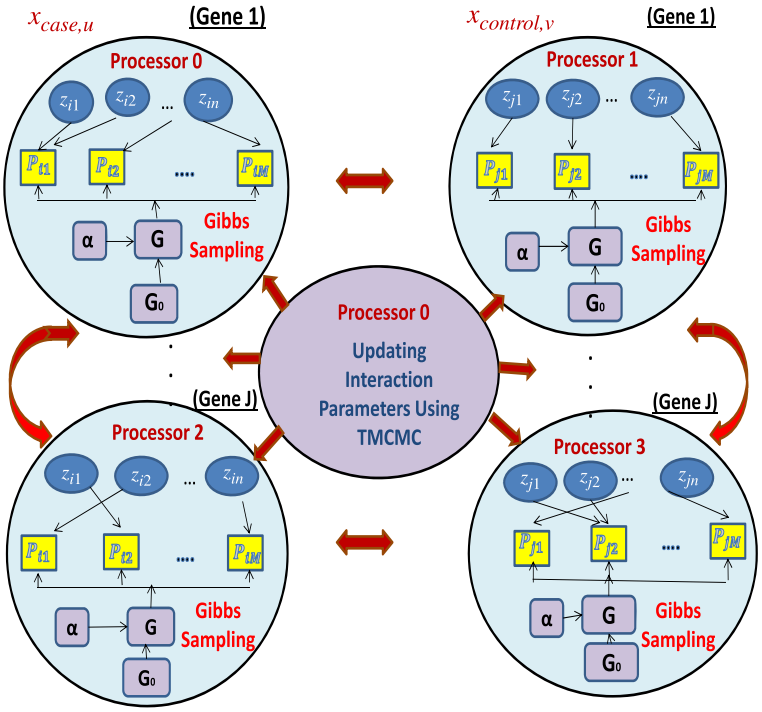

Figure 1 *Schematic diagram for our model and parallel processing idea*: *The arrows in the diagram represent dependence between the variables. The ranks of the processors updating the sets of parameters in parallel using Gibbs sampling are also shown. Once the other parameters are updated in parallel, the interaction parameters are updated using TMCMC by the processor with rank zero.*

where, for $k = 0, 1$, i_k is the index such that $\boldsymbol{P}_{Mi_kjk} = \{\boldsymbol{p}_{1i_kjk}, \boldsymbol{p}_{2i_kjk}, \ldots, \boldsymbol{p}_{Mi_kjk}\}$ is some measure of central tendency of $\{\boldsymbol{P}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk}\}; i = 1, \ldots, N_k\}$. Appropriate measures of central tendency, based on clusterings, is discussed in Section 3.2, with details in Section S-3.

Thus, h_{0j} and h_{1j} are the mixture distributions of the genotype of gene j associated with control and case, respectively, with $1/M$ being the component-wise mixing probabilities (recall from Section 2.2 that we had set $\pi_{mijk} = 1/M$ for all m, i, j, k). If gene j is not responsible for the case control status, then we must have $h_{0j} = h_{1j}$, else, $h_{0j} \neq h_{1j}$. Formally, to ascertain if the J genes under consideration have any effect on the case-control status, it is pertinent to test

$$H_{01} : h_{0j} = h_{1j}; \qquad j = 1, \ldots, J, \tag{3.3}$$

versus

$$H_{11} : \text{not } H_{01}. \tag{3.4}$$

To investigate the effects of environment and gene–gene interactions we shall also test, for $\ell = 1, \ldots, D$; $j = 1, \ldots, J$, and $k = 0, 1$:

$$H_{02} : \beta_{\ell jk} = 0 \quad \text{versus} \quad H_{12} : \beta_{\ell jk} \neq 0, \tag{3.5}$$

and

$$H_{03} : \phi = 0 \quad \text{versus} \quad H_{13} : \phi \neq 0. \tag{3.6}$$

The cases that can possibly arise and the respective conclusions are the following:

- If $\max_{1\le j\le J} d(h_{0j}, h_{1j})$ is significantly small with high posterior probability, then H_{01} is to be accepted. If h_{0j} and h_{1j} are not significantly different, then it is plausible to conclude that the jth gene is not marginally significant in the case-control study.
- Suppose that H_{01} is accepted (so that genes have no significant role) and that $\beta_{\ell jk}$ is significant, at least for some ℓ, j and k, but ϕ is insignificant. This may be interpreted as the environmental variable $\boldsymbol{E}$ having some altering effect on the jth gene, that doesn't affect the disease status. If ϕ turns out to be significant, then this would additionally imply that the environmental variable $\boldsymbol{E}$ influences gene–gene interaction, but not in a way that causes the disease.
- If H_{01} is rejected, indicating that the genes have significant roles to play in causing the disease, but none of the $\beta_{\ell jk}$ or ϕ turn out to be significant, then only genes, not $\boldsymbol{E}$, are responsible for causing the disease. In that case, the disease may be thought to be of purely genetic in nature.
- Suppose H_{01} is rejected, $\beta_{\ell j0}$ and $\beta_{\ell j1}$ turn out to be significant, but that $H_{0\ell j} : \beta_{\ell j0} = \beta_{\ell j1}$ is accepted.Then although $\boldsymbol{E}$ is insignificant with respect to the marginal effect of gene j, it affects the disease status by triggering gene–gene interaction in some genes if ϕ turns out to be significant.
- If H_{01} is rejected, $\beta_{\ell jk}$ is significant for some ℓ, j, k, and ϕ is insignificant, then the presence of $\boldsymbol{E}$ has altering effect on some genes, which, in turn, cause the disease. In this case, since ϕ is insignificant, $\boldsymbol{E}$ does not seem to influence gene–gene interaction.
- If H_{01} is rejected, $\beta_{\ell jk}$ is insignificant for all ℓ, j, k, but ϕ is significant, then significant effect of $\boldsymbol{E}$ on altering the marginal effect of genes is to be ruled out, and one may conclude that the underlying cause of the disease is gene–gene interaction, which has been adversely affected by the environmental variable.
- If H_{01} is rejected, $\beta_{\ell jk}$ is significant for some ℓ, j, k, and ϕ is also significant, then the environmental variable has possibly significantly affected both the marginal and also gene–gene interaction adversely to cause the disease.

3.2 Hypothesis testing based on clustering modes

For $k = 0, 1$, let i_k denote the index of the "central" clusterings of $\boldsymbol{P}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk}\}$, $i = 1, \ldots, N_k$. The concept of central clustering has been introduced by Mukhopadhyay, Bhattacharya and Dihidar (2011). Significant divergence between the two clusterings of $\boldsymbol{P}_{Mi_0jk=0} = \{\boldsymbol{p}_{1i_0jk=0}, \boldsymbol{p}_{2i_0jk=0}, \ldots, \boldsymbol{p}_{Mi_0jk=0}\}$ and $\boldsymbol{P}_{Mi_1jk=1} = \{\boldsymbol{p}_{1i_1jk=1}, \boldsymbol{p}_{2i_1jk=1}, \ldots, \boldsymbol{p}_{Mi_1jk=1}\}$, for $j = 1, \ldots, J$. clearly indicates that the jth gene is marginally significant. Once i_0 and i_1 are determined, we shall consider the clustering distance between $\boldsymbol{P}_{Mi_0jk=0}$ and $\boldsymbol{P}_{Mi_1jk=1}$, denoted by $\hat{d}(\boldsymbol{P}_{Mi_0jk=0}, \boldsymbol{P}_{Mi_1jk=1})$, as a suitable measure of divergence. We shall be particularly interested in

$$d^* = \max_{1\le j\le J} \hat{d}(\boldsymbol{P}_{Mi_0jk=0}, \boldsymbol{P}_{Mi_1jk=1}). \tag{3.7}$$

In Section S-3 of the supplement we include a brief discussion of the aforementioned methodology.

BB point out that although significantly large divergence between clusterings indicate rejection of the null hypothesis, insignificant clustering distance need not necessarily provide strong enough evidence in favour of the null. In other words, even if the clustering distance is insignificant, it is important to check if the parameter vectors being compared are significantly different. In this regard, BB propose an appropriate divergence measure based on Euclidean distances of the logit transformations of the minor allele frequencies. The necessary ideas in our current context are discussed in Section S-3.1 of the supplement. In our case, in order to compute the Euclidean distance, we first compute the averages $\bar{p}_{mijk} = \sum_{r=1}^{L_j} p_{m,ijkr}/L_j$,

then consider their logit transformations $\text{logit}(\bar{p}_{mijk}) = \log\{\bar{p}_{mijk}/(1-\bar{p}_{mijk})\}$. Then, we compute the Euclidean distance between the vectors

$$\text{logit}(\bar{\boldsymbol{P}}_{Mi_0jk=0}) = \{\text{logit}(\bar{p}_{1i_0jk=0}), \text{logit}(\bar{p}_{2i_0jk=0}), \ldots, \text{logit}(\bar{p}_{Mi_0jk=0})\}$$

and

$$\text{logit}(\bar{\boldsymbol{P}}_{Mi_1jk=1}) = \{\text{logit}(\bar{p}_{1i_1jk=1}), \text{logit}(\bar{p}_{2i_1jk=1}), \ldots, \text{logit}(\bar{p}_{Mi_1jk=1})\}.$$

We denote the Euclidean distance associated with the jth gene by

$$d_{E,j} = d_{E,j}(\text{logit}(\bar{\boldsymbol{P}}_{Mi_0jk=0}), \text{logit}(\bar{\boldsymbol{P}}_{Mi_1jk=1})),$$

and denote $\max_{1\le j\le J} d_{E,j}$ by d_E^*.

3.3 Formal Bayesian hypothesis testing procedure integrating the above developments

In our problem, we need to test the following for reasonably small choices of ε's:

$$H_{0,d^*} : d^* < \varepsilon_{d^*} \quad \text{versus} \quad H_{1,d^*} : d^* \ge \varepsilon_{d^*}; \tag{3.8}$$

$$H_{0,d_E^*} : d_E^* < \varepsilon_{d_E^*} \quad \text{versus} \quad H_{1,d_E^*} : d_E^* \ge \varepsilon_{d_E^*}; \tag{3.9}$$

$$H_{0,\beta_{\ell jk}} : |\beta_{\ell jk}| < \varepsilon_{\ell jk} \quad \text{versus} \quad H_{1,\beta_{\ell jk}} : |\beta_{\ell jk}| \ge \varepsilon_{\ell jk}, \quad \text{for } \ell = 1, \ldots, D;\ j = 1, \ldots, J;\ k = 0, 1; \tag{3.10}$$

$$H_{0,\phi} : \phi < \varepsilon_{\phi} \quad \text{versus} \quad H_{1,\phi} : \phi \ge \varepsilon_{\phi}. \tag{3.11}$$

If H_0 is rejected in (3.8) or in (3.9), we could also test if the jth gene is influential by testing, for $j = 1, \ldots, J$, $H_{0,\hat{d}_j} : \hat{d}_j < \varepsilon_{\hat{d}_j}$ versus $H_{1,\hat{d}_j} : \hat{d}_j \ge \varepsilon_{\hat{d}_j}$, where $\hat{d}_j = \hat{d}(\boldsymbol{P}_{Mi_0jk=0}, \boldsymbol{P}_{Mi_1jk=0})$; we could also test $H_{0,d_{E,j}} : d_{E,j} < \varepsilon_{d_{E,j}}$ versus $H_{1,d_{E,j}} : d_{E,j} \ge \varepsilon_{d_{E,j}}$.

To test if gene–gene interactions are significant, one may test, following BB, $H_{0,j,j^*} : |\boldsymbol{A}_{jj^*}| < \varepsilon_{A_{jj^*}}$ versus $H_{1,j,j^*} : |\boldsymbol{A}_{jj^*}| \ge \varepsilon_{A_{jj^*}}$, for $j^* \ne j$, $\boldsymbol{A}_{jj^*}$ being the (j, j^*)th element of $\boldsymbol{A}$. If H_{1,j,j^*} is accepted for some (or many) $j^* \ne j$, then this would indicate significant interaction between the j^*th and the jth genes.

As argued in BB, here also it is easily seen that our testing procedure is equivalent to Bayesian multiple testing procedures that minimize the Bayes risk of additive "0-1" and "0-1-c" loss functions (see BB for the details; see also Berger (1985)). Since it is well-known that Bayesian multiple testing methods automatically provide multiplicity control through the inherent hierarchy (see, for example, Scott and Berger (2010)), separate error control is not necessary. A brief, schematic representation of the hierarchy of the hypothesis tests is shown in Figure 2.

Our choices of the ε's are based on the idea of null model introduced in BB. In a nutshell, we first specify an appropriate null model, which, for example, is the same model as ours but with $\boldsymbol{A}$ and $\tilde{\boldsymbol{\Sigma}}$ set to identity matrices to reflect the null hypotheses of "no interaction" and the same mixture distributions under cases and controls for each gene for no genetic effect. From the null model thus specified, we then generate case-control genotype data and fit our general Bayesian model to this "null data" and set ε to be the 55th percentile of the relevant posterior distribution. The rationale and details of this procedure are provided in BB (particularly in Section S-7 of their supplement).

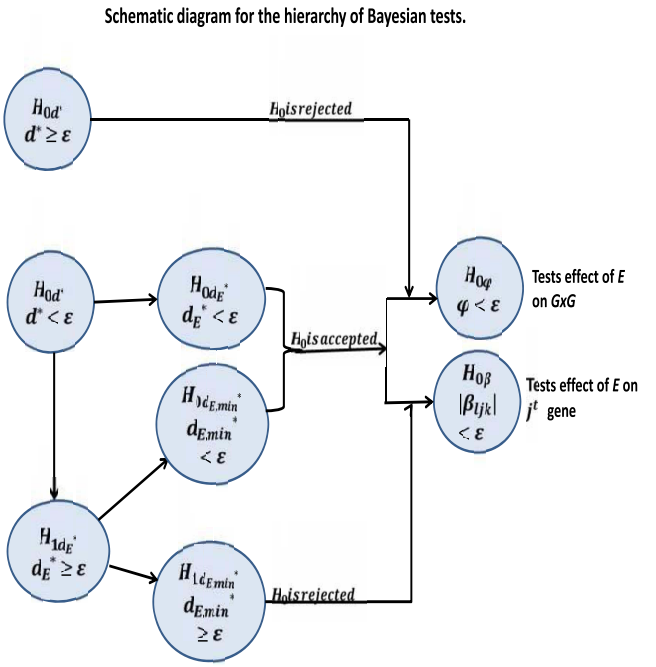

Figure 2 *Schematic diagram for our Bayesian testing idea.*

4 Simulation studies

For simulation studies, we first generate biologically realistic genotype data sets under stratified population with known G × G and G × E set ups from the GENS2 software of Pinelli et al. (2012). We consider simulation studies in 5 different true model set-ups with two genes, one environmental variable and 5 sub-populations: (a) presence of gene–gene and gene–environment interaction, (b) absence of genetic or environmental effect, (c) absence of genetic and gene–gene interaction effects but presence of environmental effect, (d) presence of genetic and gene–gene interaction effects but absence of environmental effect, and (e) independent and additive genetic and environmental effects.

As we demonstrate, our model and methodologies successfully identify the marginal effects of the genes, along with the G × G and G × E, and the number of sub-populations. Here we briefly summarize the results of our experiments while the details are provided in Section S-4 of the supplement.

In case (a), both the clustering and the Euclidean metric suggest significant effects of both the genes. Significant interaction between the two genes is also suggested by our test regarding $\boldsymbol{A}_{12}$. In the true, data-generating model, although the environmental variable exerts significant marginal effect on the genes, it does not influence gene–gene interaction. In keeping with this, our test on ϕ led to acceptance of the null hypothesis of no significance. Significance marginal effect of the environmental variable on gene 1 is borne out by the test on $\beta_{\ell jk}$'s, as β_{111} turned out to be significant.

In case (b) there is no gene–gene or gene-environmental interaction effect. As elucidated in Section S-4.2, the test based on the Euclidean metric, which turned out to be more appropriate than the clustering metric in this case, rightly indicated insignificance of the genetic effects. This of course leads to the conclusion that the environmental variable can not have any negative impact on the genes to trigger the disease. However, our Bayesian nonparametric model can not be used to test if the environmental variable directly affects the case-control status, without affecting the genes; see Section S-4.2 for the explanation. Hence, given that

genetic effects are absent, we used the logistic regression to test the same, and obtained clear acceptance of the null hypothesis of no significance of the environmental variable. In other words, our results are in accordance with the true scenario.

Case (c) is a continuation of case (b) with the true model dictating the effect of the environmental variable on case-control, although no genetic effects are present. The test based on logistic regression suggested no environmental effect on gene–gene interaction. Thus, setting $\phi = 0$ in our Bayesian nonparametric model, we tested for the effects of the genes, obtaining results that clearly suggested insignificance of the genetic effects. The tests with $\beta_{\ell jk}$ demonstrated no environmental effect on the individual genes. However, the best logistic regression model based on the Akaike Information Criterion (AIC) consists of the environmental variable. On the basis of this and the result that all the genes insignificant, we conclude that the environmental variable is the only factor responsible for the case-control status. Thus, our results are consistent with the true scenario.

In case (d), our prescribed tests easily identified that the genetic effects and the gene–gene interaction are significant, and that the environmental variable has no effect on the individual genes or gene–gene interaction. We used test based on logistic regression as before to reach the conclusion that the environmental effect has no effect the case-control outcome.

In the final simulation study case (e), note that our Bayesian model does not support the assumption of additive genetic and environmental effects and hence is not expected to perform well under this case. Resorting eventually to logistic model, we obtained the AIC-based best model that consists of the additive marginal effects of the first gene and the environmental variable, along with an additive intercept. This is broadly in keeping with the data-generating mechanism. We find that with respect to our Bayesian model the additive effect has been wholly transformed into the environmental effect, and that the environmental variable is much more influential compared to the genes in the sense of directly affecting case-control status without affecting the genes.

Note that it is very important to identify the so-called disease predisposing loci (DPL), which are the SNPs that are responsible for influencing the risk of the disease. In cases (a), (d) and (e), where genes play significant roles, the DPL for both the genes have been identified with precision by our model and methods, in spite of the highly complex dependent structure induced by the gene–gene and gene–environment interactions. Furthermore, in all the cases (a)–(e), the true number of sub-populations are correctly identified. Thus, our model and methods perform quite encouragingly.

5 Application of our model and methodologies to a real, case-control dataset on myocardial infarction

MI (more commonly, heart attack), has been subjected to much investigation for detecting the underlying genetic causes, the possible environmental factors and their interactions. Application of our ideas to a case-control genotype dataset on early-onset of myocardial infarction (MI) from MI Gen study, obtained from the dbGaP database (http://www.ncbi.nlm.nih.gov/gap), led to some interesting insights into gene–environment and gene–gene interactions on incorporating sex as the environmental factor.

5.1 Data description

The MI Gen data obtained from dbGaP broadly represents a mixture of four sub-populations: Caucasian, Han Chinese, Japanese and Yoruban. For our analysis, we considered a set of SNPs that are found to be individually associated with different cardiovascular end points

like LDL cholesterol, smoking, blood pressure, body mass etc. in various GWA studies published in NHGRI catalogue and augmented this set further with another set of SNPs found to be marginally associated with MI in the MIGen study (see Lucas et al. (2012)). Our study also includes SNPs that are reported to be associated with MI in various other studies; see Erdmann, Linsel-Nitschke and Schunkert (2010), Qi et al. (2011) and Wang et al. (2004). In all, we obtained 271 SNPs. Unfortunately, only 33 of them turned out to be common to the SNPs of our original MI dataset on genotypes, which has been mapped on to the genes using the Ensembl human genome database. However, we included in our study all the SNPs associated with the genes containing the 33 common SNPs. Specifically, our study involves the genotypic information on 32 genes covering 1251 loci, including the 33 previously identified loci for 200 individuals. We chose this relatively small number of individuals to ensure computational feasibility. However, even this data set, along with our model and prior, yielded results that are not only compatible with, but also complement the results established in the literature.

Categorization of the case-control genotype data into the four sub-populations, each of which are likely to represent several further and rather varied sub-populations genetically, implies that the maximum number of mixture components must be fixed at some value much higher than 4. As before, we set $M = 30$ and $\alpha_{jk} = 10$ for every (j, k), to facilitate data-driven inference.

We chose a similar set-up for the null model. That is, we chose the same number of genes and the same number of loci for each gene, the same number of cases and controls, the same value $M = 30$, but $\alpha_{jk} = 1.5$ for every (j, k), as in our simulation studies. We use the same priors as in the real data set-up except that we set $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ to be identity matrices to ensure that the genetic interaction is not present and set the same mixture distribution under cases and controls for each gene to ensure the absence of genetic effects.

5.2 Remarks on incorporation of the sex variable in our model

In our case, $\boldsymbol{E}_i = E_i$, a one-dimensional binary variable, where $E_i = 1$ if the ith individual is male and $E_i = 0$ if female. Hence, $\boldsymbol{\beta}_{jk} = \beta_{jk}$ is a scalar quantity. In (2.9) and (2.10) we considered the environmental variable to be continuous, but remarked that the model can be easily extended to include categorical variables. Indeed, in this case the exponentials of (2.9) and (2.10) can be thought of as binary regressions with sex as the covariate.

As regards $\mathcal{E}_{ij}$ of (2.12), we first consider $a_0 + a_1 E_i$ as a binary regression, and then write

$$\mathcal{E}_{ij} = \exp\bigl(-\|(a_0 + a_1 E_i) - (a_0 + a_1 E_j)\|^2\bigr) = \exp\bigl[-a_1^2 (E_i - E_j)^2\bigr], \tag{5.1}$$

with $b = a_1^2$ being the smoothness parameter. Observe that for the same sex, $\mathcal{E}_{ij} = 1$ while for different sex, $\mathcal{E}_{ij} = \exp(-b) < 1$.

5.3 Remarks on model implementation

We first obtain the number of parameters to be updated by TMCMC in our case; other unknowns associated with the mixtures, to be updated using Gibbs steps in parallel. Note that in our case, the interaction matrix $\boldsymbol{A}$ is of order $32 \times 32 = 1024$, and the associated Cholesky decomposition then consists of $33 \times 16 = 528$ parameters. Also, $\boldsymbol{\lambda}$ is a $NJ = 200 \times 2 = 400$-dimensional random vector and $\boldsymbol{\Sigma}$ is of order $N \times N = 200 \times 200$, so that its Cholesky decomposition consists of $201 \times 100 = 20100$ parameters. Furthermore, $\{(u_r, v_r) : r = 1, \ldots, L\}$, where $L = 207$, consists of $2 \times 207 = 414$ parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ consist of 64 parameters each, and there are two more parameters b and ϕ. So, in all, there are 21,572 parameters to be updated simultaneously in a single block using TMCMC.

We implemented our parallel MCMC algorithm detailed in S-2 of the supplement on a VMware consisting of 50 double-threaded, 64-bit physical cores, each running at 2493.990 MHz. In spite of the large number of parameters associated with the interaction part, our mixture of additive and additive-multiplicative TMCMC still ensured reasonable performance.

Our parallel MCMC algorithm takes about 11 days to yield 100,000 iterations in our aforementioned VMware machine. We discard the first 50,000 iterations as burn-in. Informal convergence diagnostics such as trace plots exhibited adequate mixing properties of our parallel algorithm.

5.4 Results of the real data analysis

5.4.1 *Effect of the sex variable.* It turned out that $\varepsilon_\phi = 1.043069$ and $P(\phi < \varepsilon_\phi|\text{Data}) \approx 1$, so that ϕ is clearly insignificant, indicating no differential effect of sex on the genetic interactions. The posterior probabilities $P(|\beta_{1j1} - \beta_{1j0}| < \varepsilon|\text{Data})$ are shown in Figure 3. As before, ε is the 55th percentile of the posterior distribution of $|\beta_{1j1} - \beta_{1j0}|$ under the null model. Under the 0-1 loss function, the above posterior probability exceeding 0.5 indicates significant environmental effect on the jth gene. From the figure it is interesting to note that there is significant differential effect due to sex on the marginal effects of several genes although sex does not affect the genetic interactions significantly.

5.4.2 *Influence of genes and gene–gene interactions on MI based on our study.* Our Bayesian hypotheses testing using the clustering metric yielded $P(d^* < \epsilon_1|\text{Data}) \approx 0.35202$ while that with the Euclidean distance we obtained $P(d_E^* < \epsilon_2|\text{Data}) \approx 0.51078$. In other words, it seems rather debatable whether or not the genes have significant overall effect on MI. This is in sharp contrast with the results obtained by BB where both clustering metric and Euclidean distance confirmed significant overall genetic influence on MI. However, both the posterior probabilities are substantially large, practically indicating that the genes are not very significant.

As far as testing of significance of the individual genes are concerned, it turned out that under the clustering metric, except genes *SMARCA*4, *RBMS*1, *COL*4*A*1, *RP*11-306*G*20.1, *MRAS*, *SLC*22*A*1, *CDKAL*1, *PCSK*9, *ADAMTS*9-*AS*2, and *AP*006216.5, the rest turned out to be significant, while with respect to the Euclidean metric the only insignificant genes are *AP*006216.10, *CELSR*2, *MRAS*, *PCSK*9, *OR*4*A*48*P* and *BUD*13. The posterior probabilities of the null hypotheses (of no significant genetic influence) are shown in Figure S-6 of the supplement. The figure reveals that the posterior probabilities of no significant genetic influence,

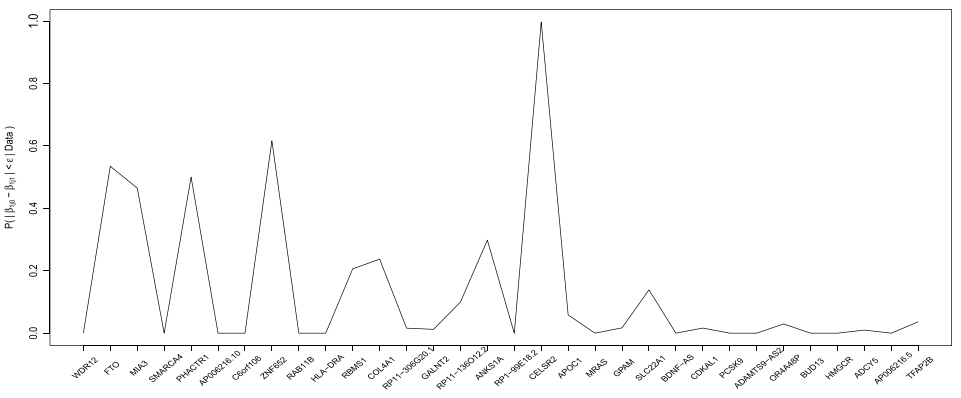

Figure 3 *Index plots of posterior probabilities of no environmental effect with respect to* $|\beta_{1j0} - \beta_{1j1}| < \varepsilon$, *for* $j = 1, \ldots, 32$.

although generally did not cross 0.5, are not adequately small to reflect very strong evidence against the null hypotheses. This is consistent with the result on overall genetic significance that we obtained.

The actual gene–gene correlations based on medians of the posterior covariances, are shown in Figure S-7 of the supplement. The color intensities correspond to the absolute values of the correlations. Consistent with the figure, all the tests on interaction turned out to support the hypotheses of no interaction.

Thus, individual genes have impact on MI but not gene–gene interactions. Moreover, the relatively weak evidences against the null suggest that external factors, in our case sex, may be playing a bigger role in explaining case-control with respect to MI. As such, given our data set of size 200 with 77 cases, the empirical conditional probability of a male given case is 0.3766234, while the empirical conditional probability of a male given control is 0.504065, indicating that with respect to our data, females seem to be more at risk compared to males. Coherency of Bayesian models in general is instrumental in reflecting this information in our inference in the way of downplaying the genes, suggesting at the same time that the only external factor, namely, sex, must have more important effect.

A detailed investigation of the DPL detected by our model and methods, and the role of SNP-SNP interactions behind such DPL, is carried out in Section S-5 of the supplement, and a discussion on the posterior distribution of the number of distinct mixture components is provided in Section S-6 of the supplement.

5.5 Discussion of our Bayesian methods and GWAS in light of our findings

Our results of Bayesian analysis of the MI data set demonstrate that sex plays more significant role than the genes in triggering the disease, and in particular, do not support gene–gene interaction. In these regards, our results significantly differ from those obtained by BB, who do not consider the sex variable in their model. Since as per our inference sex seems to be far more influential compared to the genes with respect to MI, there is internal consistency of our more general gene–gene and gene–environment interaction model with the gene–gene interaction model of BB. It is important to note that Lucas et al. (2012) analyzed the same MI dataset using logistic regression and reached the same conclusion as ours that there is no significant gene–gene interaction. Since two completely different methods of analyses are in such strong agreement, it is pertinent to presume that the data contains enough information on the lack of gene–gene interaction. However, as we demonstrated, SNP-SNP correlations have important roles to play in determining the DPL. These are responsible for suppression of the SNPs considered influential in the literature by implicit induction of negative correlations between Euclidean distances between cases and controls for the associated SNPs. Thus, even though the genes did not turn out to be as significant, it is clear that sophisticated nonparametric modeling of gene–gene and SNP-SNP interactions is of utmost importance.

Finally, since in GWA studies are usually conducted by testing one SNP at a time, it is important to clarify if the same is permissible with our Bayesian nonparametric model while accounting for gene–gene and gene–environment interactions. We assert that this is indeed the case. To elucidate, observe that our method of DPL detection, which uses the Euclidean distance between SNP-wise case and control, can be easily formalized to create a SNP-wise Bayesian hypothesis testing problem, the null hypothesis being that the SNP-wise Euclidean distance is below some sufficiently small threshold. The Bayesian testing procedure is akin to the distance based testing approach discussed in Section 3.3.

6 Summary and conclusion

In this paper, we have extended the Bayesian semiparametric gene–gene interaction model of BB to realistically include the case of gene–environment interactions. Careful attention

has been paid to the fact that in the absence of mutation, the environmental variable does not affect the marginal genotypic distributions, in spite of influencing gene–gene interaction. Needless to mention, our model considers dependence between SNPs as well to account for LD effects, in addition to gene–gene, gene–environment and dependencies between individuals. Besides, our model, via DP, facilitates learning about the number of genotypic sub-populations associated with the individuals and the genes, while accounting for the environmental effect at the same time.

We extend the Bayesian hypotheses testing methods introduced in BB to enable test for significances of marginal genetic and environmental effects, gene–gene interactions, effect of environment on gene–gene interaction and mutational effect. The basis for our tests are extensions of the clustering metric based tests proposed by BB to account for the environmental variables, in conjunction with the tests based on Euclidean metric. We recommended careful application of our tests based on the clustering metric, followed by re-confirmation with respect to the Euclidean metric.

On the Bayesian computational side, we propose a powerful parallel processing algorithm that takes advantage of the conditional independence structures built within our model through the DP based mixture framework for parallelisation, and is complemented by the efficiency of TTMCMC, which updates the interaction parameters within a single processor.

We validate our model and methodologies with applications to biologically realistic datasets generated from under 5 different set-ups characterized by different combinations and structures associated with gene–gene and gene–environment interactions. Adequate performance of our model and methods are demonstrated in every situation. Additionally, our ideas correctly captured the true number of genetic sub-populations in each case, and attempted to capture the DPL adequately even in the face of highly complex dependence structures.

We apply our model and methods to the MI Gen data set also studied by BB and because of inclusion of the sex variable, succeeded in obtaining results that are quite compatible with those reported in the literature. Although the gene–gene interactions turned out to be insignificant, the SNP-SNP correlations associated with case-control Euclidean distances facilitated understanding the mismatch of our DPL with those reported in the literature as having significant impact on MI. Interestingly, our Bayesian approach allowed us obtain insightful results even with a sample consisting of only 200 individuals, showing the importance of building sophisticated models and prior structures, and efficient computational methods and technologies.

Acknowledgments

We are sincerely grateful to the Editor-in-Chief and the referee whose comments and suggestions have led to substantial improvement of our manuscript.

Supplementary Material

Supplement to "Effects of gene–environment and gene–gene interactions in case-control studies: A novel Bayesian semiparametric approach" (DOI: 10.1214/18-BJPS413SUPP; .pdf). In Section S-1 we provide further discussion regarding the effects of environment on gene-gene interactions. In Section S-2 we detail a parallel MCMC algorithm for implementing our Bayesian model, and in Section S-3 we include a discussion of clustering metric, clustering mode and Euclidean distance based divergence measures. We present the details of our simulation studies in Section S-4 and in Section S-5 we explain at length the roles of the disease predisposing loci detected by our Bayesian analysis of the real, myocardial infarction data. With respect to the real data, we present and analyze the posterior distributions of the number of distinct components in Section S-6.

References

Ahn, J., Mukherjee, B., Ghosh, M. and Gruber, S. B. (2013). Bayesian semiparametric analysis of two-phase studies of gene–environment interaction. *Annals of Applied Statistics* **7**, 543–569. MR3086430 https://doi.org/10.1214/12-AOAS599

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer. MR0804611 https://doi.org/10.1007/978-1-4757-4286-2

Bhattacharjee, S., Wang, Z., Ciampa, J., Kraft, P., Chanock, S., Yu, K. and Chatterjee, N. (2010). Using principal components of genetic variation for robust and powerful detection of gene–gene interactions in case-control and case-only studies. *American Journal of Human Genetics* **86**, 331–342.

Bhattacharya, D. and Bhattacharya, S. (2016). A Bayesian semiparametric approach to learning about gene–gene interactions in case-control studies. Preprint. Available at http://arxiv.org/abs/1411.7571. MR3860648 https://doi.org/10.1080/02664763.2018.1444741

Bhattacharya, D. and Bhattacharya, S. (2020). Supplement to "Effects of gene–environment and gene–gene interactions in case-control studies: A novel Bayesian semiparametric approach." https://doi.org/10.1214/18-BJPS413SUPP.

Brockwell, A. (2006). Parallel Markov chain Monte Carlo simulation by pre-fetching. *Journal of Computational and Graphical Statistics* **15**, 246–261. MR2269370 https://doi.org/10.1198/106186006X100579

Calderhead, B. (2014). A general construction for parallelizing Metropolis–Hastings algorithms. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 17408–17413.

Chen, Y., Freitas, N. D., Eskelin, M., Fang, J. and Welling, M. (2016). Herded Gibbs sampling. *Journal of Machine Learning Research* **17**, 263–291. MR3491104

De Iorio, M., Favaro, S. and Teh, Y. W. (2015). Bayesian inference on population structure: From parametric to nonparametric modeling. In *Nonparametric Bayesian Inference in Biostatistics*, 135–151. Cham: Springer. MR3411018

Dutta, S. and Bhattacharya, S. (2014). Markov chain Monte Carlo based on deterministic transformations. *Statistical Methodology* **16**, 100–116. Also available at http://arxiv.org/abs/1106.5850. Supplement available at http://arxiv.org/abs/1306.6684.

Erdmann, J., Linsel-Nitschke, P. and Schunkert, H. (2010). Genetic causes of myocardial infarction. *Deutsches Ärzteblatt International* **107**, 694–699.

Hunter, D. J. (2005). Gene environment interactions in human diseases. *Nature Reviews Genetics* **6**, 287–298.

Jacob, P., Robert, C. P. and Smith, M. H. (2011). Using parallel computation to improve independent Metropolis–Hastings based estimation. *Journal of Computational and Graphical Statistics* **3**, 616–635. MR2878993 https://doi.org/10.1198/jcgs.2011.10167

Khouri, M. J. (2005). Do we need genomic research for the prevention of common diseases with environmental causes? *American Journal of Epidemiology* **161**, 799–805.

Ko, Y.-A., Saha Chaudhuri, P., Vokonas, P. S., Park, S. K. and Mukherjee, B. (2013). Likelihood ratio tests for detecting gene environment interaction in longitudinal studies. *Genetic Epidemiology* **37**, 581–591.

Lucas, G., Lluis-Ganella, C., Subirana, I., Masameh, M. D. and Gonzalez, J. R. (2012). Hypothesis-based analysis of gene–gene interaction and risk of myocardial infraction. *PLoS ONE* **7**, e41730.

Majumdar, A., Bhattacharya, S., Basu, A. and Ghosh, S. (2013). A novel Bayesian semiparametric algorithm for inferring population structure and adjusting for case-control association tests. *Biometrics* **69**, 164–173. MR3058063 https://doi.org/10.1111/biom.12004

Mapp, C. (2003). The role of genetic factors in occupational asthma. *European Respiratory Journal* **21**, 173–178.

Martino, L., Elvira, V. and Camps-Valls, G. (2018). The recycling Gibbs sampler for efficient learning. *Digital Signal Processing* **74**, 1–13. MR3754555 https://doi.org/10.1016/j.dsp.2017.11.012

Martino, L., Elvira, V., Luengo, D., Corander, J. and Louzada, F. (2016). Orthogonal parallel MCMC methods for sampling and optimization. *Digital Signal Processing* **58**, 64–84.

Mather, K. and Caligary, P. (1976). Genotype x environmental interactions. *Heredity* **36**, 41–48.

Mukherjee, B., Ahn, J., Gruber, S. B. and Chatterjee, N. (2012). Testing gene environment interaction in large-scale association studies. *American Journal of Epidemiology* **175**, 177–190.

Mukherjee, B., Ahn, J., Gruber, S. B., Ghosh, M. and Chatterjee, N. (2010). Bayesian sample size determination for case-control studies of gene–environment interaction. *Biometrics* **66**, 934–948. MR2758230 https://doi.org/10.1111/j.1541-0420.2009.01357.x

Mukherjee, B., Ahn, J., Gruber, S. B., Moreno, V. and Chatterjee, N. (2008). Testing gene–environment interaction from case-control data: A novel study of type-I error, power and designs. *Genetic Epidemiology* **32**, 615–626.

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. MR2526617 https://doi.org/10.1111/j.1541-0420.2007.00953.x

Mukhopadhyay, S., Bhattacharya, S. and Dihidar, K. (2011). On Bayesian "central clustering": Application to landscape classification of Western Ghats. *Annals of Applied Statistics* **5**, 1948–1977. MR2884928 https://doi.org/10.1214/11-AOAS454

Ottman, R. (2010). Gene environment interactions: Definitions and study designs. *Pubmed* **6**, 764–770.

Pinelli, M., Scala, G., Amato, R., Cocozza, S. and Miele, G. (2012). Simulating gene–gene and gene–environment interactions in complex diseases: Gene–environment iNteraction simulator 2. *BMC Bioinformatics* **13**, 132.

Purcell, S. (2002). Variance components models for gene–environment interaction in twin analysis. *Twin Research* **5**, 554–571.

Qi, L., Ma, J., Qi, Q., Hartiala, J., Allayee, H. and Campos, H. (2011). Genetic risk score and risk of myocardial infarction in hispanics. *Circulation* **123**, 374–380.

Sanchez, B., Kang, S. and Mukherjee, B. (2012). A latent variable approach to study of gene–environment interactions in the presence of multiple correlated exposures. *Biometrics* **68**, 466–476. MR2959613 https://doi.org/10.1111/j.1541-0420.2011.01677.x

Scott, G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38**, 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

Scott, S. A. (2011). Personalizing medicine with clinical pharmacogenetics. *Genetics in Medicine* **13**, 987–995.

Wang, Q., Rao, S., Shen, G.-Q., Li, L., Moliterno, D. J., Newby, L. K., Rogers, W. J., Cannata, R., Zirzow, E., Elston, R. C. and Topol, E. J. (2004). Premature myocardial infarction novel susceptibility locus on chromosome 1P34-36 identified by genomewide linkage analysis. *Circulation* **74**, 262–271.

Wang, X., Elston, R. C. and Zhu, X. (2010). The meaning of interaction. *Human Heredity* **70**, 269–277.

Wright, A. F., Carothers, A. D. and Campbell, H. (2002). Gene–environment interactions—The BioBank UK study. *Pharmacogenomics Journal* **2**, 75–82.

Interdisciplinary Statistical Research Unit
Indian Statistical Institute
203, B. T. Road
Kolkata 700108
India
E-mail: sourabh@isical.ac.in