Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 31-41

A Nonparametric Structure Based on Hierarchical Dirichlet Processes for Studying Gene-Gene and Gene-Environment Interactions

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: 15 June 2023; Revised: 01 July 2023; Accepted: 04 July 2023

Abstract

It has been established in many studies that genes interact in complex networks among themselves and with various environmental factors to cause diseases. In this article, we discuss how realistic statistical models for case-control genotype data can be developed using nonparametric Bayesian techniques founded on hierarchies of Dirichlet process based mixture models for studying such complex interaction structures. Suitable Bayesian hypothesis testing procedures need to be developed for uncovering the roles of genes, environment and their interactions in case-control studies. Empowered with an efficient TMCMC based parallelisable algorithm, application of our ideas to data simulated under five different setups of disease-gene-environment association as well as a real, Myocardial Infarction (MI) dataset yielded interesting results that not only agrees with the existing works in this area, but also gives some novel insights into the genetic interactions underlying the disease.

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

1. Introduction

Present day biomedical research is pointing towards the significance of interactions between genes and the environment in causing complex diseases. According to Hunter (2005), considering the contributions of genes and environment to a disease separately, ignoring their interactions, might lead to incorrect estimation of the disease proportion that is explained by these factors. The additive linear models or the logistic model based approaches, (see for example Ahn *et al.* (2013), Wen and Stephens (2014) and Liu, Ma and Amos (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 existing Bayesian techniques like BEAM, EpiBN study interaction by identifying the SNPs that influence the disease risk given particular allele combinations, ignoring the genes as functional units. In a nutshell, none of the existing methods, classical or Bayesian, attempts simultaneous modelling of the uncertainties associated with the genes as the functional units along with the interactions, both at SNP and gene level through unified statistical models.

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. The Bayesian semiparametric model proposed by Bhattacharya and Bhattacharya (2020 a) takes care of the

Corresponding Author: Durba Bhattacharya
E-mail: durba@sxccal.edu

above mentioned problems by proposing a model based on Dirichlet Processes (DP) and a hierarchical matrix-normal distribution, encapsulating the mechanism of dependence among genes under environmental effects with respect to genotype data arising out of a possibly stratified population.

As 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. 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). Application of our model and methods to five different simulation experiments for the validation purpose yielded quite encouraging results. Application to a real myocardial infarction (MI) case-control type dataset yielded results which broadly agree with the results reported in the literature, and also provided new and interesting insights into the mechanisms of 4 gene-gene and gene-environment interactions.

The rest of our paper is structured as follows. We introduce our HDP-based Bayesian nonparametric gene-gene and gene-environment interaction model in Section 2, and in Section 3 we extend the Bayesian hypothesis testing procedures proposed in Bhattacharya and Bhattacharya (2020 a) to learn about the roles of genes, environmental variables and their interactions in case-control studies, with respect to our current HDP model. In Section 4 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 section S-3 of the supplement in Bhattacharya and Bhattacharya (2020 b). In Section 5 we analyse the real MI dataset using our ideas, demonstrating quite interesting and insightful outcome. Finally, we summarize our work with concluding remarks in Section 6.

2. Bayesian nonparametric model based on hierarchies of Dirichlet process for gene-gene and gene-environment interactions

2.1. Case-control genotype data

For $s = 1, 2$ denoting the two chromosomes, let ${x^s}_{ijkr} = 1$ and ${x^s}_{ijkr} = 0$ indicate the presence and absence of the minor allele of the i-th individual belonging to the k-th group, 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 E_i denote a set of environmental variables associated with the i-th individual. We now proceed to model this case-control genotype and the environmental data using our Bayesian semiparametric model, described in the next few sections.

2.2. Mixture models based on Dirichlet processes

Let $x_{ijkr} = ({x^1}_{ijkr}, {x^2}_{ijkr})$ and $L = \max(L_1, \ldots, L_J)$. We assume that for every triplet (i, j, k), $X_{ijk} = (x_{ijk1}, \ldots, x_{ijkL})$ have the mixture distribution

$$[X_{ijk}] = \sum_{m=1}^{M} \pi_{mijk} \prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr}) \quad (1)$$

where $f(.|p_{mijkr})$ is a Bernoulli mass function given by:

$$f(x_{ijkr}|p_{mijkr}) = p_{mijkr}{}^{x^1_{ijkr}+x^2_{ijkr}}(1-p_{mijkr})^{2-(x^1_{ijkr}+x^2_{ijkr})} \quad (2)$$

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.

Allocation variables z_{ijk}, with probability distribution

$$[z_{ijk} = m] = \pi_{mijk}, \quad (3)$$

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

$$[X_{ijk}|z_{ijk}] = \prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr}) \quad (4)$$

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 $p_{mijk} = (p_{mijk1}, \ldots, p_{mijkL})$, we assume that

$$p_{1ijk}, p_{2ijk}, \ldots, p_{Mijk} \overset{iid}{\sim} G_{ijk}; \quad (5)$$

$$G_{ijk} \sim \text{DP}(\alpha_{G,ik}, G_{0,jk}) \quad (6)$$

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

$$\log(\alpha_{G,ik}) = \mu_G + \beta_G{}^T E_{ik}, \quad (7)$$

where E_{ik} is a d-dimensional vector of continuous environmental variable for the i-th individual in the k-th group, β_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 model the dependence between the genes and case-control status

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

$$G_{0,jk} \overset{iid}{\sim} DP(\alpha_{G_0 k}, H_k); j = 1, \ldots, J, \quad (8)$$

$$\text{where } \log(\alpha_{G_0,k}) = \mu_{G_0} + \beta_{G_0}{}^T \overline{E}_k, \quad (9)$$

$$\text{with } \overline{E}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} E_{ik} \quad (10)$$

We postulate the last level of hierarchy as

$$H_k \overset{iid}{\sim} \text{DP}(\alpha_H \widetilde{H});\ k = 0,1 \tag{11}$$

where $\log(\alpha_H) = \mu_H + \beta_H{}^T \bar{\bar{E}}$, (12)

with $\bar{\bar{E}} = \frac{\bar{E}_0 + \bar{E}_1}{2}$ (13)

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

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

Under $\widetilde{H}$, where $\nu_1, \nu_2 > 0$.

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). For detailed discussion on the dependence structure induced by our hdp-based model see Section 3 of Bhattacharya and Bhattacharya (2020 b).

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

3.1. Formulation of the tests and interpretation of their results

To test if genes have any effect on case-control, we formulate the following hypotheses:

$$H_{01}: h_{0j} = h_{1j};\ j = 1, \ldots, J, \tag{15}$$

versus

$$H_{11}: \text{not } H_{01}, \tag{16}$$

$$\text{where } h_{0j}(.) = \frac{1}{M}\sum_{m=1}^{M}\prod_{r=1}^{L_j} f(.\,|p_{mi_0jk=0}{}^r) \tag{17}$$

$$h_{1j}(.) = \frac{1}{M}\sum_{m=1}^{M}\prod_{r=1}^{L_j} f(.\,|p_{mi_1jk=1}{}^r) \tag{18}$$

In the above, for $k = 0, 1$, i_k is the index such that $P_{Mi_kjk} = \{p_{1i_kjk}, p_{2i_kjk}, \ldots, p_{Mi_kjk}\}$ is an appropriate measure of central tendency (see Section 4.2.1 of Bhattacharya and Bhattacharya (2020 b)) of $\{P_{Mijk} = \{p_{1ijk}, p_{2ijk}, \ldots, p_{Mijk}\};\ i = 1, \ldots, N_k$.

3.1.1. Bayesian test for the significance of the environmental variables

To check if the environmental variables are significant, we shall test the following:

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

$$H_{02l}: \beta_{Gl} = 0 \text{ versus } H_{12l}: \beta_{Gl} \neq 0, \tag{19}$$

$$H_{03l}: \beta_{G_0l} = 0 \text{ versus } H_{12l}: \beta_{G_0l} \neq 0, \tag{20}$$

and $H_{04l}: \beta_{Hl} = 0$ versus $H_{14l}: \beta_{Hl} \neq 0$. (21)

3.1.2. Bayesian test for significance of gene-gene interaction

In order to test for gene-gene interaction, it is necessary to first reasonably define a measure of gene-gene interaction influenced by environmental variables.

For our purpose, we first define

$$\overline{p}_{mijk} = \frac{\sum_{r=1}^{L_j} p_{mijkr}}{L_j} \tag{22}$$

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

$$C(i, j_1, j_2, k) = cov(\text{logit}(\overline{p}_{z_{ij_1k}ij_1k}), \text{logit}(\overline{p}_{z_{ij_2k}ij_2k}), \tag{23}$$

as quantification of 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}. 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 (23) using the generated iid samples conditionally on the remaining parameters to approximate (23).

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$j_1, j_2 \in \{1, \ldots, J\}$,

$$H_{05ij_1j_2k}: C(i, j_1, j_2, k) = 0 \text{ versus } H_{15ij_1j_2k}: C(i, j_1, j_2, k) \neq 0. \tag{24}$$

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. In case H_{01} is rejected, we go forward to perform various tests related to gene-gene and gene-environment interactions, enlisted in Sections 3.1.1. and 3.1.2. above. For interpretations and detailed discussion on the tests see Section 4.1.4 of Bhattacharya and Bhattacharya (2020 b).

3.2. Methodologies for implementing the Bayesian tests

3.2.1. Hypothesis testing based on clustering modes

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.

For $k = 0, 1$, let i_k denote the index of the central clusterings of $P_{Mijk} = \{p_{1ijk}, p_{2ijk}, \ldots, p_{Mijk}\}$, $i = 1, \ldots, N_k$. We then study the divergence between the two clusterings of $P_{Mi_0jk=0} = \{p_{1i_0jk=0}, p_{2i_0jk=0}, \ldots, p_{Mi_0jk=0}\}$and $P_{Mi_1jk=1} = \{p_{1i_1jk=1}, p_{2i_1jk=1}, \ldots, p_{Mi_1jk=1}\}$, for $j = 1, \ldots, J$.

Significantly large clustering distance between $P_{Mjk=0}$ and $P_{Mjk=1}$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 a)) 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 $\overline{p}_{mijk} = \sum_{r=1}^{L_j} p_{\mathrm{m}ijkr}/L_j$, then consider their logit transformations $\mathrm{logit}(\overline{p}_{mijk}) = \log\{\overline{p}_{mijk}/(1-\overline{p}_{mijk})\}$. Then, we compute the Euclidean distance between the vectors

$\mathrm{logit}(\overline{P}_{Mi_0jk=0}) = \{\mathrm{logit}(\overline{p}_{1i_0jk=0}), \mathrm{logit}(\overline{p}_{2i_0jk=0}),\ldots, \mathrm{logit}(\overline{p}_{\mathrm{M}i_0jk=0})\}$ and

$\mathrm{logit}(\overline{P}_{Mi_1jk=1}) = \{\mathrm{logit}(\overline{p}_{1i_1jk=1}), \mathrm{logit}(\overline{p}_{2i_1jk=1}),\ldots, \mathrm{logit}(\overline{p}_{\mathrm{M}i_1jk=1})\}$

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

$$d_{E,j} = d_{E,j}(\mathrm{logit}(\overline{P}_{Mi_0jk=0}), \mathrm{logit}(\overline{P}_{Mi_1jk=1}))$$

and denote $\underset{1 \le j \le J}{max}\ d_{E,j}$ by $d^*_{E,j}$.

3.2.2. 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:

$H_{0,d^*}: d^* < \varepsilon_{d^*}$ versus $H_{1,d^*}: d^* \ge \varepsilon_{d^*}$; (25)

$H_{0,d^*_E}: d^*_E < \varepsilon_{d^*_E}$ versus $H_{1,d^*_E}: d^*_E \ge \varepsilon_{d^*_E}$; (26)

For l = 1,2,...,*d*

$H_{0,\beta_{Gl}}: |\beta_{Gl}| < \varepsilon_{Gl}$ versus $H_{1,\beta_{Gl}}: |\beta_{Gl}| \ge \varepsilon_{Gl}$ (27)

$H_{0,\beta_{Gl}}: |\beta_{G_0l}| < \varepsilon_{G_0l}$ versus $H_{1,\beta_{G_0l}}: |\beta_{G_0l}| \ge \varepsilon_{G_0l}$ (28)

$H_{0,\beta_{Hl}}: |\beta_{\mathrm{H}l}| < \varepsilon_{Hl}$ versus $H_{1,\beta_{\mathrm{H}l}}: |\beta_{\mathrm{H}l}| \ge \varepsilon_{Hl}$ (29)

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

$H_{0,\mathrm{C(i},j_1,j_2,\mathrm{k)}}: |\mathrm{C(i},j_1,j_2,\mathrm{k)}| < \varepsilon_{\mathrm{C(i},j_1,j_2,\mathrm{k)}}$ versus $H_{1,\beta_{\mathrm{C(i},j_1,j_2,\mathrm{k)}}}: |\beta_{\mathrm{C(i},j_1,j_2,\mathrm{k)}}| \ge \varepsilon_{\mathrm{C(i},j_1,j_2,\mathrm{k)}}$ (30)

If H_0 is rejected in (25) or in (26), 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 $\hat{d}_j = \hat{d}(\overline{P}_{Mi_0jk=0}, \overline{P}_{Mi_1jk=1})$; we could also test $H_{0,d_{E,j}}: d_{E,j} < \varepsilon_{d_{E,j}}$ versus $H_{1,d_{E,j}}: d_{E,j} \ge \varepsilon_{d_{E,j}}$. For the null model and choice of ε see Bhattacharya and Bhattacharya (2020 b).

4. 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

under 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 S-3 of the supplement of Bhattacharya and Bhattacharya (2020 b). Here we briefly summarize the results of our experiments. In case (a), we correctly obtained clear significance of the influence of genetic effects. Also, $\beta_{\mathrm{H}l}$ turned out to be very significant, demonstrating significant overall impact of the environmental variable on gene-gene interaction. 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 (2020A). 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.

5. Application of 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.

5.1. Data description

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 sub-populations: 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 a) 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 a)) for the details and the relevant references.

5.2. Remarks on model implementation

Our parallel MCMC algorithm detailed in Section S-2 of the supplement of Bhattacharya and Bhattacharya (2020 b), takes about 7 days to generate 30,000 iterations on our VMware consisting of 1 TB RAM, 60 double-threaded, 64-bit physical cores, each running at 2.5 GHz; 50 such cores were available to us. We discard the first 10, 000 iterations as burn-in, using the subsequent 20,000 iterations for our Bayesian inference. Convergence is studied using informal convergence diagnostics such as trace plots. Some instances are provided in Section S-3 of the supplement of Bhattacharya and Bhattacharya (2020 b).

5.3. Results of the real data analysis

5.3.1. Effect of the sex variable

We obtain $P(|\beta_{Gl}| < \varepsilon_{Gl}$ |Data) ≈ 0, $P(|\beta_{G_0 l}| < \varepsilon_{G_0 l}$|Data) ≈ 0 and $P(|\beta_{\mathrm{H}l}| < \varepsilon_{Hl}$|Data) ≈ 1. In other words, although $\bar{\bar{E}}$ (here E being the sex variable) is insignificant, both E_{ik} and $\overline{E}_k$ are very significant. Thus, in this study, sex seems to play an important role in influencing gene-gene interaction.

5.3.2. Roles of individual genes

With the clustering metric we obtained $P(d^* < \varepsilon_{d^*}) \approx 0.030$ while that with the Euclidean distance we obtained $P({d^*}_E < \varepsilon_{{d^*}_E}|\text{Data}) \approx 0.540$. 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. None of the individual genes turned out to be significant, for either the clustering metric or the Euclidean metric. The posterior probabilities of the null hypotheses (of no significant genetic influence) with respect to the clustering metric is shown in Figure 1.

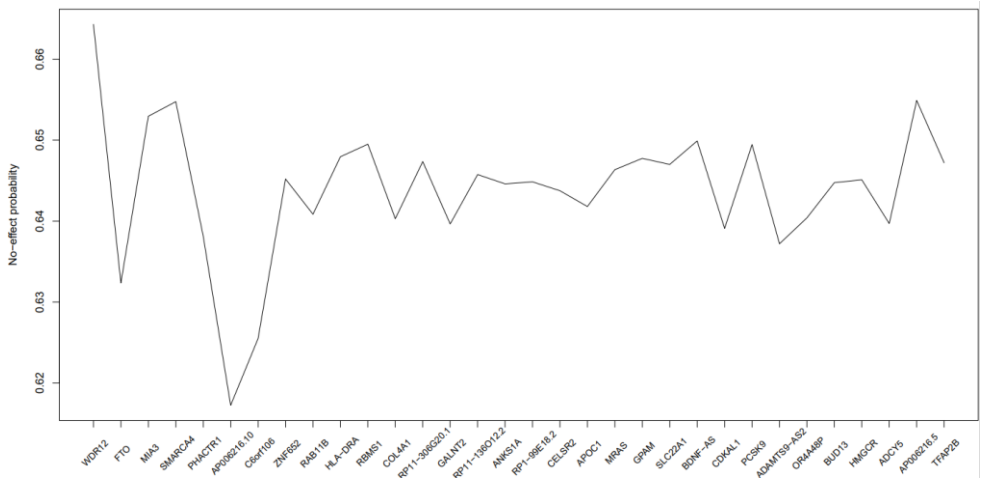

Figure 1: Posterior probability of no genetic effect with respect to clustering metric

5.3.3. Gene-gene interactions

Figures 2(a) to 2 (d) show the typical gene-gene correlations representative of cases and controls in males and females. The colour 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. Our tests indicate that for most of the subjects, at least one of the genes AP006216.10 and C6orf106 interact 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 might be associated with 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, following this study, the gene-gene interactions need to be investigated further for their possible beneficial effect on the subjects against MI.

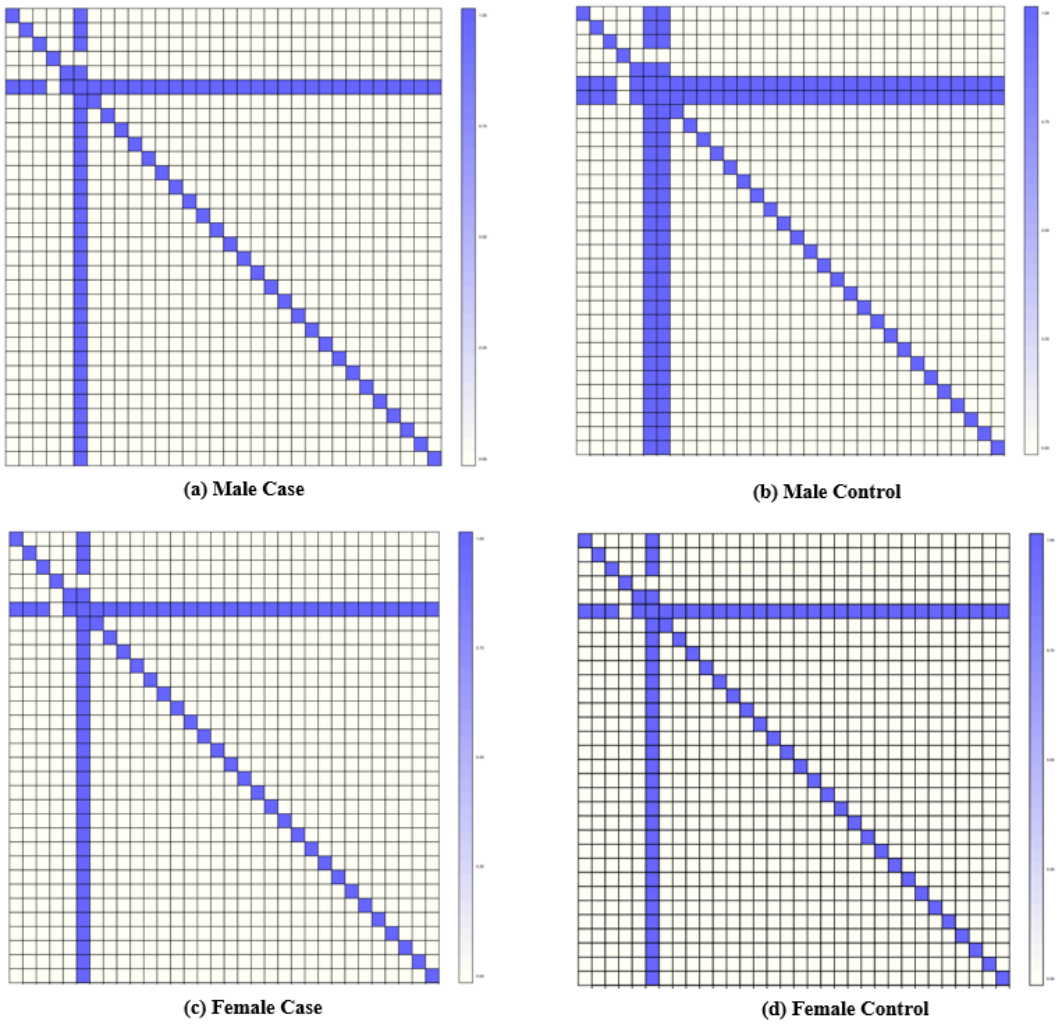

Figure 2: Typical gene-gene posterior correlation plot for male cases and controls and female cases and controls

5.3.4. Posteriors of the number of sub-populations

Figure 3 shows the posteriors of the number of sub-populations for the males and females associated with respectively. Observe that the posteriors are quite similar, 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.

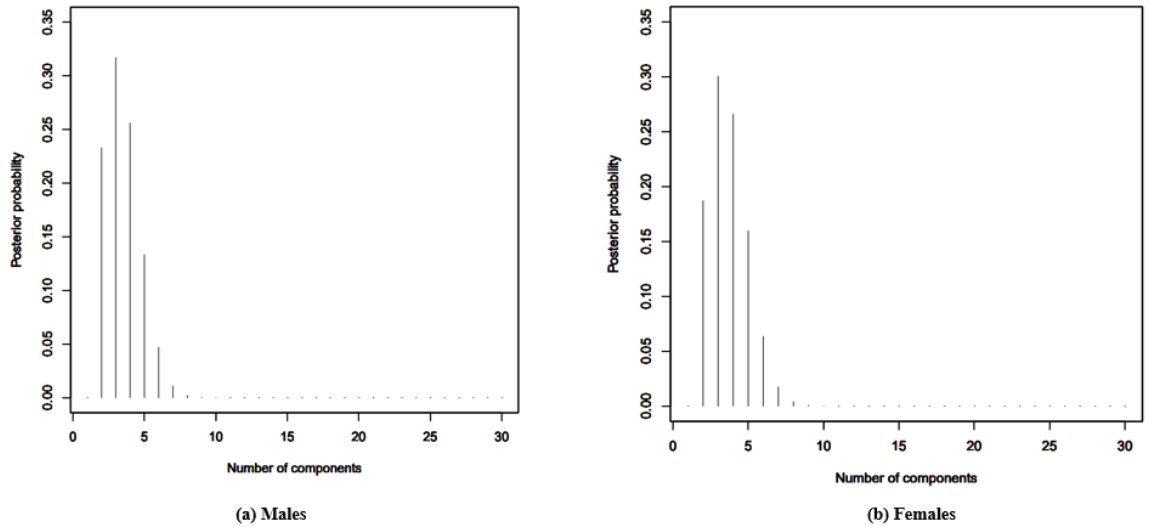

Figure 3: Typical posteriors of the number of components for males and females

6. 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 the existing work in this area, in the sense of much clear interpretability and accounting for subject-specific gene-gene interactions. We propose a novel parallel MCMC algorithm to implement our model (Sections S-1and S-2 of the supplementary material of Bhattacharya and Bhattacharya (2020 b)), that combines powerful technology with conditionally independent structures inherent within our HDP based model and efficient TMCMC methods. Applications of our ideas to biologically realistic datasets generated under 5 different setups 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 real MI dataset yielded results that are broadly in agreement with the previous works on the same dataset. For example, in accordance with Bhattacharya and Bhattacharya (2020A) (see also Lucas *et al.* (2012)) we obtained strong impact of the sex variable, weak gene-gene correlations but no significant effect of the individual genes. But special mention must be reserved for our original finding that two genes, AP006216.10 and C6orf106, tend to fight the disease by their positive interaction with the remaining genes. Another interesting discovery that emerged from our analyses is that only in male cases all the gene-gene interactions were insignificant. These two findings seem to confirm the general belief that as compared to females, males are more vulnerable to heart attack.

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. (2020 a). 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. Also available at "https://arxiv.org/abs/1601.03519".

Bhattacharya, D. and Bhattacharya, S. (2020 b). A non-gaussian, nonparametric structure for gene-gene and gene-environment interactions in case-control studies based on hierarchies of Dirichlet processes. arXiv:1704.07349v2 [stat.AP].

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.

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

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.

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.

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**; https://doi.org/10.1186/1471-2105-13-132.

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.

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

Yi, N., Kaklamani, V. G. and Pasche, B. (2011). Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. *Annals of Human Genetics*, **75**, 90–104.