Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 327–360
http://www.ssca.org.in/journal

Gene-Gene and Gene-Environment Interactions in Case-Control Studies Based on Hierarchies of Dirichlet Processes

Durba Bhattacharya[1] **and Sourabh Bhattacharya**[2]
[1]*Department of Statistics, St. Xavier's College (Autonomous), Kolkata*
[2]*Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata*

Received: 29 May 2024; Revised: 30 August 2024; Accepted: 06 September 2024

Abstract

It is becoming increasingly clear that complex interactions among genes and environmental factors play crucial roles in triggering complex diseases. Thus, understanding such interactions is vital, which is possible only through statistical models that adequately account for such intricate, albeit unknown, dependence structures. In this article, we propose and develop a novel nonparametric Bayesian model for case-control genotype data using hierarchies of Dirichlet processes that offers a more realistic and nonparametric dependence structure among the genes, induced by the environmental variables. In this regard, we propose a novel and highly parallelisable MCMC algorithm that is rendered quite efficient by the combination of modern parallel computing technology, effective Gibbs sampling steps, retrospective sampling and Transformation based Markov Chain Monte Carlo (TMCMC). We devise appropriate Bayesian hypothesis testing procedures to detect the roles of genes and environment in case-control studies. Applying our ideas to 5 biologically realistic case-control genotype datasets simulated under distinct set-ups, we obtain encouraging results in each case. We finally apply our ideas to a real, myocardial infarction dataset, and obtain interesting results on gene-gene and gene environment interaction, that broadly agree with the results reported in the literature, but provide further important insights.

Key words: Case-control study; Hierarchical Dirichlet process; Gene-gene and gene-environment interaction; Myocardial Infarction; Parallel processing; Transformation based MCMC.

1. Introduction

In spite of much research on gene-gene interaction, including genome-wide association studies (GWAS), it has become increasingly clear that gene-gene interaction alone is insufficient for explaining most complex diseases. Investigating environmental factors independently of the genetic factors is not sufficient either – biomedical research points towards the importance of interactions between genes and the environment in explaining complex diseases. Indeed, according to Hunter (2005) (see also Mather and Caligary (1976)), considering only the separate contributions of genes and environment to a disease, ignoring their

Corresponding Author: Durba Bhattacharya
Email: durba@sxccal.edu

interactions, will lead to incorrect estimation of the disease proportion (the "population attributable fraction") that is explained by genes, the environment, and their joint effect. In particular, environmental exposures are expected to influence gene-gene interactions of the individuals. A comprehensive overview of gene-environment interaction with various examples is provided in Bhattacharya and Bhattacharya (2020).

Since no simple relationship exists between the genes and environment, it is clear that linear or additive models, as are mostly used so far, are inadequate for modeling gene-environment interactions. Also, the logistic model based approaches, (see for example Ahn *et al.* (2013), Wen and Stephens (2014) and Liu *et al.* (2015)) resting on Fisher's definition of interaction result in the inclusion of a large number of interaction terms even with a moderate number of genetic and environmental factors.

The fact that the genetic data may arise from a stratified population with an unknown number of subpopulations makes the problem all the more demanding. Wen and Stephens (2014), in their attempt to study the genetic association with respect to genetic data arising from multiple potentially-heterogeneous subgroups, assume the number of subgroups to be known in advance. Also, the problem of quantifying the strength of heterogeneity, as acknowledged by Wen and Stephens (2014), remains unanswered due to the above considerations and the need of an appropriate prior. The Bayesian semiparametric model proposed by Bhattacharya and Bhattacharya (2020) takes care of the above mentioned problems by proposing a model based on Dirichlet Processes (DP) and a hierarchical matrix-normal distribution, which encapsulates the mechanism of dependence among genes under environmental effects with respect to genotype data arising out of a possibly stratified population. In a somewhat similar spirit, Urbut *et al.* (2019) and Yang *et al.* (2024) propose mixture of multivariate nornal distributions with appropriate covariance matrices relevant for the phenomenon under study.

We now elaborate on a possible drawback of the dependence structure induced by the modeling strategy of Bhattacharya and Bhattacharya (2020), which motivated us to develop our present work based on Hierarchical Dirichlet Processes.

In their model, the relevant gene-gene covariance matrix for individual i is $\tilde{\sigma}_{ii}\boldsymbol{A}$, where $\boldsymbol{A}$ is the gene-gene interaction matrix common to all the individuals in the absence of environmental variables, and $\tilde{\sigma}_{ii} = \sigma_{ii} + \phi$, with σ_{ii} being the i-th diagonal element of a symmetric, positive definite matrix not associated with the environmental variable, and ϕ is a non-negative parameter, to be interpreted as the effect of the environmental variable $\boldsymbol{E}$ on gene-gene interaction. Note that Bhattacharya and Bhattacharya (2020) assumed that the covariance matrices for all the individuals are affected in the same way by the environmental variable, which seems to be a limitation of the covariance structure. The environmental variables may affect the gene-gene interactions of individuals differently depending on the extent and type of their exposure to the environmental factors.

In this article, we introduce a novel Bayesian nonparametric model for gene-gene and gene-environment interactions for case-control genotype data that solves the issues detailed above. Our model represents the individual genotype data as finite mixtures based on Dirichlet processes as before, but instead of the hierarchical matrix normal distribution, we introduce a hierarchy of Dirichlet processes that create appropriate nonparametric dependence among the genes induced by the environment, case-control dependence, and de-

pendence among the individuals. As we show, our modeling strategy satisfies all the desirable properties, bypassing the drawbacks of the matrix-normal based model of Bhattacharya and Bhattacharya (2020). The key idea of inducing such nonparametric dependence is to ensure that the minor allele frequencies associated with every sub-population, individual, gene and case/control status share a global pool of random parameters, in such a manner that only the dependence structure is influenced by the environmental variables, not the marginal distributions of the minor allele frequencies. The last point is important biologically and so, it requires care to model such intricate dependence.

Although our hierarchical Dirichlet process (HDP) model has parallels with the HDP introduced by Teh *et al.* (2006), our HDP has one more level of hierarchy compared to Teh *et al.* (2006). Moreover, the aforementioned special and intricate dependence structure has not been considered in any previous HDP application.

Exploiting conditional independence structures of our Bayesian model, we develop a novel and highly parallelisable Markov Chain Monte Carlo (MCMC) methodology that combines the efficiencies of modern parallel computing infrastructure, Gibbs steps, retrospective sampling methods, and Transformation based Markov Chain Monte Carlo (TMCMC). For the hypothesis testing procedures, we essentially adopt and extend the ideas provided in Bhattacharya and Bhattacharya (2020). Application of our model and methods to five simulation experiments for the validation purpose yielded quite encouraging results, and application to a real myocardial infarction (MI) case-control type dataset yielded results that are broadly in agreement with the results reported in the literature, but provided new and interesting insights into the mechanisms of gene-gene and gene-environment interactions.

The rest of our paper is structured as follows. We introduce our model in Section 2, and in Section 3 discuss the relevant dependence structures induced by our model. In Section 4 we extend the Bayesian hypothesis testing procedures proposed in Bhattacharya and Bhattacharya (2020) to learn about the roles of genes, environmental variables and their interactions in case-control studies. In Section 5 we briefly discuss the results of application of our model and methodologies to 5 biologically realistic simulated data sets, the details of which are provided in the Annexure, described below. In Section 6 we analyze the real MI dataset using our ideas, demonstrating quite interesting and insightful outcome. Finally, we summarize our work with concluding remarks in Section 7.

Additional details are provided in the Annexure, whose sections have the prefix "A-" when referred to in this paper.

2. A new Bayesian nonparametric model for gene-gene and gene-environme nt interactions

2.1. Case-control genotype data

For $s = 1, 2$ denoting the two chromosomes, let $y^s_{ijkr} = 1$ and $y^s_{ijkr} = 0$ indicate the presence and absence of the minor allele of the i-th individual belonging to the k-th group (either control or case), for $k = 0, 1$, with $k = 1$ denoting case; at the r-th locus of j-th gene, where $i = 1, \ldots, N_k$; $r = 1, \ldots, L_j$ and $j = 1, \ldots, J$; let $N = N_1 + N_2$. Let $\boldsymbol{E}_i$ denote a set of environmental variables associated with the i-th individual. In what follows, we model this case-control genotype and the environmental data using our Bayesian nonparametric model,

described in the next few sections.

2.2. Mixture models based on Dirichlet processes

Let $\boldsymbol{y}_{ijkr} = (y^1_{ijkr}, y^2_{ijkr})$, and if $L = \max\{L_1, \ldots, L_J\}$, let $\boldsymbol{Y}_{ijk} = (\boldsymbol{y}_{ijk1}, \boldsymbol{y}_{ijk2}, \ldots, \boldsymbol{y}_{ijkL_j})$ and $\tilde{\boldsymbol{Y}}_{ijk} = (\tilde{\boldsymbol{y}}_{ijk,L_j+1}, \ldots, \tilde{\boldsymbol{y}}_{ijkL})$, where $\tilde{\boldsymbol{Y}}_{ijk}$ are unobserved and assumed to be missing. We introduce these unobserved variables to match the number of loci for all the genes, which is required so that the vectors of minor allele frequencies come from the distribution having the same dimension. This "dimension-matching" is required for the theoretical development of our modeling ideas; see (5) and (6).

We assume that for every triplet (i, j, k), $\boldsymbol{X}_{ijk} = (\boldsymbol{x}_{ijk1}, \ldots, \boldsymbol{x}_{ijkL}) = (\boldsymbol{Y}_{ijk}, \tilde{\boldsymbol{Y}}_{ijk})$ have the mixture distribution

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

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

$$f\left(\boldsymbol{x}_{ijkr} | p_{mijkr}\right) = \left\{p_{mijkr}\right\}^{x^1_{ijkr}+x^2_{ijkr}} \left\{1 - p_{mijkr}\right\}^{2-(x^1_{ijkr}+x^2_{ijkr})}. \tag{2}$$

In the above, M denotes the *maximum* number of mixture components and p_{mijkr} stands for the minor allele frequency at the r-th locus of the j-th gene for the i-th individual of the k-th case/control group. Note that minor allele frequency is the frequency at which the second most common allele occurs in a given population.

Allocation variables z_{ijk}, with probability distribution

$$[z_{ijk} = m] = \pi_{mijk}, \tag{3}$$

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

$$[\boldsymbol{X}_{ijk} | z_{ijk}] = \prod_{r=1}^{L} f\left(\boldsymbol{x}_{ijkr} | p_{z_{ijk}ijkr}\right). \tag{4}$$

Following Majumdar *et al.* (2013), Bhattacharya and Bhattacharya (2018), we set $\pi_{mijk} = 1/M$, for $m = 1, \ldots, M$, and for all (j, k).

Letting $\boldsymbol{p}_{mijk} = (p_{mijk1}, p_{mijk2}, \ldots, p_{mijkL})$, we next assume that

$$\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk} \overset{iid}{\sim} \boldsymbol{G}_{ijk}; \tag{5}$$

$$\boldsymbol{G}_{ijk} \sim \mathrm{DP}\left(\alpha_{G,ik} \boldsymbol{G}_{0,jk}\right), \tag{6}$$

where $\mathrm{DP}\left(\alpha_{G,ik} \boldsymbol{G}_{0,jk}\right)$ stands for Dirichlet process with expected probability measure $\boldsymbol{G}_{0,jk}$ having precision parameter $\alpha_{G,ik}$, with

$$\log(\alpha_{G,ik}) = \mu_G + \boldsymbol{\beta}_G^T \boldsymbol{E}_{ik}, \tag{7}$$

where $\boldsymbol{E}_{ik}$ is a d-dimensional vector of continuous environmental variable for the i-th individual in the k-th group, $\boldsymbol{\beta}_G$ is a d-dimensional vector of regression coefficients, and μ_G is the intercept term. The model can be easily extended to include categorical environmental variables along with the continuous ones.

2.3. Hierarchical Dirichlet processes to induce dependence between the genes and case-control status

We further assume that for $k = 0, 1$,

$$\boldsymbol{G}_{0,jk} \overset{iid}{\sim} DP\left(\alpha_{G_0,k}\boldsymbol{H}_k\right);\ j = 1, \ldots, J, \tag{8}$$

where

$$\log(\alpha_{G_{0,k}}) = \mu_{G_0} + \boldsymbol{\beta}_{G_0}^T \bar{\boldsymbol{E}}_k, \tag{9}$$

with

$$\bar{\boldsymbol{E}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \boldsymbol{E}_{ik}. \tag{10}$$

We postulate the last level of hierarchy as

$$\boldsymbol{H}_k \overset{iid}{\sim} DP\left(\alpha_H \tilde{\boldsymbol{H}}\right);\ k = 0, 1, \tag{11}$$

where

$$\log(\alpha_H) = \mu_H + \boldsymbol{\beta}_H^T \bar{\bar{\boldsymbol{E}}}, \tag{12}$$

with

$$\bar{\bar{\boldsymbol{E}}} = \frac{\bar{\boldsymbol{E}}_0 + \bar{\boldsymbol{E}}_1}{2}. \tag{13}$$

We specify the base probability measure $\tilde{\boldsymbol{H}}$ as follows: for $m = 1, \ldots, M$, $i = 1, \ldots, N_k$, $k = 0, 1$, and $r = 1, \ldots, L$,

$$p_{mijkr} \overset{iid}{\sim} \text{Beta}\left(\nu_1, \nu_2\right), \tag{14}$$

under $\tilde{\boldsymbol{H}}$, where $\nu_1, \nu_2 > 0$.

This completes the specification of a hierarchy of Dirichlet processes to build dependence among the genes and the distributions of genotypes of cases-controls given data. Note that our model consists of one more level of hierarchy of Dirichlet processes than considered in the applications of Teh *et al.* (2006), who introduce hierarchical Dirichlet processes (HDP). Specifically, for given k and H_k, our hierarchy levels are comparable to that of Teh *et al.* (2006), but our extra level of hierarchy comes from (11), which creates dependence between case and control; details and reasons for insisting on such dependence structure are provided in Section 3.3.

Moreover, our likelihood based on Dirichlet processes ensuring at most M mixture components, is significantly different from those considered in the applications of Teh *et al.* (2006), which are based on the traditional DP mixture; see Mukhopadhyay *et al.* (2011), Mukhopadhyay *et al.* (2012), Mukhopadhyay and Bhattacharya (2013) for details on the conceptual, computational and asymptotic advantages of our modeling style over the traditional DP mixture.

2.4. The Chinese restaurant analogy

An extended version of the Chinese restaurant metaphor used by Teh *et al.* (2006) may be considered to illustrate our model. For $k = 0, 1$, the set of random probability measures $\{\boldsymbol{G}_{0,jk}; j = 1, \ldots, J\}$ can be associated with J restaurants. Letting τ_{ijk} denote the number of tables at the j-th restaurant associated with the i-th individual, we denote by $\boldsymbol{\phi}_{lijk}$ the dish being served at table l of the j-th restaurant for the i-th individual. Note that $\{\boldsymbol{\phi}_{lijk}; l = 1, \ldots, \tau_{ijk}; i = 1, \ldots, N_k\}$ is a set of *iid* realizations from $\boldsymbol{G}_{0,jk}$. Thus, we have different sets of realizations from $\boldsymbol{G}_{0,jk}$ for different individuals i.

For $k = 0, 1$, we also let $\boldsymbol{\Xi}_{R_k k} = \{\boldsymbol{\xi}_{1k}, \ldots, \boldsymbol{\xi}_{R_k k}\}$ denote a set of R_k *iid* realizations from $\boldsymbol{H}_k$. Then it follows that for $l = 1, \ldots, \tau_{ijk}$, $i = 1, \ldots, N_k$, and for $j = 1, \ldots, J$, $\boldsymbol{\phi}_{lijk} \in \boldsymbol{\Xi}_{R_k k}$. In other words, $\boldsymbol{\Xi}_{R_k k}$ is the set of distinct elements in the set $\{\boldsymbol{\phi}_{lijk}; l = 1, \ldots, \tau_{ijk}; i = 1, \ldots, N_k; j = 1, \ldots, J\}$, and, from the Chinese restaurant perspective, is the set of global dishes among all the restaurants, given k.

Finally, let $\boldsymbol{\zeta}_S = \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_S\}$ denote a set of S *iid* realizations from $\tilde{\boldsymbol{H}}$. Then it follows that $\boldsymbol{\zeta}_S$ is the set of distinct elements in $\{\boldsymbol{\Xi}_{R_k k} : k = 0, 1\}$. In other words, $\boldsymbol{\zeta}_S$ is the set of global dishes served in all the restaurants, irrespective of $k = 0$ or $k = 1$.

3. Discussion of the dependence structure induced by our HDP-based model

3.1. Dependence among individuals

It follows from the discussion in Section 2.4 that $\{\boldsymbol{\phi}_{lijk}; l = 1, \ldots, T_{mijk}; i = 1, \ldots, N_k\} \in \{\boldsymbol{\xi}_{1k}, \ldots, \boldsymbol{\xi}_{R_{mk}k}\}$, where $\boldsymbol{\xi}_{1k}, \ldots, \boldsymbol{\xi}_{R_{mk}k} \overset{iid}{\sim} \boldsymbol{H}_k$. This shows that $\{\boldsymbol{\phi}_{lijk}; l = 1, \ldots, T_{mijk}; i = 1, \ldots, N_k\}$ in (15) are shared among the individuals, thus creating dependence among the subjects.

For more precise insights regarding the dependence structure, let us first marginalize over $\boldsymbol{G}_{ijk}$ to obtain the joint distribution of $\boldsymbol{P}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \ldots, \boldsymbol{p}_{Mijk}\}$ using the following Polya urn distributions: given $\boldsymbol{G}_{0,jk}$, $\boldsymbol{p}_{1ijk} \sim \boldsymbol{G}_{0,jk}$, and for $m = 2, \ldots, M$,

$$[\boldsymbol{p}_{mijk} | \boldsymbol{p}_{lijk}; l < m] = \frac{\alpha_{G,ik}}{\alpha_{G,ik} + m - 1} \boldsymbol{G}_{0,jk}(\boldsymbol{p}_{mijk}) + \frac{1}{\alpha_{G,ik} + m - 1} \sum_{t=1}^{T_{mijk}} \tilde{n}_{tmijk} \delta_{\boldsymbol{\phi}_{tijk}}(\boldsymbol{p}_{mijk}), \tag{15}$$

where $\sum_{t=1}^{T_{mijk}} \tilde{n}_{tmijk} = m - 1$. Here $\tilde{n}_{tmijk} = \#\{l < m : \boldsymbol{p}_{lijk} = \boldsymbol{\phi}_{tijk}\}$.

Since conditionally on $\boldsymbol{G}_{0,jk}$, the marginal distribution of $\boldsymbol{p}_{mijk}$, for $m = 1, \ldots, M$ and $i = 1, \ldots, N_k$, is $\boldsymbol{G}_{0,jk}$, the marginal is unaffected by the environmental variable, but the joint distribution of $\boldsymbol{P}_{Mijk}$ implied by the Polya urn distributions (15) shows that the dependence structure of $\boldsymbol{P}_{Mijk}$ is influenced by the regression on $\boldsymbol{E}_{ik}$ through $\alpha_{G,ik}$. This is a very desirable property of our modeling approach, since, in reality, the population minor allele frequencies for the case-control group are not expected to be affected by environmental variables, although environmental exposure is expected to influence dependence among individuals and gene-gene interactions in individuals. Note that marginal distributions depending upon environmental variables may be envisaged only under mutation, but since it is an extremely rare phenomenon and the type of case control type genotype data we are dealing with is not appropriate for such studies, we do not include mutational effects in our

model.

3.2. Dependence among the genes

We now show that the gene-gene interactions of the i-th individual are affected by $\boldsymbol{E}_{ik}$, but not the marginal effects of the genes.

Dependence among the genes for the i-th individual is induced by $\{\boldsymbol{\phi}_{tijk}; t = 1, \ldots, \tau_{ijk}; j = 1, \ldots, J\}$, where, for $t = 1, \ldots, \tau_{ijk}$, $\boldsymbol{\phi}_{tijk} \overset{iid}{\sim} \boldsymbol{G}_{0,jk}$, with $\boldsymbol{G}_{0,jk} \sim DP\left(\alpha_{G_0,k}\boldsymbol{H}_k\right)$. In fact, marginalizing over $\boldsymbol{G}_{0,jk}$ yields the following Polya urn scheme for $\{\boldsymbol{\phi}_{tijk}; t = 1, \ldots, \tau_{ijk}\}$:

$$[\boldsymbol{\phi}_{tijk}|\boldsymbol{\phi}_{lijk}; l < t] = \frac{\alpha_{G_0,k}}{\alpha_{G_0,k} + t - 1}\boldsymbol{H}_k\left(\boldsymbol{\phi}_{tijk}\right) + \frac{1}{\alpha_{G_0,k} + t - 1}\sum_{l=1}^{R_{tk}} \bar{n}_{ltik}\delta_{\boldsymbol{\xi}_{lk}}\left(\boldsymbol{\phi}_{tijk}\right), \tag{16}$$

where $\bar{n}_{ltik} = \#\left\{\ell < t : \boldsymbol{\phi}_{\ell ijk} = \boldsymbol{\xi}_{lk}\right\}$. Note that $\sum_{l=1}^{R_{tk}} \bar{n}_{ltik} = t - 1$.

It is clear from (16) that $\{\boldsymbol{\phi}_{tijk}; j = 1, \ldots, J\}$ share $\{\boldsymbol{\xi}_{lk}; l = 1, \ldots, R_k\}$, so that the latter set creates dependence among the genes. Moreover, it is also clear from (16) that the dependence structure does not depend directly upon $\boldsymbol{E}_{ik}$, but upon $\bar{\boldsymbol{E}}_k$, through the regression of $\log(\alpha_{G_0,k})$ on $\bar{\boldsymbol{E}}_k$; see (9). In other words, the gene-gene dependence structure of any individual is not directly influenced by the corresponding environmental variable. However, the dependence structure is also influenced by $\bar{n}_{ltik}$, which depends upon the i-th individual in the k-th case-control group through τ_{ijk}, which is directly influenced by $\boldsymbol{E}_{ik}$ through $\alpha_{G,ik}$. Thus, as is desirable, our modeling style induces gene-gene interactions that are specific to the individuals and are influenced by the corresponding environmental variables and the averages of the environmental variables within the case-control groups that the individuals belong to.

It is also interesting to observe that in spite of the individual-specific gene-gene interactions, the marginal distributions of $\boldsymbol{\phi}_{tijk}$ remains $\boldsymbol{G}_{0,jk}$ for the non-marginalized version and $\boldsymbol{H}_k$ for the marginalized version characterized by (16), signifying that the individual genes are not affected by $\boldsymbol{E}_{ik}$.

3.3. Case-control dependence

Finally, we note that

$$[\boldsymbol{\xi}_{sk}|\boldsymbol{\xi}_{lk}; l < s] = \frac{\alpha_H}{\alpha_H + s - 1}\tilde{\boldsymbol{H}}\left(\boldsymbol{\xi}_{sk}\right) + \frac{1}{\alpha_H + s - 1}\sum_{l=1}^{S_{sk}} \breve{n}_{lsk}\delta_{\boldsymbol{\zeta}_l}\left(\boldsymbol{\xi}_{sk}\right), \tag{17}$$

where $\breve{n}_{lsk} = \#\left\{\ell < s : \boldsymbol{\xi}_{\ell k} = \boldsymbol{\zeta}_l\right\}$ and $\sum_{l=1}^{S_{sk}} \breve{n}_{lsk} = s - 1$. So, $\{\boldsymbol{\xi}_{sk}; s = 1, \ldots, R_k; k = 0, 1\}$ share $\{\boldsymbol{\zeta}_l; l = 1, \ldots, S\}$, creating dependence between case and control status. Dependence between case and control status are likely to be caused by various implicit factors and environmental variables that are not accounted for in the study. These factors and environmental variables may be insignificant individually, but together may exert non-negligible influence on cases and controls.

A schematic diagram of our HDP-based model and the dependence structure is depicted in Figure 1. We remark that in a much simpler set-up, the original HDP proposed

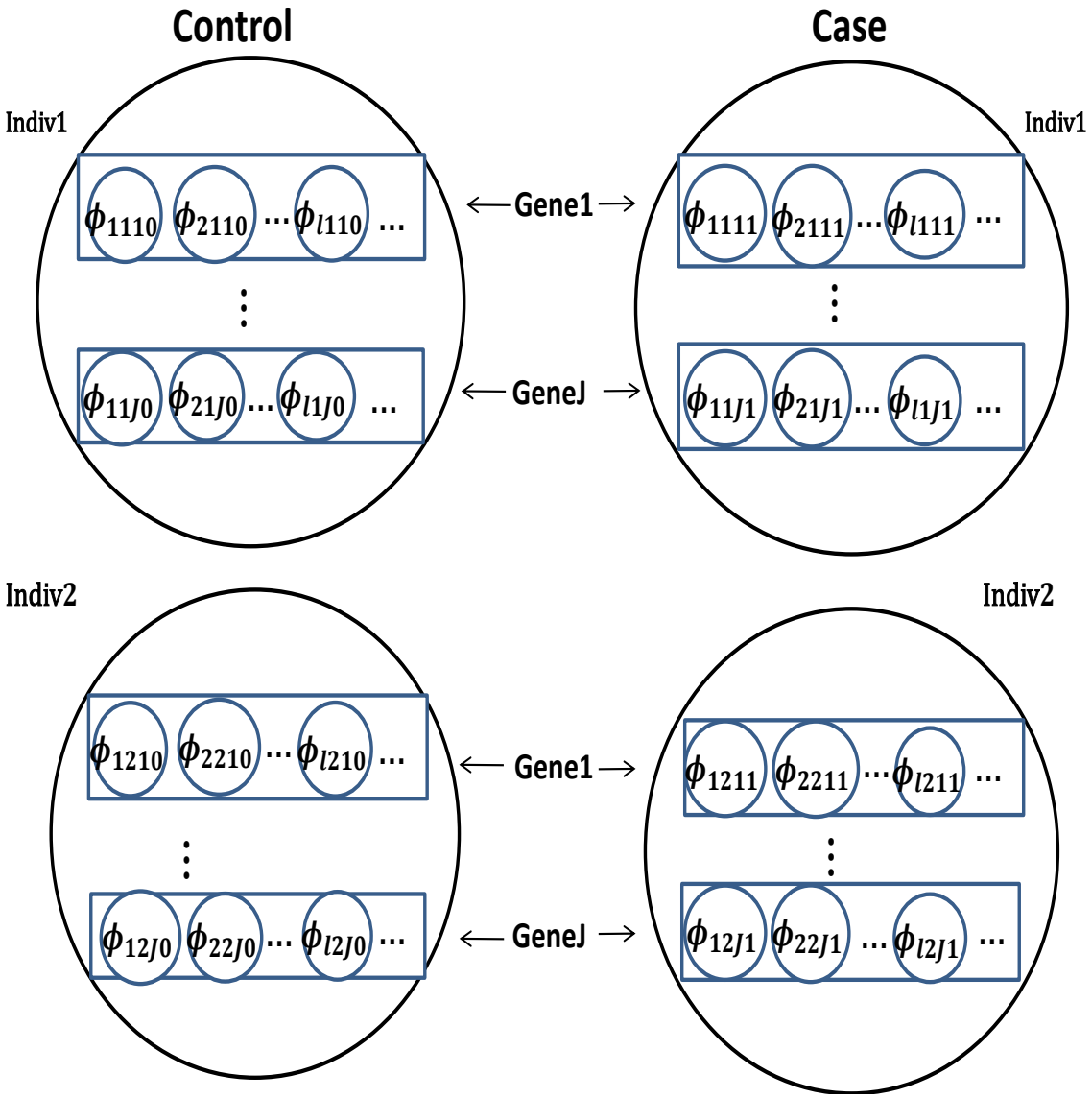


Figure 1: Schematic diagram of our HDP-based Bayesian model.

in Teh *et al.* (2006) has also been used by De Iorio *et al.* (2015) for inferring population admixture, allowing for correlations between loci due to linkage disequilibrium.

In Section A-1 we propose an MCMC procedure for the inferential purpose, and in Section A-2 we provide a parallel algorithm for implementing the MCMC method.

4. Detection of the roles of environment, genes and their interactions with respect to our HDP based model

4.1. Formulation of the tests and interpretation of their results

4.1.1. Bayesian test for the impact of the genes on case-control

To test if genes have any effect on case-control, we formulate as in Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), the following hypotheses:

$$H_{01} : h_{0j} = h_{1j};\ j = 1, \ldots, J, \tag{18}$$

versus

$$H_{11} : \text{not } H_0, \tag{19}$$

where

$$h_{0j}(\cdot) = \frac{1}{M} \sum_{m=1}^{M} \prod_{r=1}^{L_j} f\left(\cdot | p^r_{mi_0jk=0}\right); \tag{20}$$

$$h_{1j}(\cdot) = \frac{1}{M} \sum_{m=1}^{M} \prod_{r=1}^{L_j} f\left(\cdot | p^r_{mi_1jk=1}\right). \tag{21}$$

In the above, 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 4.2.1.

4.1.2. Bayesian test for significance of the environmental variables

To check if the environmental variables are significant, we shall test the following: for $\ell = 1, \ldots, d$,

$$H_{02\ell} : \beta_{G,\ell} = 0 \text{ versus } H_{12\ell} : \beta_{G,\ell} \neq 0, \tag{22}$$

$$H_{03\ell} : \beta_{G_0,\ell} = 0 \text{ versus } H_{13\ell} : \beta_{G_0,\ell} \neq 0, \tag{23}$$

and

$$H_{04\ell} : \beta_{H,\ell} = 0 \text{ versus } H_{14\ell} : \beta_{H,\ell} \neq 0. \tag{24}$$

4.1.3. Bayesian test for significance of gene-gene interaction

In our HDP based nonparametric model there is no readily available quantification of gene-gene interaction unlike the models of Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020). Thus, in order to test for gene-gene interaction, it is necessary to first reasonably define such a measurement.

A measure of gene-gene interaction influenced by environmental variables

For our purpose, we first define

$$\bar{p}_{mijk} = \frac{\sum_{r=1}^{L_j} p_{mijkr}}{L_j}. \tag{25}$$

With the above definition, for subject i belonging to case-control group k, we consider the following covariance

$$C(i, j_1, j_2, k) = cov\left(\text{logit}(\bar{p}_{z_{ij_1k}ij_1k}), \text{logit}(\bar{p}_{z_{ij_2k}ij_2k})\right), \tag{26}$$

as quantification of subject-wise gene-gene dependence that accounts for population memberships of subject i with respect to genes j_1 and j_2, through z_{ij_1k} and z_{ij_2k}, where for any $p \in (0,1)$, $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. Thus, gene-gene interaction associated with our model is subject-specific.

While implementing our model using our parallelised MCMC methodology, we simulate $C(i, j_1, j_2, k)$ at each iteration by generating $\{p_{mijkr} : r = 1, \ldots, L_j\}$ as many times as required from the respective full conditionals holding the remaining parameters fixed, and then compute the empirical covariance corresponding to (26) using the generated *iid* samples conditionally on the remaining parameters to approximate (26).

Formulation of the Bayesian tests for gene-gene interactions

To test for subject-wise gene-gene interaction, we consider the following tests: for $i = 1, \ldots, N_k$, $k = 0, 1$, and for $j_1, j_2 \in \{1, \ldots, J\}$,

$$H_{05ij_1j_2k} : C(i, j_1, j_2, k) = 0 \text{ versus } H_{15ij_1j_2k} : C(i, j_1, j_2, k) \neq 0. \tag{27}$$

4.1.4. Interpretations of the results of the above tests

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

- For some appropriate divergence measure d between two distributions, if $\max_{1 \leq j \leq J} d(h_{0j}, h_{1j})$, 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 j-th 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 at least one of $\beta_{G,\ell}$ or $\beta_{G_0,\ell}$ or $\beta_{H,\ell}$ is significant, at least for some ℓ. This may be interpreted as the environmental variable $\boldsymbol{E}$ having some altering effect on all the genes in a way that doesn't affect the disease status. If $C(i, j_1, j_2, k)$ turns out to be significant, then this would additionally imply that the environmental variable $\boldsymbol{E}$ influences interaction between genes j_1 and j_2 for the i-th individual, but not in a way that is responsible for the case/control status.
- If H_{01} is rejected, indicating that the genes are significant, but none of the $\beta_{G,\ell}$, $\beta_{G_0,\ell}$, $\beta_{H,\ell}$ or $C(i, j_1, j_2, k)$ are significant, then only the genes, not $\boldsymbol{E}$, are responsible for the disease. In that case, one may conclude that the disease is of purely genetic nature.

- Suppose that H_{01} is rejected, none of $\beta_{G,\ell}$, $\beta_{G_0,\ell}$, $\beta_{H,\ell}$ is significant, but $C(i, j_1, j_2, k)$ is significant for at least some i, j_1, j_2, k. Then the environmental variable is not significant, and the case/control status of the individuals associated with significant gene-gene interactions can be attributed to purely genetic causes triggered by gene-gene interactions of the individuals.

- Now suppose that H_{01} is rejected, and at least one of $\beta_{G,\ell}$, $\beta_{G_0,\ell}$, $\beta_{H,\ell}$ is significant, but none of the subject-wise gene-gene interactions is significant. Then the environmental variable $\boldsymbol{E}$ does not significantly affect the interactions to determine the case/control status, and marginal effects of the individual genes are responsible for the case/control status of an individual.

- If, on the other hand, H_{01} is rejected, at least one of $\beta_{G,\ell}$, $\beta_{G_0,\ell}$, $\beta_{H,\ell}$ is significant, and $C(i, j_1, j_2, k)$ is significant for at least some i, j_1, j_2, k, then the environmental variable is significant and is responsible for influencing gene-gene interactions within the individuals with significant $C(i, j_1, j_2, k)$, which, in turn, affects the case/control status of the individuals.

4.2. Methodologies for implementing the Bayesian tests

4.2.1. Hypothesis testing based on clustering modes

As in Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), here we exploit the concept of "central" clustering introduced by Mukhopadhyay *et al.* (2011). Briefly, central clustering may be interpreted as a suitable measure of central tendency of a set of clusterings. Mukhopadhyay *et al.* (2011) particularly consider the mode(s) of the set of clusterings, and provide methods for appropriately obtaining the mode(s) using a suitable metric that they propose to quantify distances between any two clusterings. Their proposed metric is also computationally inexpensive, which makes the concept based on central clusterings extremely useful in practice.

For $k = 0, 1$, let i_k denote the index of the central clusterings of $\boldsymbol{P}_{Mijk} = \{\boldsymbol{p}_{1ijk}, \boldsymbol{p}_{2ijk}, \ldots, \boldsymbol{p}_{Mijk}\}$, $i = 1, \ldots, N_k$. We then study the 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 $j = 1, \ldots, J$. A schematic diagram illustrating the idea can be found in Bhattacharya and Bhattacharya (2020).

Significantly large divergence between the two clusterings clearly indicates that the j-th gene is marginally significant.

4.2.2. Enhancement of clustering metric based inference using Euclidean distance

As argued in Bhattacharya and Bhattacharya (2018), significantly large clustering distance between $\boldsymbol{P}_{Mjk=0}$ and $\boldsymbol{P}_{Mjk=1}$ indicates rejection of H_0, but insignificant clustering distance does not necessarily provide strong evidence in favour of the null. In this regard, Bhattacharya and Bhattacharya (2018) (see also Bhattacharya and Bhattacharya (2020)) argue

that the Euclidean distance is an appropriate candidate to be tested for significance before arriving at the final conclusion. Briefly, 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

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

and

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

We denote the Euclidean distance associated with the j-th gene by

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

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

4.2.3. 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:

$$H_{0,d^*}: d^* < \varepsilon_{d^*} \text{ versus } H_{1,d^*}: d^* \ge \varepsilon_{d^*}; \tag{28}$$

$$H_{0,d_E^*}: d_E^* < \varepsilon_{d_E^*} \text{ versus } H_{1,d_E^*}: d_E^* \ge \varepsilon_{d_E^*}; \tag{29}$$

for $\ell = 1, \ldots, d$,

$$H_{0,\beta_{G,\ell}}: |\beta_{G,\ell}| < \varepsilon_{G,\ell} \text{ versus } H_{1,\beta_{G,\ell}}: |\beta_{G,\ell}| \ge \varepsilon_{G,\ell}, \tag{30}$$

$$H_{0,\beta_{G_0,\ell}}: |\beta_{G_0,\ell}| < \varepsilon_{G_0,\ell} \text{ versus } H_{1,\beta_{G_0,\ell}}: |\beta_{G_0,\ell}| \ge \varepsilon_{G_0,\ell}, \tag{31}$$

$$H_{0,\beta_{H,\ell}}: |\beta_{H,\ell}| < \varepsilon_{H,\ell} \text{ versus } H_{1,\beta_{H,\ell}}: |\beta_{H,\ell}| \ge \varepsilon_{H,\ell}, \tag{32}$$

and, for $i = 1, \ldots, N_k$, $k = 0, 1$, $j_1, j_2 \in \{1, \ldots, J\}$,

$$H_{0,C_{i,j_1,j_2,k}}: |C_{i,j_1,j_2,k}| < \varepsilon_{C,ij_1j_2k} \text{ versus } H_{1,C_{i,j_1,j_2,k}}: |C_{i,j_1,j_2,k}| \ge \varepsilon_{C,ij_1j_2k}, \tag{33}$$

If H_0 is rejected in (28) or in (29), we could also test if the j-th 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 $H_{1,d_{E,j}}: d_{E,j} \ge \varepsilon_{d_{E,j}}$.

4.2.4. Null model and choice of ε

To obtain the null posterior distribution, we fit our HDP-based Bayesian model to the dataset generated from the HDP-based model where the genes are independent and not influenced by the environmental variable, and where there is no difference between the probabilities associated with case and control. For the null data we chose the same number of genes, the same number of loci for each gene, and the same number of cases and controls as the non-null data. We also choose the same value M as in the non-null model, but set $\beta_G = \beta_{G_0} = \beta_H = 0$. To generate the data from the null model, we first

simulate, independently for $j = 1, \ldots, J$, the set $\{\boldsymbol{p}_{m1j0} : m = 1, \ldots, M\}$, using the Polya urn scheme involving $\tilde{\boldsymbol{H}}$ and α_H, and set $\{\boldsymbol{p}_{m1j1} : m = 1, \ldots, M\} = \{\boldsymbol{p}_{m1j0} : m = 1, \ldots, M\}$, so that there is no difference between the probabilities associated with case and control, and that the genes are independent. Since the simulation method is independent of the environmental variable, it is clear that the genes are not influenced by the environment. Given the probabilities $\{\boldsymbol{p}_{m1j1} : m = 1, \ldots, M\}$ and $\{\boldsymbol{p}_{m1j0} : m = 1, \ldots, M\}$, we then simulate the data using our Bernoulli model. To the data thus generated, we fit our full HDP-based Bayesian model, to obtain the null posterior.

As in Bhattacharya and Bhattacharya (2018) here also we specify ε's as $F^{-1}(0.55)$, where F is the distribution function of the relevant benchmark null posterior distribution. Bhattacharya and Bhattacharya (2018) showed that the choice $F^{-1}(0.55)$, rather than the median, ensures that the correct null hypothesis is accepted under the "$0-1$" loss. Note that, for the median, the posterior probability of the true null is 0.5, while under the "$0-1$" loss, the true null will be accepted if its posterior probability is greater than $1/2$.

5. Simulation studies

For simulation studies, we first generate realistic biological data for stratified population with known gene-environment interaction from the GENS2 software of Pinelli *et al.* (2012). To this data, we then apply our model and methodologies in an effort to detect gene-environment interaction effects that are present in the data. We consider simulation studies in 5 different true model set-ups: (a) presence of gene-gene and gene-environment interaction, (b) absence of genetic or gene-environmental interaction 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. The details of our simulation experiments are provided in Section A-3 of the supplement. Here we briefly summarize the results of our experiments.

In case (a), we correctly obtained clear significance of the influence of genetic effects. Also, β_H turned out to be very significant, demonstrating significant overall impact of the environmental variable on the genes. Interestingly, as one may expect, there are more instances of significant gene-gene interactions in the case group compared to the control group. The posteriors of the number of sub-populations gave high probabilities to the correct number of sub-populations in all the 5 simulation experiments. Quite importantly, we demonstrate in cases (a), (d) and (e) where the genes are relevant, that our HDP model can detect disease predisposing loci (DPL) with more precision compared to the matrix-normal-inverse-Wishart model for gene-environment interactions proposed in Bhattacharya and Bhattacharya (2020). In case (b) using our ideas in conjunction with significance testing in a simple logistic regression framework, we are correctly able to conclude that the genetic or gene-environmental effects are insignificant. As in Bhattacharya and Bhattacharya (2020), the right conclusion is arrived at in case (c) by utilizing our ideas in conjunction with the Akaike Information Criterion (AIC) in the context of simple logistic regression. Using our Bayesian testing procedure along with the aid of logistic regression, we have been able to correctly obtain insignificance of the environmental variable and significance of the genes. In this experiment, we found no gene-gene interaction in the control group and found two (marginal) instances of gene-gene interaction among the cases. As regards case (e), we note as in Bhattacharya and

Bhattacharya (2020) that additivity of genetic and environmental effects is not supported even by our current HDP-based Bayesian model. In spite of this, we correctly obtained significance of the environmental variable and precisely obtained the DPLs. But the lack of the additivity criterion in our model seems to have forced gene-environment interactions. Bhattacharya and Bhattacharya (2020) report similar results, who obtained, after eventually resorting to logistic regression, the AIC-based best model consisting of the additive marginal effects of the first gene and the environmental variable, along with an additive intercept, which is broadly consistent with the data-generating mechanism.

6. Application of our HDP based ideas to a real, case-control dataset on Myocardial Infarction

We now consider application of our model and methods to a case-control dataset on early-onset of myocardial infarction (MI) from MI Gen study, obtained from the dbGaP database **http://www.ncbi.nlm.nih.gov/gap**. The same dataset has been analyzed by Bhattacharya and Bhattacharya (2018) without considering the sex variable as the covariate, and by Bhattacharya and Bhattacharya (2020), who incorporate the sex variable in their gene-environment interaction model. Although Bhattacharya and Bhattacharya (2018) obtained significant genetic and gene-gene interaction effects, their later study after considering sex as the environmental variable, revealed strong effects of the sex variable but no significant gene-gene interaction, although many of the genes turned out to be individually significant. In our current HDP based analysis, we once again obtain strong effects of the sex variable, but in contrast with Bhattacharya and Bhattacharya (2020), although we obtain significant genetic effects, none of the genes turned out to be significant individually. Moreover, the subject-wise gene-gene interactions, although of small magnitude, turned out to be significant in some cases, and interestingly (and apparently counter-intuitively) seem to be instrumental in counter-acting the disease rather than provoking it.

6.1. Data description

We recall that the MI Gen data obtained from dbGaP consists of observations on presence/absence of minor alleles at 727478 SNP markers associated with 22 autosomes and the sex chromosomes of 2967 cases of early-onset myocardial infarction, 3075 age and sex matched controls. The average age at the time of MI was 41 years among the male cases and 47 years among the female cases. The data broadly represents a mixture of four subpopulations: Caucasian, Han Chinese, Japanese and Yoruban. Using the Ensembl human genome database (**http://www.ensembl.org/**) we could categorize 446765 markers out of 727478 with respect to 37233 genes.

As in Bhattacharya and Bhattacharya (2020) we considered 32 genes covering 1251 loci, for 200 individuals. These 1251 loci include 33 SNPs that are believed to be associated with MI and also those that are believed to be associated with different cardiovascular end points like LDL cholesterol, smoking, blood pressure, body mass, etc. Other than the 33 SNPs, the remaining 1218 SNPs are not known to be associated with the disease. See Bhattacharya and Bhattacharya (2020) for the details and the relevant references.

Since the four broad sub-populations are not unlikely to admit further genetic subdivisions, it makes sense to set the maximum number of mixture components, M, to a

value much larger than 4. As before, we set $M = 30$; we also set $\nu_1 = \nu_2 = 1$, so that $\tilde{\boldsymbol{H}}$ is the uniform distribution on $[0, 1]$. As in the simulation experiments, here also the structures $\alpha_{G,ik} = 0.1 \times \exp(100 + \mu_G + \beta_G E_{ik})$, $\alpha_{G_0,k} = 0.1 \times \exp\left(100 + \mu_{G_0} + \beta_{G_0}\bar{E}_k\right)$ and $\alpha_H = 0.1 \times \exp\left(100 + \mu_H + \beta_H \bar{\bar{E}}\right)$, where $\mu_G, \mu_{G_0}, \mu_H \overset{iid}{\sim} U(0, 1)$ and $\beta_G, \beta_{G_0}, \beta_H \overset{iid}{\sim} U(-1, 1)$, ensured adequate number of sub-populations and satisfactory mixing of MCMC. For the null data and model, we follow the same procedure as discussed in Section 4.2.4.

6.2. Remarks on model implementation

Our parallel MCMC algorithm detailed in Section A-2 takes about 7 days to generate 30,000 iterations on our VMware consisting of 50 double-threaded, 64-bit physical cores, each running at 2493.990 MHz. We discard the first $10,000$ iterations as burn-in, using the subsequent 20,000 iterations for our Bayesian inference. Satisfactory mixing properties are indicated by informal convergence diagnostics such as trace plots.

6.3. Results of the real data analysis

6.3.1. Effect of the sex variable

We obtain $P(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}) \approx 0$, $P(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}) \approx 0$ and $P(|\beta_H| < \varepsilon_{\beta_H}|\text{Data}) \approx 1$. In other words, although $\bar{\bar{E}}$ (here E being the sex variable) is insignificant, both E_{ik} and $\bar{E}_k$ are very significant. Thus, in this study, sex seems to play an important role in influencing the genes.

6.3.2. Roles of individual genes

With the clustering metric we obtained $P(d^* < \epsilon_1|\text{Data}) \approx 0.030$ while that with the Euclidean distance we obtained $P(d_E^* < \epsilon_2|\text{Data}) \approx 0.540$. That is, the maximum of the gene-wise clustering metrics turns out to be significant, while the maximum of the gene-wise Euclidean metrics is seen to be insignificant. The same ambiguity was also obtained by Bhattacharya and Bhattacharya (2020). The tests of the marginal genes are expected to shed some light regarding this dilemma. The posterior probabilities of the null hypotheses (of no significant genetic influence) reveal that none of the individual genes are significant, for either the clustering metric or the Euclidean metric. Our result is not much different from that of Bhattacharya and Bhattacharya (2020) who also note that their marginal probabilities, at least for the clustering metric, are not significantly small to provide strong enough evidences against the nulls.

Now, at least from the clustering metric perspective, it is necessary to explain the issue that all the genes are insignificant individually but still the maximum of the gene-wise clustering metric values is significant. The key to this issue seems to be the correlations between the distances, which are induced by gene-gene interactions. We explain this phenomenon using a bivariate normal example. Let (X_1, X_2) have a bivariate normal distribution with means 0, variances 1, and correlation ρ. Figure 2 depicts the median of $\max\{X_1, X_2\}$ as a function of ρ, which is seen to be increasing as ρ decreases from 1 to -1. On the other hand, the medians of the marginal distributions of X_1 and X_2 remain zero, irrespective of the value of ρ. Thus, it seems that gene-gene interaction does have some role to play in this study.

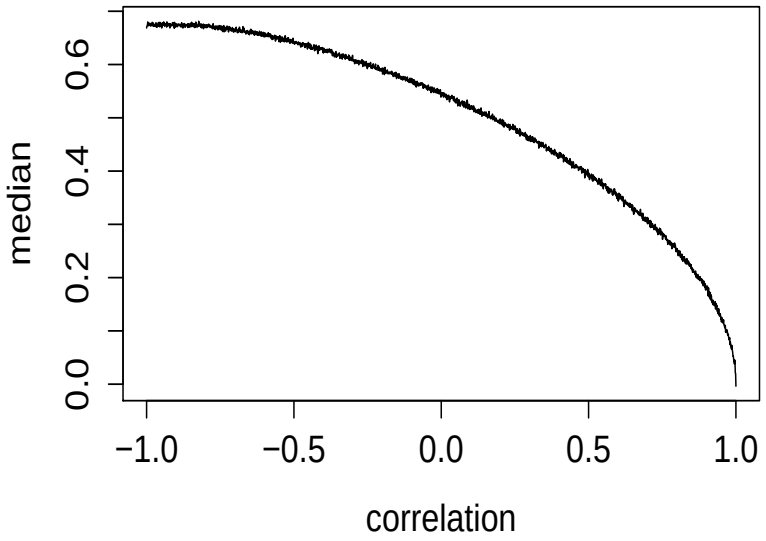


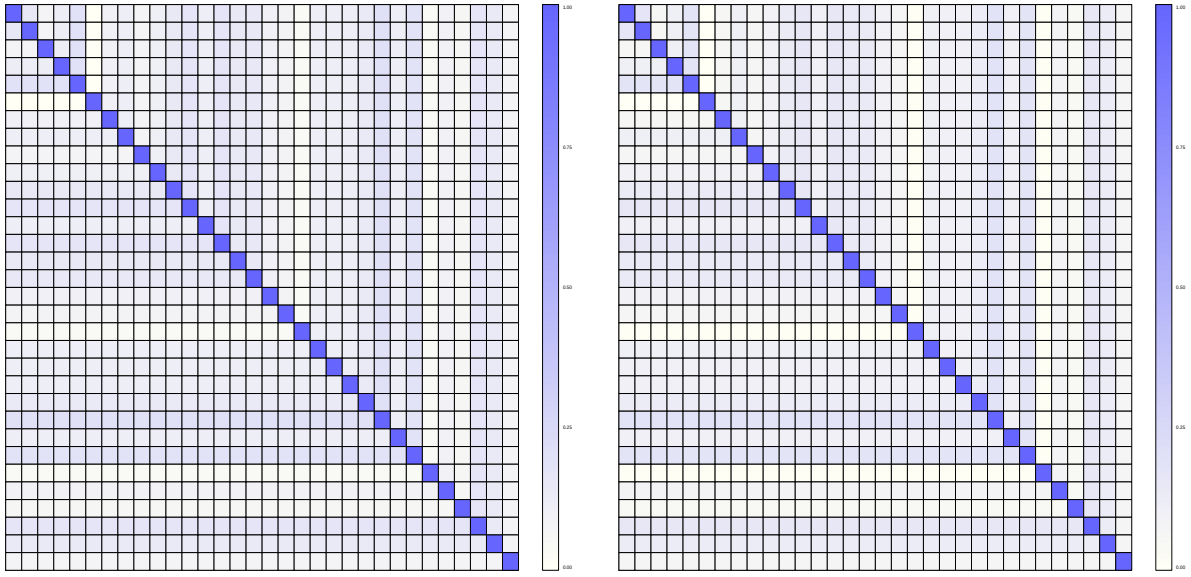
Figure 2: Bivariate normal example: Plot of the median of $\max\{X_1, X_2\}$ **with respect to the correlation** ρ**.**

6.3.3. Gene-gene interactions

Unlike Bhattacharya and Bhattacharya (2020), where there is a single gene-gene correlation structure for all the individuals, our current model has provision for subject-specific gene-gene correlations. Figures 3 and 4 show the typical gene-gene correlations representative of cases and controls in all males and females respectively. Essentially, the pictures represent the gene-gene correlation patterns for all the subjects. The color intensities correspond to the absolute values of the correlations. Although the correlations are small in all the subjects, the tests of hypotheses reveal some interesting structures. Figures 5 and 6 represent the all possible interacting patterns found in the study. Panel (a) of Figure 5 represents 9 male cases where no gene-gene interaction is significant. Panel (b) shows the genetic interaction pattern in some male cases where $AP006216.10$ and $C6orf106$, interact with all the other genes. Panel (c) shows the results of significance tests of gene-gene interactions for some male cases, for whom only $AP006216.10$ interacts with all the other genes in the study. A representative interaction pattern for the male controls shown in panel (d), indicates that only $C6orf106$ or only $AP006216.10$ interacts with every gene, but in a few subjects both $AP006216.10$ and $C6orf106$ interact with all the genes.

Even for the females, the two genes, $AP006216.10$ and $C6orf106$, play significant roles in gene-gene interactions. Indeed, in our data, unlike the 9 male cases, there is no female for whom all gene-gene interactions are insignificant. The relevant representative plots for the females, given by Figure 6, shows that for all the female cases, only $AP006216.10$ interacts with the other genes. For the female controls, either only $AP006216.10$ or only $C6orf106$ interacts with the other genes, or both $AP006216.10$ and $C6orf106$ interact significantly with the other genes included in the study.

The messages gained from our analysis seem to be somewhat counter-intuitive but perhaps quite insightful. Our tests indicate that the genes have insignificant marginal effect. Thus, some external or non-genetic factors might have some significant role to play. But for most of the subjects, at least one of the genes $AP006216.10$ and $C6orf106$ interact

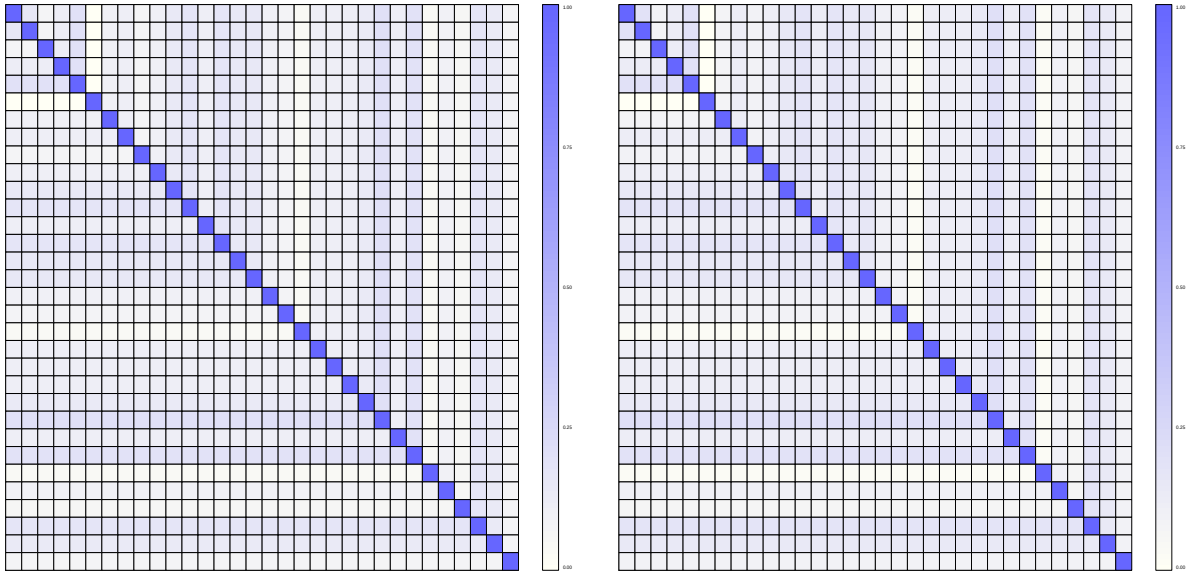


(a) Male case. (b) Male control.

Figure 3: Typical median gene-gene posterior correlation plot for male cases and male control.

with every other gene. The subjects, for whom no significant genetic interactions involving $AP006216.10$ and $C6orf106$ were detected, turned out to be male cases, indicating that the lack of genetic interaction in these males failed to get them any preventive measure against MI. On the other hand, the interactions of the genes $AP006216.10$ and $C6orf106$ with all the genes seemed to reduce the risk of the disease for the other subjects. Thus, in this study, the gene-gene interactions seem to have a beneficial effect on the subjects. It also seems that only a small proportion of males are prone to the risk of having no beneficial gene-gene interactions.

Note that our results are broadly consistent with those obtained by Bhattacharya and Bhattacharya (2020) but are more precise and informative. Indeed, they also noted relatively small impact of the individual genes and small gene-gene correlations. Our current ideas and analyses also support their conclusion that external factors (in particular, sex) are perhaps playing a bigger role in explaining case-control with respect to MI. We recall (see Bhattacharya and Bhattacharya (2020)) that with respect to the data that we used, the empirical conditional probability of a male given case is about 0.38, and that of a male given control is about 0.50, so that females seem to be more at risk, given our data. The inherent coherence of the Bayesian paradigm upholds the sex factor by attaching little importance to the individual genes. However, in contrast with Bhattacharya and Bhattacharya (2020) who found no interacting genes, here it turns out that the genes $AP006216.10$ and $C6orf106$ in interaction with other genes generally lower the risk of the individuals with respect to MI. Importantly, each of the few males having no such interactions turned out to be a case. This seems to be roughly in accordance with the popular belief that males are more susceptible to MI than females. Our Bayesian model coherently weaves together the prior and the data and brings out this information in spite of the data-driven information that females are more prone to MI than males. We also note that Lucas *et al.* (2012), who analyzed the same MI



(a) Female case. (b) Female control.

Figure 4: Typical median gene-gene posterior correlation plot for female cases and female controls.

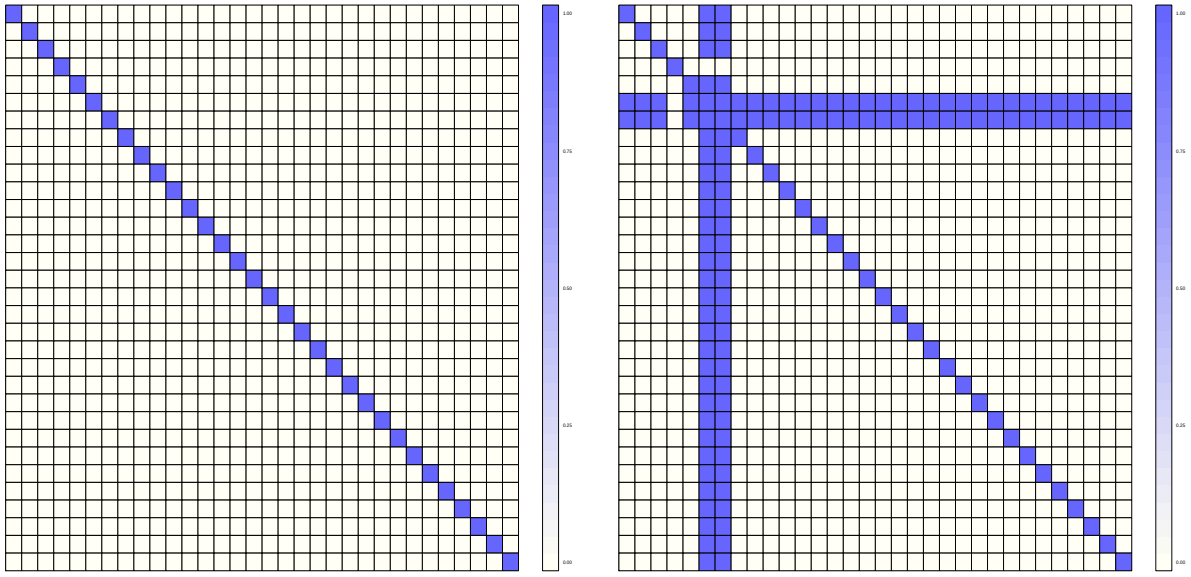
dataset using logistic regression, reached the conclusion that there is no significant gene-gene interaction. Thus, their result completely supports that of Bhattacharya and Bhattacharya (2020) and are also very much in keeping with our current results.

6.3.4. Posteriors of the number of sub-populations

The posterior distributions of the number of sub-populations for the males and females turned out to be quite similar, irrespective of case and control, with the mode at 3 and 4 components receiving the next highest probability. Thus, the 4 sub-populations, irrespective of sex, are well-supported by our model, showing that these can not be further sub-divided genetically. This is not unexpected, since the roles of the individual genes are not significant in our study. Our result broadly agrees with Bhattacharya and Bhattacharya (2020) who obtained for different genes, the modes at 5 components, with 4 components receiving the next highest posterior mass.

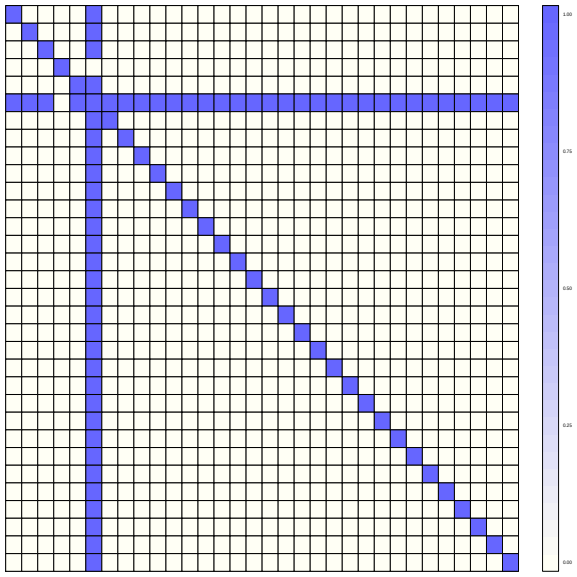
7. Summary and conclusion

In this paper, we have proposed a novel Bayesian nonparametric gene-gene and gene-environment interaction model based on hierarchies of Dirichlet processes. This model is a significant improvement over that of Bhattacharya and Bhattacharya (2020) in the sense of much clear interpretability and accounting for subject-specific gene-gene interactions. Moreover, the interactions arise as natural by-products of our nonparametric structure based on HDP, and are not based on matrix normal distributions, as in Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), and hence, are more realistic. We propose a novel parallel MCMC algorithm to implement our model, that combines powerful technology with conditionally independent structures inherent within our HDP based

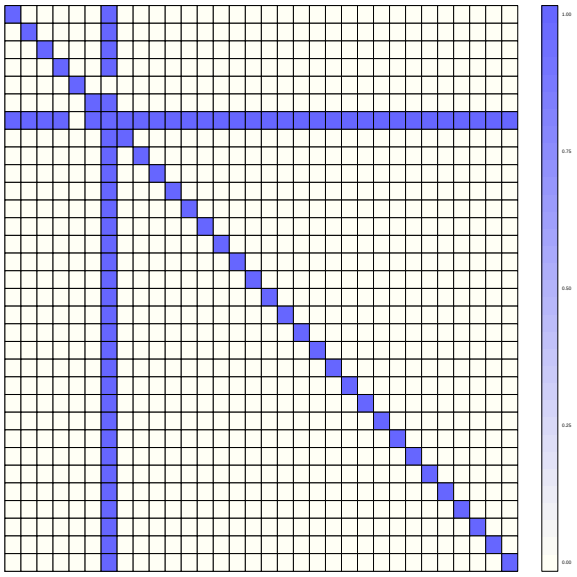


(a) Male case.

(b) Male case.

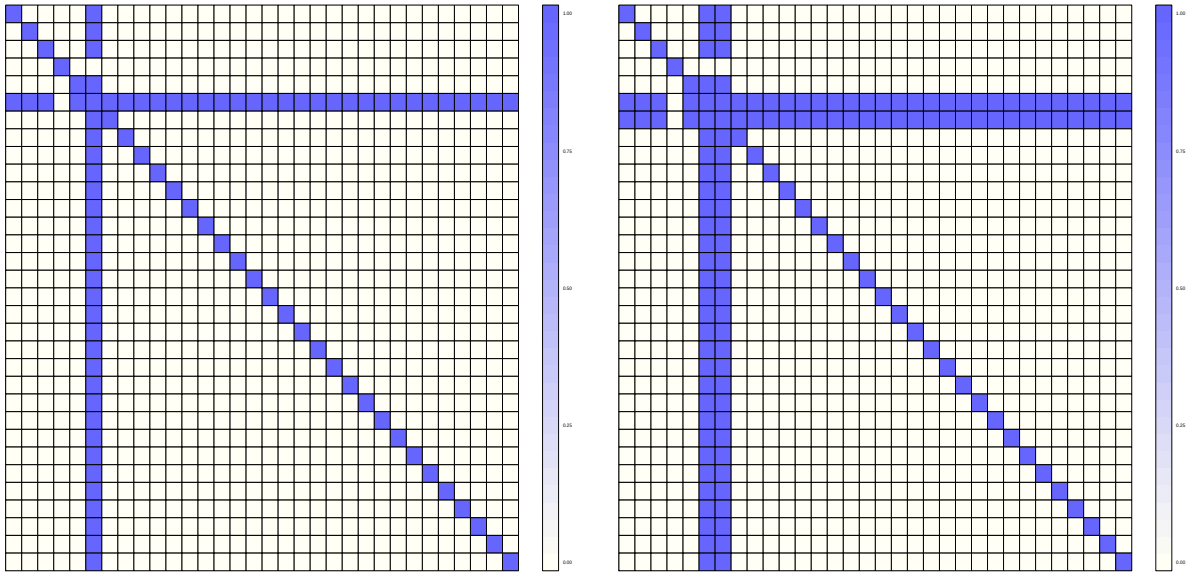


(c) Male case.



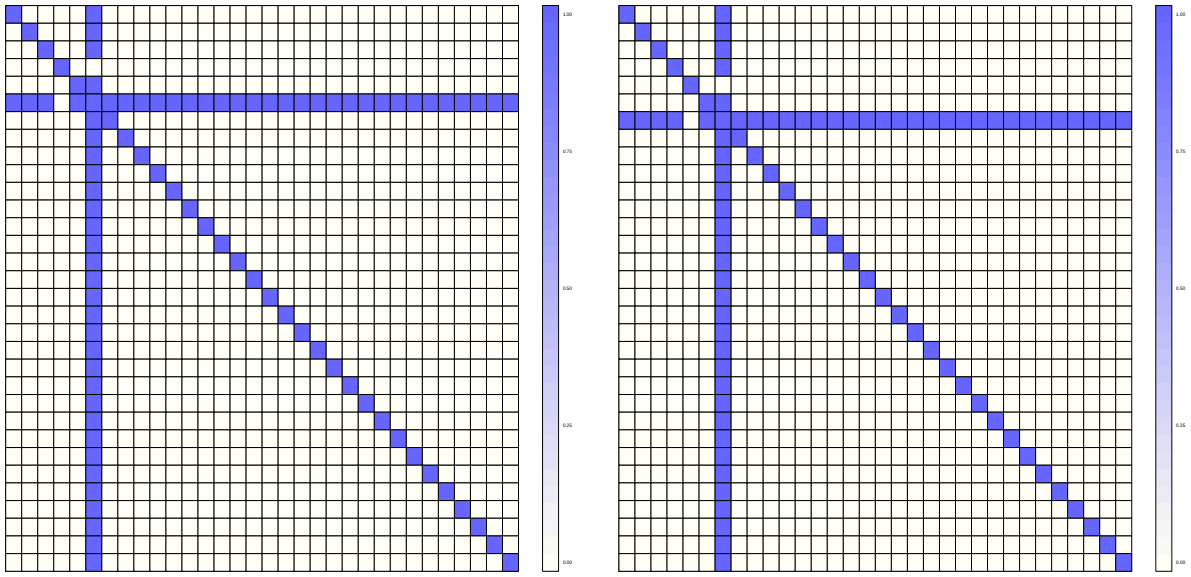
(d) Male control.

Figure 5: Presence/absence of gene-gene interactions for typical male cases and controls: Blue denotes presence and white represents absence of gene-gene interaction.



(a) Female case.

(b) Female control.



(c) Female control.

(d) Female control.

Figure 6: Presence/absence of gene-gene interactions for typical female cases and controls: Blue denotes presence and white represents absence of gene-gene interaction.

model and efficient TMCMC methods. The Bayesian tests of hypotheses that we employ in this paper are are appropriately modified versions of those proposed in Bhattacharya and Bhattacharya (2020).

Applications of our ideas to biologically realistic datasets generated under 5 different set-ups characterized by different combinations and structures associated with gene-gene and gene-environment interactions demonstrated encouraging performance of our model and methods. Our analysis of the MI dataset showed strong impact of the sex variable, which is consistent with the results of Bhattacharya and Bhattacharya (2020). Our tests showed no effect of the individual genes, which is also in keeping with Bhattacharya and Bhattacharya (2020) who obtained relatively weak marginal effects. But most interestingly, even though we obtained very weak gene-gene correlations in accordance with Bhattacharya and Bhattacharya (2020) and Lucas *et al.* (2012), our tests on gene-gene interaction showed that two genes, $AP006216.10$ and $C6orf106$, generally interact with all the other genes in a beneficial way so as to fight the disease. Moreover, the only situations where all the gene-gene interactions turned out to be insignificant, were the male cases, showing that the usual belief that males are more prone to heart attack than females may hold some value from this perspective.

Although many standard methods are commonly used in GWAS to identify the genetic and the environmental effects, there are several reasons that point towards the fact that our approach is not comparable with the existing methods.

So far, due to insufficient computational resources, we are compelled to restrict focus on a relatively small portion of the data. For improving our computing infrastructure, we have already taken the initiative of procuring supercomputing facilities from the Govt. of India, a project led, on behalf of Indian Statistical Institute, by the second author of this paper. With such a facility, we will be able to analyze the entire MI dataset with much ease.

Acknowledgment

We are sincerely grateful to the Editor-in-Chief and the anonymous reviewer, whose comments have led to significant improvement of our manuscript.

References

Ahn, J., Mukherjee, B., Gruber, S. B., and Ghosh, M. (2013). Bayesian semiparametric analysis for two-phase studies of gene-environment interaction. *The Annals of Applied Statistics*, **7**, 543–569.

Bhattacharya, D. and Bhattacharya, S. (2018). A Bayesian semiparametric approach to learning about gene-gene interactions in case-control studies. *Journal of Applied Statistics*, **45**, 1–23.

Bhattacharya, D. and Bhattacharya, S. (2020). Effects of gene-environment and gene-gene interactions in case-control studies: A Novel Bayesian Semiparametric Approach. *Brazilian Journal of Probability and Statistics*, **34**, 71–89.

De Iorio, M., Elliott, L. T., Favaro, S., Adhikari, K., and Teh, Y. W. (2015). Modeling population structure under hierarchical Dirichlet processes. Available at "https://arxiv.org/abs/1503.08278".

Dey, K. K. and Bhattacharya, S. (2017). On geometric ergodicity of additive and multiplicative transformation based Markov chain Monte Carlo in high dimensions. *Brazilian Journal of Probability and Statistics*, **31**, 569–617. Also available at "http://arxiv.org/pdf/1312.0915.pdf".

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

Liu, C., Ma, J., and Amos, C. I. (2015). Bayesian variable selection for hierarchical gene-environment and gene-gene interactions. *Human Genetics*, **134**, 23–36.

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**, 1–8.

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.

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

Mukhopadhyay, S. and Bhattacharya, S. (2013). Bayesian MISE convergence rates of mixture models based on the Polya urn model: Asymptotic comparisons and choice of prior parameters. Available at http://arxiv.org/abs/1205.5508.

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.

Mukhopadhyay, S., Roy, S., and Bhattacharya, S. (2012). Fast and efficient Bayesian semiparametric curve-fitting and clustering in massive data. *Sankhya. Series B*, **71**, 77–106.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet processes hierarchical models. *Biometrika*, **95**, 169–186.

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**.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

Urbut, S. M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, **51**, 187–195.

Wen, X. and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogenous subgroups: From meta-analyses to gene-environment iInteractions. *Annals of Applied Statistics*, **8**, 176–203.

Yang, Y., Carbonetto, P., Gerard, D., and Stephens, M. (2024). Improved methods for empirical Bayes multivariate multiple testing and effect size estimation. Available at http://arxiv.org/abs/2406.08784.

ANNEXURE

A-1. An MCMC method using Gibbs sampling and TMCMC

A-1.1. Full conditionals

Full conditional of $\boldsymbol{H}_k$

First observe that for $k = 0, 1$, the full conditional of $\boldsymbol{H}_k$ is given by

$$[\boldsymbol{H}_k|\cdots] \sim DP\left(\alpha_H + n_{\cdot k}, \frac{\alpha_H \tilde{\boldsymbol{H}} + \sum_{s=1}^{S} n_{sk}\delta_{\boldsymbol{\eta}_s}}{\alpha_H + n_{\cdot k}}\right), \tag{34}$$

where $n_{sk} = \#\{r \in \{1, \ldots, R_k\} : \boldsymbol{\xi}_{rk} = \boldsymbol{\eta}_s\}$ and $n_{\cdot k} = \sum_{s=1}^{S} n_{sk}$.

Full conditional of $\boldsymbol{G}_{0,jk}$

Similarly, the full conditional of $\boldsymbol{G}_{0,jk}$ is given, for $j = 1, \ldots, J$ and $k = 0, 1$, by

$$[\boldsymbol{G}_{0,jk}|\cdots] \sim DP\left(\alpha_{G_0,k} + n_{\cdot jk}, \frac{\alpha_{G_0,k}\boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk}\delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k} + n_{\cdot jk}}\right), \tag{35}$$

where $n_{ljk} = \#\{(t, i) \in \{1, \ldots, \tau_{ijk}\} \times \{1, \ldots, N_k\} : \boldsymbol{\phi}_{tijk} = \boldsymbol{\xi}_{lk}\}$ and $n_{\cdot jk} = \sum_{l=1}^{R_k} n_{ljk}$.

The full conditionals of $\boldsymbol{H}_k$ and $\boldsymbol{G}_{0,jk}$ given by (34) and (35) indicate generating the infinite-dimensional random probability measures using Sethuraman's characterization of Dirichlet processes (see Sethuraman (1994)). However, in our case, forming the infinite-dimensional Sethuraman's construction is not necessary; instead, it will be required to simulate from the random probability measures having distributions (34) and (35). Such simulations are possible using the retrospective method (see Papaspiliopoulos and Roberts (2008)) which avoids dealing with infinitely many objects.

Full conditional of $\boldsymbol{p}_{mijk}$

The associated Polya urn distribution of $\boldsymbol{p}_{mijk}$ given $\boldsymbol{P}_{Mijk}\backslash\{\boldsymbol{p}_{mijk}\}$, derived by marginalizing over $\boldsymbol{G}_{ijk}$, is the following:

$$[\boldsymbol{p}_{mijk}|\boldsymbol{P}_{Mijk}\backslash\{\boldsymbol{p}_{mijk}\}] = \frac{\alpha_{G,ik}}{\alpha_{G,ik} + M - 1}\boldsymbol{G}_{0,jk}(\boldsymbol{p}_{mijk}) + \frac{1}{\alpha_{G,ik} + M - 1}\sum_{m2\neq m=1}^{M}\delta_{\boldsymbol{p}_{m2ijk}}(\boldsymbol{p}_{mijk}) \tag{36}$$

where $M_{tijk} = \#\{m2 \in \{1, \ldots, M\}\backslash\{m\} : \boldsymbol{p}_{m2ijk} = \boldsymbol{\phi}_{tijk}\}$ and $\delta_{\boldsymbol{\phi}_{tijk}}(\cdot)$ denotes point mass at $\boldsymbol{\phi}_{tijk}$.

Given $z_{ijk} = m$, on combining the Polya urn distribution with the likelihood $\prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr})$ we obtain the following full conditional of $\boldsymbol{p}_{mijk}$:

$$[\boldsymbol{p}_{mijk}|\cdots] \propto \alpha_{G,ik}\prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr})\boldsymbol{G}_{0,jk}(\boldsymbol{p}_{mijk}) + \sum_{t=1}^{\tau_{ijk}} M_{tijk}\prod_{r=1}^{L} f(x_{ijkr}|\phi_{tijkr})\delta_{\boldsymbol{\phi}_{tijk}}(\boldsymbol{p}_{mijk}). \tag{37}$$

Note that in (37), $\boldsymbol{G}_{0,jk}$, drawn from (35), is not available in closed form and only admits the form dictated by Sethuraman's construction, given, almost surely, by

$$\boldsymbol{G}_{0,jk} = \sum_{l=1}^{\infty} \tilde{p}_l \delta_{\tilde{\boldsymbol{\xi}}_{ljk}}, \tag{38}$$

where $\tilde{p}_1 = V_1$, $\tilde{p}_l = V_l \prod_{s<l}(1-V_s)$, for $l \geq 2$, with $V_1, V_2, \ldots \overset{iid}{\sim} \text{Beta}\,(\alpha_{G_0,k} + n_{\cdot jk}, 1)$, and for $l = 1, 2, \ldots$, $\tilde{\boldsymbol{\xi}}_{ljk} \overset{iid}{\sim} \frac{\alpha_{G_0,k}\boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk}\delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k}+n_{\cdot jk}}$.

In (37), the posterior proportional to $\prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr})\boldsymbol{G}_{0,jk}\,(\boldsymbol{p}_{mijk})$, which we denote by $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$, is the discrete distribution that puts mass $C_{ijk}\tilde{p}_t \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{tjkr})$ to the point $\tilde{\boldsymbol{\xi}}_{tjk}$, for $t = 1, 2, \ldots$, where

$$C_{ijk} = \left(\sum_{t=1}^{\infty} \tilde{p}_t \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{tjkr})\right)^{-1} \tag{39}$$

is the normalizing constant. Combining these with (37) it follows that

$$[\boldsymbol{p}_{mijk}|\cdots] = \alpha_{G,ik}\bar{C}C_{ijk}^{-1}[\boldsymbol{G}_{0,jk}\,(\boldsymbol{p}_{mijk})\,|\boldsymbol{X}_{ijk}] + \bar{C}\sum_{t=1}^{\tau_{ijk}} M_{tijk} \prod_{r=1}^{L} f(x_{ijkr}|\phi_{tijkr})\delta_{\boldsymbol{\phi}_{tijk}}\,(\boldsymbol{p}_{mijk})\,, \tag{40}$$

where

$$\bar{C} = \left[\alpha_{G,ik}C_{jk}^{-1} + \sum_{t=1}^{\tau_{ijk}} M_{tijk} \prod_{r=1}^{L_j} f(x_{ijkr}|\phi_{tijkr})\right]^{-1}$$

is the normalizing constant of $[\boldsymbol{p}_{mijk}|\cdots]$.

A-1.2. Retrospective sampling methods

Retrospective method for simulating from $[\boldsymbol{p}_{mijk}|\cdots]$

From (40) it follows that, to draw from $[\boldsymbol{p}_{mijk}|\cdots]$, it is required to simulate from $[\boldsymbol{G}_{0,jk}\,(\boldsymbol{p}_{mijk})\,|\boldsymbol{X}_{ijk}]$ with probability proportional to C_{ijk}^{-1}. However, since C_{ijk} involves an infinite series, its calculation is infeasible. The same issue also prevents the traditional simulation methods to draw from the discrete distribution $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$. In this case, the retrospective sampling method proposed in Section 3.5 of Papaspiliopoulos and Roberts (2008) is the appropriate method for our purpose. We first briefly discuss the role of such method in simulating from $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$, and then argue that a by-product of the method can be used to estimate C_{ijk} arbitrarily accurately.

Retrospective method to draw from $[\boldsymbol{G}_{0,jk}\,(\boldsymbol{p}_{mijk})\,|\boldsymbol{X}_{ijk}]$

Note that the retrospective method requires $\prod_{r=1}^{L} f(x_{ijkr}|\phi_{tijkr})$ in our case to be uniformly bounded for all $\boldsymbol{\phi}_{tijk}$, which holds in our case, as $f(x_{ijkr}|\phi_{tjkr})$ represents the Bernoulli distribution, which is bounded above by 1. We briefly describe the method as follows. Let

$$c_\ell(K) = \sum_{a=1}^{K} \tilde{p}_a \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{ajkr}) \tag{41}$$

and

$$c_u(K) = c_\ell(K) + (1 - \sum_{a=1}^{K} \tilde{p}_a). \tag{42}$$

Let us also define $\breve{p}_{\ell,a}(K) = \tilde{p}_a \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{ajkr})/c_\ell(K)$ and $\breve{p}_{u,a}(K) = \tilde{p}_a \prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{ajkr})/c_u(K)$. To simulate from $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$ we first generate $U \sim \text{Uniform}(0,1)$, and given U, choose $\tilde{\boldsymbol{\xi}}_{tjk}$ when

$$\sum_{a=1}^{t-1} \breve{p}_{u,a}(K) \leq U \leq \sum_{a=1}^{t} \breve{p}_{\ell,a}(K). \tag{43}$$

In fact, K needs to be increased and $\tilde{p}_t$ and $\prod_{r=1}^{L} f(x_{ijkr}|\tilde{\xi}_{ajkr})$ simulated retrospectively, till (43) is satisfied for some $t \leq K$.

Retrospective method for estimating C_{ijk} arbitrarily accurately

By choosing K to be large enough, the quantities $c_\ell(K)$ and $c_u(K)$ given by (41) and (42), respectively, can be made arbitrarily close. In other words, for any $\epsilon > 0$, there exists $K_0 \geq 1$ such that $|c_\ell(K) - c_u(K)| < \epsilon$, for $K \geq K_0$. Thus, for any such $K \geq K_0$, one may approximate C_{ijk} with $[c_\ell(K)]^{-1}$. In practice, it is only required to simulate $\tilde{U} \sim \text{Uniform}(0,1)$ and simulate from $[\boldsymbol{G}_{0,jk}(\boldsymbol{p}_{mijk})|\boldsymbol{X}_{ijk}]$ if $\tilde{U} \leq \bar{C}C_{ijk}^{-1}$. For sufficiently small ϵ and for finite number of simulations, it will generally hold that $\tilde{U} \leq \bar{C}C_{ijk}^{-1}$ if and only if $\tilde{U} \leq \bar{C}_\epsilon c_\ell(K)$, for $K \geq K_0$, where

$$\bar{C}_\epsilon = \left[c_\ell^{-1}(K) + \sum_{t=1}^{\tau_{ijk}} M_{tijk} \prod_{r=1}^{L} f(x_{ijkr}|\phi_{tijkr}) \right]^{-1}.$$

Retrospective method to simulate from $\frac{\alpha_{G_0,k}\boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk}\delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k}+n_{\cdot jk}}$

The retrospective simulation method requires simulation of $\tilde{\boldsymbol{\xi}}_{ljk} \overset{iid}{\sim} \frac{\alpha_{G_0,k}\boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk}\delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k}+n_{\cdot jk}}$, for $l = 1, 2, \ldots$. This requires simulation from $\boldsymbol{H}_k$ with probability proportional to $\alpha_{G_0,k}$. For this, we first simulate $U \sim \text{Uniform}(0,1)$. We then simulate a realization from $\boldsymbol{H}_k$ after generating $\boldsymbol{H}_k$ from the Dirichlet process given by (34). Note that we do not have to generate the entire random probability measure $\boldsymbol{H}_k$ for this; we only need to generate as many realizations $\boldsymbol{\eta}_{lk}^*$'s from $\frac{\alpha_H \tilde{\boldsymbol{H}} + \sum_{s=1}^{S} n_{sk}\delta_{\boldsymbol{\eta}_s}}{\alpha_H + n_{\cdot k}}$ and as many $p_{lk}^* = V_{lk}^* \prod_{s<l}(1 - V_{lk}^*)$; $l = 1, 2 \ldots$, with $p_{1k}^* = V_{1k}^*$, with $V_{lk}^* \overset{iid}{\sim} \text{Beta}(\alpha_H + n_{\cdot k}, 1)$, as required to satisfy $\sum_{l=1}^{t-1} p_{lk}^* < U \leq \sum_{l=1}^{t} p_{lk}^*$, for some $t \geq 1$ (with $p_0^* = 0$). We then report $\tilde{\boldsymbol{\xi}}_{1jk} = \boldsymbol{\eta}_{tk}^*$ with probability proportional to $\alpha_{G_0,k}$ and $\tilde{\boldsymbol{\xi}}_{1jk} = \boldsymbol{\xi}_{\tilde{l}jk}$ with probability proportional to $n_{\tilde{l}jk}$, for $\tilde{l} \in \{1, \ldots, R_k\}$. We repeat this procedure for generating $\boldsymbol{\xi}_{ljk}$; $l \geq 2$, by sequentially augmenting the existing simulations of $\boldsymbol{\eta}_{lk}^*$'s and p_{lk}^*'s with new draws from $\tilde{\boldsymbol{H}}$ and $\text{Beta}(\alpha_H + n_{\cdot k}, 1)$, if needed. Indeed, note that for augmentation of p_{lk}^*'s, only extra V_{lk}^*'s need to be generated from $\text{Beta}(\alpha_H + n_{\cdot k}, 1)$.

A-1.3. Updating the allocation and proportion variables

Updating procedure for z_{ijk} and $\boldsymbol{p}_{mijk}$

The full conditional of z_{ijk} is given by the following:

$$[z_{ijk} = m|\cdots] \propto \pi_{mijk} \prod_{r=1}^{L_j} f\left(\boldsymbol{x}_{ijkr}|p_{mijkr}\right); \tag{44}$$

for $m = 1, \ldots, M$.

Recall that we have devised a method of simulating from the full conditional of $\boldsymbol{p}_{mijk}$ given the data and the remaining variables. For our convenience, we re-formulate the full conditional in terms of the dishes $\boldsymbol{\phi}_{tjk}$ and the indicators of the dishes, which we denote by t_{mijk}, where $t_{mijk} = t$ if and only if $p_{mijk} = \phi_{tijk}$; $t = 1, \ldots, \tau_{ijk}$.

Now let $\tau_{ijk}^{(m)}$ denote the number of elements in $\boldsymbol{P}_{Mijk} \setminus \{\boldsymbol{p}_{mijk}\}$ that arose from $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$. Also let $\boldsymbol{\phi}^{m*}_{tijk} = \left\{\phi^{m*}_{tijkr};\ r = 1, \ldots, L\right\};\ t = 1, \ldots, \tau_{ijk}^{(m)}$ denote the parameter vectors arising from $[\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]$. Further, let $\boldsymbol{\phi}^{m*}_{tijk}$ occur M_{mtijk} times.

Then we update t_{mijk} using Gibbs steps, where the full conditional distribution of t_{mijk} is given by

$$[t_{mijk} = t|\cdots] \propto \begin{cases} q^*_{t,mijk} & \text{if} \quad t = 1, \ldots, \tau_{ijk}^{(m)}; \\ q_{0,mijk} & \text{if} \quad t = \tau_{ijk}^{(m)} + 1, \end{cases} \tag{45}$$

where

$$q_{0,mijk} = \alpha_{G,ik} C^{-1}_{ijk}; \tag{46}$$

$$q^*_{t,mijk} = M_{mtijk} \prod_{r=1}^{L_j} \left\{\phi^{m*}_{tijkr}\right\}^{n_{1mijkr}} \left\{1 - \phi^{m*}_{tijkr}\right\}^{n_{2mijkr}}. \tag{47}$$

In (46) and (47), n_{1mijkr} and n_{2mijkr} denote the number of "a" and "A" alleles, respectively, at the r-th locus of the j-th gene of the i-th individual, associated with the m-th mixture component. In other words, $n_{1mijkr} = x^1_{ijkr} + x^2_{ijkr}$ and $n_{2mijkr} = 2 - \left(x^1_{ijkr} + x^2_{ijkr}\right)$.

Let $n^*_{1tijkr} = \sum_{m:t_{mijk}=t} n_{1mijkr}$ and $n^*_{2tijkr} = \sum_{m:t_{mijk}=t} n_{2mijkr}$. Then, for $t = 1, \ldots, \tau_{ijk};\ r = 1, \ldots, L_j;\ j = 1, \ldots, J$ and $k = 0, 1$, update $\boldsymbol{\phi}^*_{tijk}$ by simulating from its full conditional distribution, given by

$$[\boldsymbol{\phi}^*_{tijk}|\cdots] \sim [\boldsymbol{G}_{0,jk}|\boldsymbol{X}_{ijk}]. \tag{48}$$

The above simulation from $[\boldsymbol{\phi}^*_{tijk}|\cdots]$ is to be carried out by the retrospective method as discussed above.

A-1.4. Updating the missing data

Updating $\tilde{\boldsymbol{Y}}_{ijk}$

From (4) it follows that

$$[\tilde{\boldsymbol{Y}}_{ijk}|z_{ijk}] = \prod_{r=L_j+1}^{L} f\left(\boldsymbol{y}_{ijkr}|p_{z_{ijk}ijkr}\right). \tag{49}$$

Hence, given the other unknowns, $\tilde{\boldsymbol{Y}}_{ijk}$ can be updated by simply simulating from the Bernoulli distributions given by (49).

A-1.5. Relevant factor aggregations for updating the fixed-dimensional parameters

Relevant factors for updating μ_G and $\boldsymbol{\beta}_G$

Let

$$\mathcal{L}_G(\mu_G, \boldsymbol{\beta}_G) = \prod_{k=0}^{1} \prod_{i=1}^{N_k} \prod_{j=1}^{J} \prod_{m=2}^{M} [\boldsymbol{p}_{mijk} | \boldsymbol{p}_{lijk}; l < m],$$

where $[\boldsymbol{p}_{mijk} | \boldsymbol{p}_{lijk}; l < m]$ is given by (15). Let $\pi_G(\mu_G, \boldsymbol{\beta}_G)$ denote the prior on $(\mu_G, \boldsymbol{\beta}_G)$. Note that $\pi_G(\mu_G, \boldsymbol{\beta}_G)\mathcal{L}_G(\mu_G, \boldsymbol{\beta}_G)$ is the product of the only factors in the joint model consisting of μ_G and $\boldsymbol{\beta}_G$.

Relevant factors for updating μ_{G_0} and $\boldsymbol{\beta}_{G_0}$

Now let

$$\mathcal{L}_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0}) == \prod_{k=0}^{1} \prod_{i=1}^{N_k} \prod_{j=1}^{J} \prod_{t=2}^{\tau_{ijk}} [\boldsymbol{\phi}_{tijk} | \boldsymbol{\phi}_{lijk}; l < t],$$

where $[\boldsymbol{\phi}_{tijk} | \boldsymbol{\phi}_{lijk}; l < t]$ is given by (16).

Let $\pi_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0})$ denote the prior on $(\mu_{G_0}, \boldsymbol{\beta}_{G_0})$. Then $\pi_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0})\mathcal{L}_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0})$ is the functional form associated with μ_{G_0} and $\boldsymbol{\beta}_{G_0}$.

Relevant factors for updating μ_H and $\boldsymbol{\beta}_H$

Finally, we let

$$\mathcal{L}_H(\mu_H, \boldsymbol{\beta}_H) = \prod_{k=0}^{1} \prod_{s=2}^{R_k} [\boldsymbol{\xi}_{sk} | \boldsymbol{\xi}_{lk}; l < s],$$

where $[\boldsymbol{\xi}_{sk} | \boldsymbol{\xi}_{lk}; l < s]$ is given y (17).

Let $\pi_H(\mu_H, \boldsymbol{\beta}_H)$ be the prior on $(\mu_H, \boldsymbol{\beta}_H)$. Then $\pi_H(\mu_H, \boldsymbol{\beta}_H)\mathcal{L}_H(\mu_H, \boldsymbol{\beta}_H)$ is the functional form to be considered for updating μ_H and $\boldsymbol{\beta}_H$.

A-1.6. Mixture of additive and multiplicative TMCMC for updating the fixed-dimensional parameters in a single block

We shall update all the parameters μ_G, $\boldsymbol{\beta}_G$, μ_{G_0}, $\boldsymbol{\beta}_{G_0}$, μ_H and $\boldsymbol{\beta}_H$ using a mixture of additive and multiplicative TMCMC, where all the aforementioned parameters are given either the additive move or the multiplicative move with equal probability, and where the acceptance ratio will be calculated by evaluating the functional form

$$\pi_G(\mu_G, \boldsymbol{\beta}_G)\mathcal{L}_G(\mu_G, \boldsymbol{\beta}_G) \times \pi_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0})\mathcal{L}_{G_0}(\mu_{G_0}, \boldsymbol{\beta}_{G_0}) \times \pi_H(\mu_H, \boldsymbol{\beta}_H)\mathcal{L}_H(\mu_H, \boldsymbol{\beta}_H)$$

at the numerator and the denominator corresponding to the proposed and the current values of μ_G, $\boldsymbol{\beta}_G$, μ_{G_0}, $\boldsymbol{\beta}_{G_0}$, μ_H and $\boldsymbol{\beta}_H$, with all other unknowns held fixed at their current values, multiplied by an appropriate Jacobian whenever the multiplicative move is chosen. For details regarding mixture of additive and multiplicative TMCMC, see Dey and Bhattacharya (2017).

A-2. A parallel algorithm for implementing our MCMC procedure

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. Furthermore, the parameters of the HDP are also amenable to efficient parallelization. The details are as follows.

(1) (a) In processes numbered 0 and 1, simultaneously obtain the set of distinct elements $\boldsymbol{\Xi}_{R_k,k}$; $k = 0, 1$, from $\{\phi_{tijk}; t = 1, \ldots, \tau_{ijk}; i = 1, \ldots, N_k; j = 1, \ldots, J\}$; $k = 0, 1$.

(b) Communicate $\boldsymbol{\Xi}_{R_k,k}$; $k = 0, 1$, to all the processes.

(2) (a) In process 0, obtain the set of distinct elements $\boldsymbol{\zeta}_S$ from $\{\boldsymbol{\Xi}_{R_k,k}; k = 0, 1\}$.

(b) Communicate $\boldsymbol{\zeta}_S$ to all the processes.

(3) In processes numbered 0 and 1, do the following in parallel for $k = 0, 1$:

(a) Simulate, following the retrospective method. $\boldsymbol{\eta}^*_{lk} \overset{iid}{\sim} \frac{\alpha_H \tilde{\boldsymbol{H}} + \sum_{s=1}^{S} n_{sk} \delta_{\boldsymbol{\eta}_s}}{\alpha_H + n_{\cdot k}}$; $l = 1, 2, \ldots, \mathcal{L}$, for sufficiently large $\mathcal{L}$.

(b) Communicate the simulated values to all the processes.

(3) Split $\{(j, k) : \ j = 1, \ldots, J; \ k = 0, 1\}$ in the available parallel processes.

(a) For each (j, k), simulate, following the retrospective method. $\tilde{\boldsymbol{\xi}}_{ljk} \overset{iid}{\sim} \frac{\alpha_{G_0,k} \boldsymbol{H}_k + \sum_{l=1}^{R_k} n_{ljk} \delta_{\boldsymbol{\xi}_{lk}}}{\alpha_{G_0,k} + n_{\cdot jk}}$; $l = 1, 2, \ldots, \mathcal{L}$.

(b) Communicate the simulated values to all the processes.

(4) (a) Split the triplets $\{(i, j, k) : \ i = 1, \ldots, N_k; \ j = 1, \ldots, J; \ k = 0, 1\}$ in the available parallel processes sequentially into

$$\mathcal{T}_1 = \{(i, j, 0) : \ i = 1, \ldots, N_0; \ j = 1, \ldots, J\}$$

and

$$\mathcal{T}_2 = \{(i, j, 1) : \ i = 1, \ldots, N_1; \ j = 1, \ldots, J\}.$$

(b) Then parallelise updating of the mixtures associated with $\mathcal{T}_1$, followed by those of $\mathcal{T}_2$.

(c) If, for any (i, j, k), retrospective simulation from $[\boldsymbol{G}_{0,jk} | \boldsymbol{X}_{ijk}]$ requires more than $\mathcal{L}$ simulations of $\tilde{\boldsymbol{\xi}}_{ljk}$ in step (3) (a), then increase $\mathcal{L}$ to $\mathcal{L}^*$, and

(i) For $k = 0, 1$, augment the simulations of $\{\boldsymbol{\eta}^*_{lk}; l = 1, \ldots, \mathcal{L}\}$ with new simulations $\{\boldsymbol{\eta}^*_{lk}; l = \mathcal{L} + 1, \ldots, \mathcal{L}^*\}$.

(ii) For $j = 1, \ldots, J$ and for $k = 0, 1$, augment the simulations of $\left\{\tilde{\boldsymbol{\xi}}_{ljk}; l = 1, \ldots, \mathcal{L}\right\}$ with new simulations $\left\{\tilde{\boldsymbol{\xi}}_{ljk}; l = \mathcal{L} + 1, \ldots, \mathcal{L}^*\right\}$.

(iii) Repeat (4) (a) and (4) (b).

(5) During each MCMC iteration, for each (i, j, k) in each available parallel processor, update the allocation variables z_{ijk}, the proportions $\boldsymbol{p}_{mijk}$; $m = 1, \ldots, M$, and the missing data $\tilde{\boldsymbol{Y}}_{ijk}$, using the methods proposed in Sections A-1.3 and A-1.4.

(6) Communicate the results of updating in (4) and (5) to all the processes.

(7) (a) During each MCMC iteration, update the parameters μ_G, $\boldsymbol{\beta}_G$, μ_{G_0}, $\boldsymbol{\beta}_{G_0}$, μ_H and $\boldsymbol{\beta}_H$ in a single block using a mixture of additive and multiplicative TMCMC, as proposed in Section A-1.6, in process number 0.

(b) Communicate the updated results to all the processes.

A-3. Simulation studies

For simulation studies, we first generate realistic biological data for stratified population with known gene-environment interaction from the GENS2 software of Pinelli *et al.* (2012). To this data, we then apply our model and methodologies in an effort to detect gene-environment interaction effects that are present in the data. We consider simulation studies in 5 different true model set-ups: (a) presence of gene-gene and gene-environment interaction, (b) absence of genetic or gene-environmental interaction 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 effects of the individual genes, gene-gene and gene-environment interactions, and the number of sub-populations. In all our applications, we set $M = 30$, $\nu_1 = \nu_2 = 1$, so that $\tilde{\boldsymbol{H}}$ is the uniform distribution on $[0, 1]$. We set $\alpha_{G,ik} = 0.1 \times \exp(100 + \mu_G + \beta_G E_{ik})$, $\alpha_{G_0,k} = 0.1 \times \exp\left(100 + \mu_{G_0} + \beta_{G_0}\bar{E}_k\right)$ and $\alpha_H = 0.1 \times \exp\left(100 + \mu_H + \beta_H\bar{\bar{E}}\right)$, where we assumed $\mu_G, \mu_{G_0}, \mu_H \overset{iid}{\sim} U(0, 1)$ and $\beta_G, \beta_{G_0}, \beta_H \overset{iid}{\sim} U(-1, 1)$. This structure ensured adequate number of sub-populations and satisfactory mixing of MCMC.

A-3.1.First simulation study: presence of gene-gene and gene-environment interaction

A-3.1.1.Data description

As in Bhattacharya and Bhattacharya (2020) we consider two genetic factors as allowed by GENS2 and simulated 5 data sets with gene-gene and gene-environment interaction with a one-dimensional environmental variable, associated with 5 sub-populations. One of the genes consists of 1084 SNPs and another has 1206 SNPs, with one disease pre-disposing locus (DPL) at each gene. There are 113 individuals in each of the 5 data sets, from which we selected a total of 100 individuals without replacement with probabilities assigned to the 5 data sets being $(0.1, 0.4, 0.2, 0.15, 0.15)$. Our final dataset consists of 46 cases and 54 controls. Since, in our case, the environmental variable is one-dimensional, $d = 1$.

A-3.1.2.Model implementation

We implemented our parallel MCMC algorithm on 50 cores in a 64-bit VMware with 64-bit physical cores, each running at 2793.269 MHz. Our code is written in C in conjunction with the Message Passing Interface (MPI) protocol for parallelisation.

The total time taken to implement $30,000$ MCMC iterations, where the first $10,000$ are discarded as burn-in, is approximately 20 hours. We assessed convergence informally with trace plots, which indicated adequate mixing properties of our algorithm.

A-3.1.3.Specifications of the thresholds ε's using null distributions

Following the method outlined in Section 4.2.4 and setting M to be 30, we obtain $\varepsilon_{d^*} = 0.200$, $\varepsilon_{\hat{d}_1} = 0.167$, $\varepsilon_{\hat{d}_2} = 0.167$, $\varepsilon_{d^*_E} = 0.250$, $\varepsilon_{d^*_{E,1}} = 0.185$, $\varepsilon_{d^*_{E,2}} = 0.173$, $\varepsilon_{\beta_G} = 0.874$, $\varepsilon_{\beta_{G_0}} = 0.128$, $\varepsilon_{\beta_H} = 0.219$.

A-3.1.4.Results of fitting our model

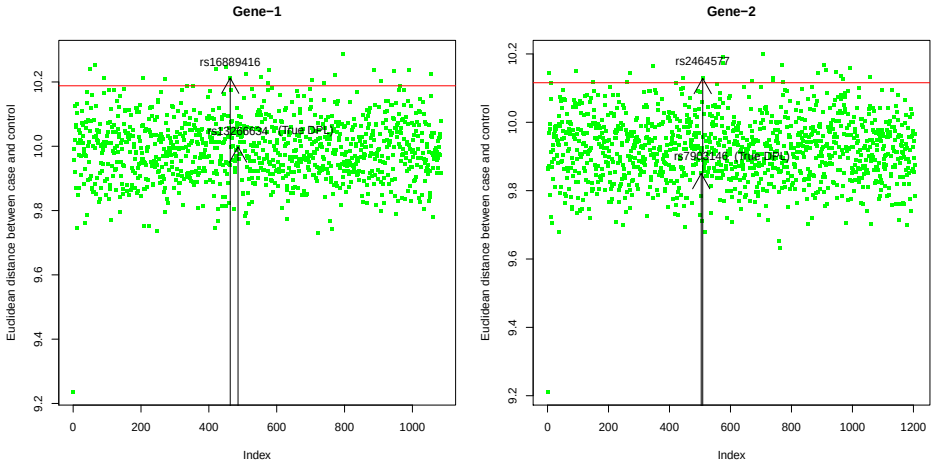
The posterior probabilities $P\left(d^* < \varepsilon_{d^*}|\text{Data}\right)$, $P\left(\hat{d}_1 < \varepsilon_{\hat{d}_1}|\text{Data}\right)$ and $P\left(\hat{d}_2 < \varepsilon_{\hat{d}_2}|\right.$ Data) empirically obtained from $20,000$ MCMC samples, turned out to be 0.378, 0.317 and 0.324, respectively. Hence, H_{0,d^*}, $H_{0,\hat{d}_1}$ and $H_{0,\hat{d}_2}$ are rejected, suggesting the influence of significant genetic effects in the case-control study.

However, $P\left(d^*_E < \varepsilon_{d^*_E}|\text{Data}\right)$, $P\left(\hat{d}_{E,1} < \varepsilon_{\hat{d}_{E,1}}|\text{Data}\right)$ and $P\left(\hat{d}_{E,2} < \varepsilon_{\hat{d}_{E,2}}|\text{Data}\right)$ are given, approximately, by 0.558, 0.561 and 0.550, respectively, which seem to contradict the results of the clustering based hypothesis tests. This can be explained as follows. Since $\boldsymbol{G}_{0,jk}$ are discrete, the parameters $\boldsymbol{p}_{mijk}$, even if generated from $\boldsymbol{G}_{0,jk}$, coincide with positive probability, so that the effective dimensionalities of logit $\left(\bar{P}_{Mi_0jk=0}\right)$ and logit $\left(\bar{P}_{Mi_1jk=1}\right)$ are drastically reduced, so that the Euclidean distance between these two vectors is substantially small. As such, the Euclidean distance fails to reject the null even if it is false. As noted in Bhattacharya and Bhattacharya (2020), even the clustering metric in this scenario is not completely satisfactory since this involves clustering distance between two empirically obtained central clusterings which may not be very accurate unless the sample sizes for case and control are very large. However, compared to the Euclidean distance, the clustering metric turns out to be far more reliable.

To check the influence of the environmental variable on the genes we compute the posterior probabilities $P\left(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}\right)$, $P\left(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}\right)$ and $P\left(|\beta_H| < \varepsilon_{\beta_H}|\text{Data}\right)$. The probabilities turned out to be 0.544, 0.550 and 0.191, respectively, showing that β_H is very significant. That is, the environmental variable has a significant overall effect on the genes.

The posterior probabilities of no gene-gene interactions for the controls and cases, showed the prominence of several gene-gene interactions in both control and case groups. As to be expected, in the case group, more instances of gene-gene interactions turned out to be significant compared to the control group.

Also, encouragingly, The posteriors of the number of sub-populations gave high prob-



(a) Index plot for the first gene (b) Index plot for the second gene.

Figure A-1: Presence of gene-gene and gene-environment interaction: Plots of the Euclidean distances $\left\{d_j^r\left(\textbf{logit}\left(\boldsymbol{p}_{i_0jk=0}^r\right), \textbf{logit}\left(\boldsymbol{p}_{i_1jk=1}^r\right)\right);\ r=1,\ldots,L_j\right\}$ against the indices of the loci, for $j=1$ (panel (a)) and $j=2$ (panel (b)).

abilities to 5, the true number of sub-populations.

A-3.1.5.Detection of DPL

The correct positions of the DPL, provided by GENS2, are $rs13266634$ and $rs7903146$, for the first and second gene respectively. Due to the LD effects implied by the highly correlated structure of our current HDP based model, the actual DPL are difficult to locate. Notably, our model is considerably more structured than those of Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), and any inappropriate dependence structure would render the task of DPL finding far more difficult than our previous models. Nevertheless, we demonstrate that our HDP model can detect DPLs with more precision compared to our previous matrix-normal-inverse-Wishart model for gene-environment interactions.

Following Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2020), and writing $\boldsymbol{p}_{ijk}^r = \{p_{mijkr} : m = 1,\ldots,M\}$, we declare the r-th locus of the j-th gene as disease pre-disposing if, for the r-th locus, the Euclidean distance $d_j^r\left(\text{logit}\left(\boldsymbol{p}_{i_0jk=0}^r\right), \text{logit}\left(\boldsymbol{p}_{i_1jk=1}^r\right)\right)$, between logit $\left(\boldsymbol{p}_{i_0jk=0}^r\right)$ and logit $\left(\boldsymbol{p}_{i_1jk=1}^r\right)$, is significantly larger than $d_j^{r2}\left(\boldsymbol{p}_{i_0jk=0}^{r2}, \boldsymbol{p}_{i_1jk=1}^{r2}\right)$, for $r2 \neq r$. We adopt the graphical method as in our previous works. The red, horizontal lines in the panels of Figure A-1 represent the cut-off value such that the points above the horizontal line are those with the highest 2% Euclidean distances. The actual DPLs of the two genes, as well as their nearest neighbours with Euclidean distances on or above the red, horizontal lines, are shown in the figures. That even such small sets of SNPs with highest 2% Euclidean distances consist of close neighbours of the true DPLs, is quite encouraging. Observe that the DPL detection is more precise for the second gene in the sense that the closest neighbour of the actual DPL above the red, horizontal line is closer to the true DPL than for the first gene.

The above results on DPL detection is also a significant improvement over Bhattacharya and Bhattacharya (2020) where highest 10% Euclidean distances were considered, suggesting that our current HDP based model is more appropriate compared to our previous matrix-normal-inverse-Wishart model for gene-environment interaction.

A-3.2.Second simulation study: no genetic or environmental effect

Here we use the same case-control genotype data set as used by Bhattacharya and Bhattacharya (2018) in their second simulation study where genetic effects are absent, consisting of 49 cases and 51 controls and 5 sub-populations with the mixing proportions $(0.1, 0.4, 0.2, 0.15, 0.15)$. We use the same environmental data set generated in our first simulation study described in Section A-3.1, which is unrelated to this genotype data.

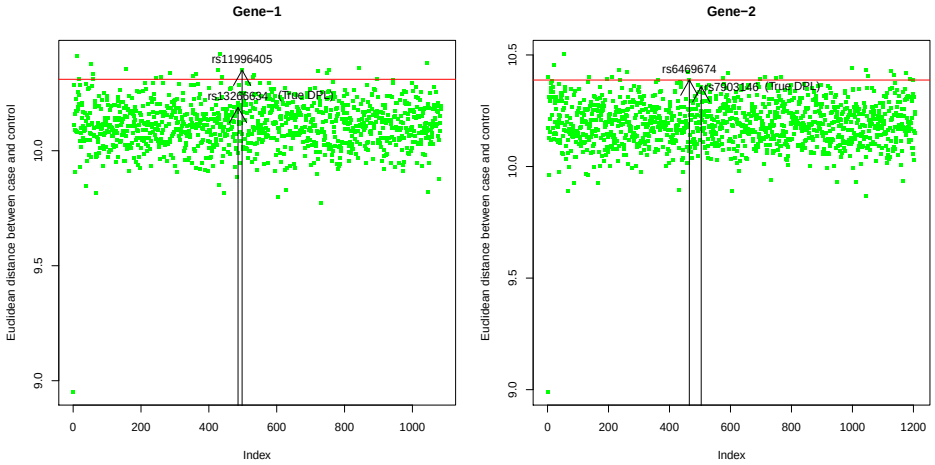
Here we obtain $P\left(d^* < \varepsilon_{d^*}|\text{Data}\right) \approx 0.407$. Although this does not cross the 0.5 benchmark, there is significant evidence in favour of the null, and falling short of 0.5 can be attributed to the slight deficiency of the distance between the two approximate central clusterings associated with case and control, as already discussed in the context of the first simulation study.

Also, in this study, $P\left(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}\right)$, $P\left(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}\right)$ and $P\left(|\beta_H| < \varepsilon_{\beta_H}|\right.$ Data) are given by 0.549, 0.550 and 0.649, respectively, suggesting insignificance of the effect of the environmental variable on gene-gene interaction. As noted in Bhattacharya and Bhattacharya (2020), however, it is not straightforward to test whether or not the environment is responsible for the case-control status. This is because we have modeled the genotype data conditionally on case-control instead of modeling the case-control status conditionally on the environmental variable. Bhattacharya and Bhattacharya (2020) use significance testing in a simple logistic regression framework to show insignificance of the environmental variable. As before, our model assigned high posterior probability to 5 sub-populations. Note that since there is no genetic effect in this study, the question of detecting DPLs does not arise here.

A-3.3.Third simulation study: absence of genetic and gene-gene interaction effects but presence of environmental effect

In this study we consider a case-control genotype data set simulated from GENS2 where case-control status depends only upon the environmental data. The number of cases here is 47 and the number of controls is 53. This is the same case-control genotype data set as used by Bhattacharya and Bhattacharya (2020) in their third simulation study.

In this case, we find that $P\left(d^* < \varepsilon_{d^*}|\text{Data}\right) \approx 0.400$, which provides reasonable evidence in favour of the null, even though the 0.5 benchmark is not crossed. Moreover, $P\left(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}\right) \approx 0.536$, $P\left(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}\right) \approx 0.518$ and $P\left(|\beta_H| < \varepsilon_{\beta_H}|\text{Data}\right) \approx 0.504$, suggesting that the environmental variable does not affect the genetic structure. Bhattacharya and Bhattacharya (2020) show by means AIC, in the context of simple logistic regression, that the best model consists of the marginal effects of the second gene and the environment. In conjunction with our HDP-based model which produces reasonable evidence in favour of accepting the hypothesis of no genetic effect, it may be possible to conclude that the environmental variable is responsible for the case-control status.



(a) Index plot for the first gene (b) Index plot for the second gene.

Figure A-2: Presence of genetic and gene-gene interaction effects but absence of environmental effect: Plots of the Euclidean distances $\left\{d_j^r\left(\textbf{logit}\left(\boldsymbol{p}_{i_0jk=0}^r\right), \textbf{logit}\left(\boldsymbol{p}_{i_1jk=1}^r\right)\right);\ r=1,\ldots,L_j\right\}$ against the indices of the loci, for $j=1$ (panel (a)) and $j=2$ (panel (b)).

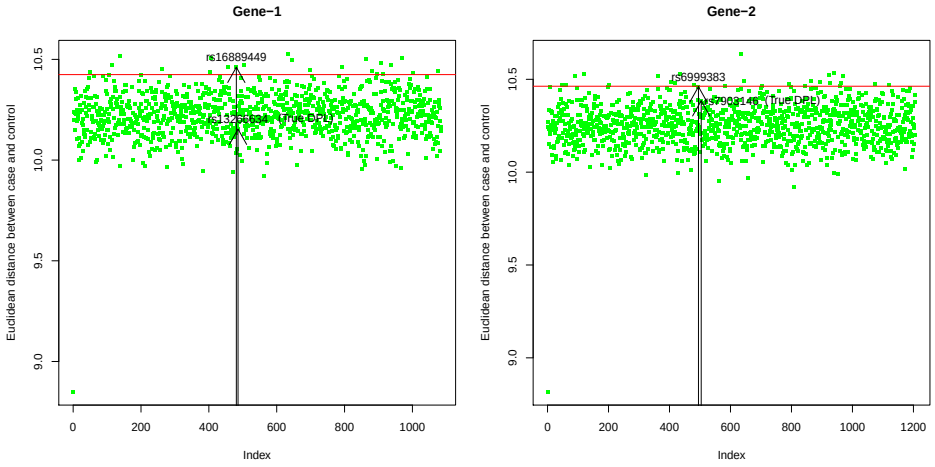
As before, 5 subpopulations get significant weight by our posterior distribution, and again, the question of DPL detection is irrelevant here since there is no genetic effect.

A-3.4.Fourth simulation study: presence of genetic and gene-gene interaction effects but absence of environmental effect

Here we use the same genotype data set as used by Bhattacharya and Bhattacharya (2018) in their first simulation study associated with genetic and gene-gene interaction effects, consisting of 41 cases and 59 controls and 5 sub-populations with the mixing proportions $(0.1, 0.4, 0.2, 0.15, 0.15)$. We use the same environmental data set generated in our first simulation study described in Section A-3.1, which is unrelated to this case-control genotype data.

Here we obtain $P\left(|\beta_G| < \varepsilon_{\beta_G}|\text{Data}\right) \approx 0.549$, $P\left(|\beta_{G_0}| < \varepsilon_{\beta_{G_0}}|\text{Data}\right) \approx 0.542$ and $P\left(|\beta_H| < \varepsilon_{\beta_H}|\text{Data}\right) \approx 0.552$, correctly suggesting insignificance of the environmental variable with respect to its effect on the genetic structure. Using logistic regression, Bhattacharya and Bhattacharya (2020) conclude that the environmental variable has no role to play in the case-control status. Furthermore, we obtain $P\left(d^* < \varepsilon_{d^*}|\text{Data}\right) \approx 0.390$, $P\left(\hat{d}_1 < \varepsilon_{\hat{d}_1}|\text{Data}\right) \approx 0.336$ $P\left(\hat{d}_2 < \varepsilon_{\hat{d}_2}|\text{Data}\right) \approx 0.324$. so that importance of genes is correctly indicated by our tests. Interestingly, study of the posterior probabilities of no gene-gene interactions for controls and cases showed no gene-gene interaction in the control group and only two (marginal) instances of gene-gene interaction among the cases.

Figure A-2 shows the plots of Euclidean distances between cases and controls for the loci of the two genes. In this case, Gene-1 has been located quite precisely, and for Gene-2 the Euclidean distance for even the true DPL is very close to the red, horizontal line, indicating encouraging performance.



(a) Index plot for the first gene (b) Index plot for the second gene.

Figure A-3: Independent and additive genetic and environmental effects: Plots of the Euclidean distances $\left\{d_j^r\left(\textbf{logit}\left(\boldsymbol{p}_{i_0jk=0}^r\right),\textbf{logit}\left(\boldsymbol{p}_{i_1jk=1}^r\right)\right);\ r=1,\ldots,L_j\right\}$ against the indices of the loci, for $j=1$ (panel (a)) and $j=2$ (panel (b)).

A-3.5.Fifth simulation study: independent and additive genetic and environmental effects

As in Bhattacharya and Bhattacharya (2020), we consider the situation where the genetic and environmental effects are independent of each other and additive; the data consists of 57 cases and 43 controls.

Note that, as in Bhattacharya and Bhattacharya (2020), in our current HDP-based Bayesian model also there is no provision for additivity of genetic and environmental effects. As such, it is not expected to capture the true data-generating mechanism accurately. Indeed, here we obtain $P\left(d^*<\varepsilon_{d^*}|\text{Data}\right)\approx 0.389$, $P\left(\hat{d}_1<\varepsilon_{\hat{d}_1}|\text{Data}\right)\approx 0.337$ and $P\left(\hat{d}_2<\varepsilon_{\hat{d}_2}|\text{Data}\right)\approx 0.331$, indicating significance of the genes. However, the test with d_E^* does not yield overwhelming evidence against the null. Our tests of gene-gene interaction indicated significant interactions for controls and particularly for cases. Also, $P\left(|\beta_G|<\varepsilon_{\beta_G}|\text{Data}\right)$, $P\left(|\beta_{G_0}|<\varepsilon_{\beta_{G_0}}|\text{Data}\right)$ and $P\left(|\beta_H|<\varepsilon_{\beta_H}|\text{Data}\right)$ are given, approximately, by 0.547, 0.550 and 0.367, the last value showing that the environmental variable does affect gene-gene interaction. The lack of the additivity provision in our model seems to have forced the gene-environment interaction in this case.

In spite of the lack of additivity of our model the Euclidean distances between cases and controls for the gene-wise SNPs are not adversely affected, and the actual DPLs are detected quite accurately; see Figure A-3. This brings forth the generality and usefulness of our nonparametric dependence structure. As before, 5 sub-populations received significant posterior probabilities.