Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 153-166

An Overview of Bayesian Semiparametric Approaches for Genetic Association Studies

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: 01 July 2025; Revised: 16 August 2025; Accepted: 19 August 2025

Abstract

In human genetics, Bayesian semiparametric approaches have proven especially effective in disease gene association studies, where genetic heterogeneity and complex interactions are common. They are particularly advantageous in stratified subpopulation settings with an unknown number of subgroups. Unlike traditional parametric models that require prespecifying the number of subpopulations, nonparametric methods such as Dirichlet Process mixture models allow the number and structure of subpopulations to be learned from the data. This flexibility enables more accurate detection of disease-associated variants while accounting for population structure, which are key challenges in complex trait analysis and precision medicine. This work provides an overview of how Dirichlet Process based mixture models can be used to flexibly model gene-gene and gene-environment interactions and identify disease-associated variants in complex, stratified populations with unknown heterogeneity.

Key words: Dirichlet process; Genetic association studies; Mixture model; Parallel computing; TMCMC.

AMS Subject Classifications: 62K05, 05B05.

1. Introduction

1.1. Gene-gene and gene-environment interaction

With recent technological advances, it is now possible to assay millions of loci in an individual's genomic DNA to identify disease-associated genes. While this capability has revolutionized genetic research, it has also introduced substantial analytical challenges, particularly in managing the massive volume of data generated. Addressing these challenges requires the development of sophisticated statistical models that integrate current biological and biochemical knowledge of disease mechanisms. Such models not only facilitate efficient

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

computation but also enable deeper insights into the complex pathways underlying multifactorial diseases.

Genome-wide association studies (GWAS) have identified numerous single nucleotide polymorphisms (SNPs) associated with complex diseases, yet they explain only a small fraction of heritable genetic variation; see Larson and Schaid (2013). A growing body of research indicates that genes often function through intricate interaction networks, which significantly shape the genetic basis of complex traits Bonetta (2010). The limited explanatory power of GWAS may stem from the absence of models that incorporate gene-gene interactions into genomic analysis Cordell (2009), thereby overlooking important biological mechanisms Yi (2010).

A major obstacle in studying genetic interactions lies in the lack of a clear definition of epistasis. Phillips (2008) distinguishes between functional and compositional biological epistasis, both of which differ from the classical statistical definition proposed by Fisher (1918) and extended by Kempthorne (1954). While VanderWeele (2009) identifies conditions for alignment between statistical and biological definitions, most statistical tests fail to reflect the biological complexity of interactions. Still, statistical models are essential for quantifying these effects Cordell (2002), Wang *et al.* (2010).

SNP-SNP interactions are often used to model gene-gene interactions in case-control studies Yi *et al.* (2011). However, SNP-level models are computationally intensive due to the large number of interaction terms required, whereas gene-level models offer dimensionality reduction at the expense of finer detail Larson and Schaid (2013), Musameh *et al.* (2015). Moreover, additive linear models can oversimplify interaction mechanisms and obscure interpretability, especially when principal components are used for reduction Wang *et al.* (2010).

These challenges are compounded by the frequent neglect of population substructure. Genetic effects can vary across subpopulations, and ignoring such heterogeneity can lead to biased inference and inflated false positives Bhattacharjee *et al.* (2010). Since the number and structure of subgroups are usually unknown, flexible models that can infer latent structure are critical.

Beyond genetic interactions, the interplay between genes and environmental factors is critical to understanding disease etiology. Although most diseases arise from a combination of genetic and environmental influences, only a small subset are purely monogenic. Environmental exposures can alter genetic risk Mapp (2003), Khouri (2005), and in certain cases, disease manifestation occurs only beyond specific environmental thresholds. Hunter (2005) emphasize that neglecting such interactions can lead to misestimation of the population-level disease burden. These interactions are particularly salient in pharmacogenetics, where treatment efficacy and safety may vary by genotype Scott (2011). Mechanistically, gene–environment interactions may act through pathways such as epigenetic modification and transcriptional regulation Purcell (2002), Ottman (2010). However, existing statistical approaches, particularly linear and log-linear models, often fail to adequately capture these complex dependencies Mukherjee *et al.* (2008), Mukherjee and Chatterjee (2008), Mukherjee *et al.* (2010), Mukherjee *et al.* (2012), Sanchez *et al.* (2012), Ahn *et al.* (2013), Ko *et al.* (2013).

These limitations point to the need for more general, data-adaptive approaches.

Bayesian nonparametric methods based on Dirichlet Process mixture models offer the flexibility to address gene-gene and gene-environment interactions while accounting for population stratification. This article surveys recent developments in this direction, building on the framework proposed in Bhattacharya and Bhattacharya (2018) and Bhattacharya and Bhattacharya (2024).

2. An overview of our Bayesian nonparametric ideas

This paper presents a Bayesian nonparametric/semiparametric framework for analyzing gene–gene interactions, fundamentally differing from traditional logistic regression approaches. Rather than modeling disease status conditional on genotype, we model genotype distributions conditional on disease status. To account for hidden population substructure, Dirichlet process-based finite mixtures (Bhattacharya, 2008) are embedded within a hierarchical model that captures interactions at both gene and SNP levels via matrix-normal priors. The framework extends naturally to gene–environment interactions through covariate-dependent priors, enabling the assessment of how environmental factors influence genetic associations.

Our Bayesian approach addresses multiple sources of uncertainty and moves beyond binary presence–absence tests by modeling the magnitude and structure of interaction effects using correlation-based measures. Disease-predisposing loci (DPLs) are detected through novel posterior-clustering-based hypothesis testing. For computational efficiency in high-dimensional settings, Transformation-based Markov Chain Monte Carlo (TMCMC) (Dutta and Bhattacharya, 2014) is employed, facilitating block updates with high acceptance rates. Combined with parallel Gibbs sampling tailored for Dirichlet process mixtures, the method achieves substantial computational gains.

We validate the methodology through simulations and apply it to a myocardial infarction case–control SNP dataset. The results corroborate known associations and reveal novel gene–gene and gene–environment interactions, illustrating the flexibility and inferential power of the proposed framework.

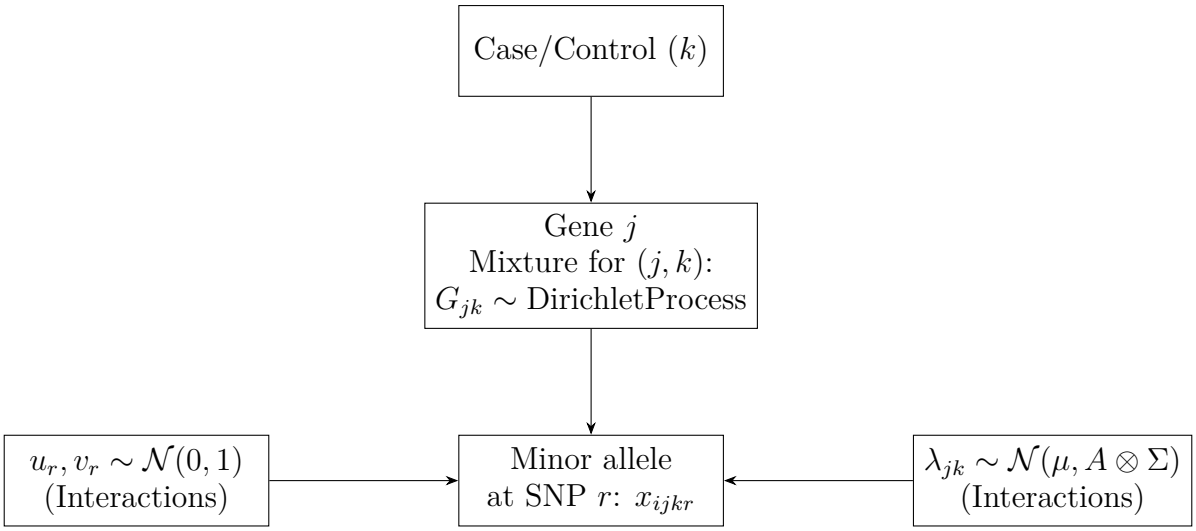

Figure 1: Schematic representation of the Bayesian model for gene–gene interactions

For each gene and case-control status, genotype data are modeled using Dirichlet process-based mixtures that capture sub-population structure. SNP-level dependencies and gene-gene interactions are introduced through a matrix-normal prior on latent interaction parameters. The modular design of the model allows efficient parallel computation: gene-specific mixture components are updated independently across processors, while the interaction parameters are updated centrally using transformation-based MCMC (TMCMC).

2.1. Case-control type genotype data

Humans have 22 pairs of autosomes and one pair of sex chromosomes in the nuclear genome. Chromosomes are composed of tightly coiled DNA, containing genes that may exist in alternative forms, known as *alleles*, at the same genetic locus. Variation in alleles can lead to phenotypic differences, and the specific allelic combination at a locus defines an individual's *genotype*. The most common genetic variation is the Single Nucleotide Polymorphism (SNP), a single base change in the DNA sequence. This study analyses SNP data from case and control cohorts in relation to a specific disease.

Let $s = 1, 2$ represent the two chromosomes. For an individual indexed by i, gene j, group k, and locus r, define $x^s_{ijkr} = 1$ if the minor allele is present, and $x^s_{ijkr} = 0$ otherwise. The indices range as follows: $i = 1, \ldots, N_k$; $j = 1, \ldots, J$; $k = 0, 1$, where $k = 1$ corresponds to the case group; and $r = 1, \ldots, L_j$. Given any (j, k), let $\boldsymbol{x}_{ijkr} = (x^1_{ijkr}, x^2_{ijkr})$, and $\boldsymbol{X}_{ijk} = (\boldsymbol{x}_{ijk1}, \boldsymbol{x}_{ijk2}, \ldots, \boldsymbol{x}_{ijkL_j})$.

2.2. Gene-gene interaction based mixture models driven by Dirichlet processes

We assume that for every triplet (i, j, k), $\boldsymbol{X}_{ijk}$ are independently distributed with mixture probability mass function with a *maximum* of M components, given by

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

where $f\left(\cdot | p_{mjkr}\right)$ is the probability mass function of independent Bernoulli distributions, given by

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

Using allocation variables z_{ijk}, with probability distribution

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

for $i = 1, \ldots, N_k$ and $m = 1, \ldots, M$, (1) can be represented as

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

We may assume appropriate Dirichlet distribution priors on $(\pi_{1jk}, \ldots, \pi_{Mjk})$ for $j = 1, \ldots, J$; $k = 0, 1$. Following Mukhopadhyay and Bhattacharya (2021), we set $\pi_{mjk} = 1/M$, for $m = 1, \ldots, M$, and for all (j, k).

Letting $\boldsymbol{p}_{mjk} = \left(p_{mjk1}, p_{mjk2}, \ldots, p_{mjkL_j}\right)$, we further assume that

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

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

where $\mathrm{DP}\left(\alpha_{jk}\boldsymbol{G}_{0,jk}\right)$ stands for Dirichlet process with expected probability measure $\boldsymbol{G}_{0,jk}$ having precision parameter α_{jk}. We assume that under $\boldsymbol{G}_{0,jk}$, for $m = 1, \ldots, M$ and $r = 1, \ldots, L_j$,

$$p_{mjkr} \overset{iid}{\sim} \mathrm{Beta}\left(\nu_{1jkr}, \nu_{2jkr}\right). \tag{7}$$

Discreteness of Dirichlet processes causes coincidences among the parameter vectors of $\boldsymbol{P}_{Mjk} = \{\boldsymbol{p}_{1jk}, \boldsymbol{p}_{2jk}, \ldots, \boldsymbol{p}_{Mjk}\}$ with positive probability, so that, with positive probability, the actual number of mixture components in (1) falls below M, the maximum number of components, the mixing probabilities taking the form M^*/M, where $1 \leq M^* \leq M$. The property of coincidences among the parameter vectors is clearly preserved by the Polya urn scheme. Notationally, we shall denote the number of distinct elements of $\boldsymbol{P}_{Mjk} = \{\boldsymbol{p}_{1jk}, \boldsymbol{p}_{2jk}, \ldots, \boldsymbol{p}_{Mjk}\}$ by τ_{jk} and that of $\boldsymbol{P}_{Mjk}\backslash\{\boldsymbol{p}_{mjk}\}$ by $\tau_{jk}^{(m)}$.

Conditioned on $\boldsymbol{G}_{jk}$, our fixed-M mixture model mimics an infinite-dimensional Dirichlet process mixture despite the non-iid nature of the data (Mukhopadhyay and Bhattacharya (2021)). The number of distinct components in $\boldsymbol{P}_{Mjk}$ can vary across (j, k) due to random duplication. This flexibility aligns with biological expectations, as genotype distributions often differ between cases and controls under genetic influence. Such heterogeneity is naturally accommodated within our framework.

2.3. Gene-gene, SNP-SNP interactions and parallel processing

To incorporate the SNP-SNP dependence, which may exist within each gene and also among the genes, The Beta parameters are modelled as ν_{1jkr} and ν_{2jkr} of (7) as follows:

For $r = 1, \ldots, L$, where $L = \max\{L_j;\ j = 1, \ldots, J\}$, and for every (j, k),

$$\nu_{1jkr} = \exp\left(u_r + \lambda_{jk}\right); \tag{8}$$

$$\nu_{2jkr} = \exp\left(v_r + \lambda_{jk}\right). \tag{9}$$

We further assume that for $r = 1, \ldots, L$,

$$u_r \overset{iid}{\sim} N(0, 1); \tag{10}$$

$$v_r \overset{iid}{\sim} N(0, 1). \tag{11}$$

The Gaussian priors on u_r and v_r with other means and variances did not yield significantly different results, establishing the prior robustness in our modeling strategy.

Subsequently, the SNP-wise dependence in a gene is modelled using matrix-normal distribution

$$\boldsymbol{\lambda} = \{\lambda_{jk};\ j = 1, \ldots, J,\ k = 0, 1\} \sim N(\boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{\Sigma}),$$

as a prior for $\boldsymbol{\lambda}$ ($\boldsymbol{\Lambda}$ in matrix form) with appropriate inverse-Wishart priors on $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$. Furthermore, the matrix-normal prior induces dependence among genes, which in turn creates dependencies among the SNPs belonging to different genes.

Given that the mixture distributions for each gene $j \in 1, \ldots, J$ and case-control group $k \in 0, 1$ are conditionally independent when the interaction parameters are known, we take advantage of this structure for efficient computation. Mixture components are updated simultaneously across multiple processors, while the interaction parameters, which govern the dependencies, are updated afterward on a single processor using a specialized TMCMC approach.

This separation in the update steps enables the method to handle large-scale data effectively while preserving the ability to capture complex gene-gene and SNP-SNP relationships.

2.4. Summary of analysis of the MI dataset

In our analysis of the real Myocardial Infarction (MI) dataset, we focused on a total of 1251 SNPs, out of which only 33 had prior evidence suggesting a possible link to the disease. The remaining 1218 SNPs had no documented association with MI and were largely considered unlikely candidates for influencing disease risk. In fact, apart from a few among the 33 literature-supported SNPs, most of the others were included not because of prior biological relevance, but to test the robustness of our model in distinguishing meaningful signals from noise. Interestingly, in several instances, the disease-predisposing loci (DPL) identified by our Bayesian approach matched those already highlighted in the literature as relevant to MI. Notable examples include SNP $rs7395662$ in gene $OR4A48P$, SNP $rs964184$ in $AP006216.10$, SNP $rs4420638$ in $APOC1$, SNP $rs1564348$ in $SLC22A1$, and SNP $rs1013442$ in $BDNF\text{-}AS$. This alignment underscores the model's ability to successfully detect true associations, thereby effectively controlling false negatives. Conversely, SNPs not identified as DPLs either by our approach or by prior studies can be reasonably regarded as unrelated to the disease, indicating that the model also maintains strong control over false positives.

3. Extension to gene-environment interactions

Our Bayesian hierarchical mixture framework integrates the mechanisms by which gene–environment interactions, as well as the isolated and joint effects of genes, contribute to disease susceptibility, while accommodating potential population stratification. A distinctive feature of the model is its ability to infer the number of latent genetic subpopulations.

To capture the influence of environmental variables, the proposed semiparametric specification employs Dirichlet process-based finite mixtures at the individual level, jointly modeling genetic profiles and case–control status. These mixtures are linked through a structured dependence encoded via hierarchical matrix-normal distributions, enabling the model to account for correlations induced by environmental exposures. The framework extends the gene–gene interaction model and Bayesian hypothesis testing methodology developed in Section 2 to detect the effects of genes, environmental factors, and their interactions.

Computation is performed via a parallel MCMC scheme that leverages the model's conditional independence structure, combining Gibbs sampling with Transformation-based MCMC (TMCMC) for efficient high-dimensional updates. Environmental covariates influence individual-level Dirichlet process mixtures, allowing for subject-specific modulation

of genotype distributions. The prior hierarchy accommodates locus-specific, gene-specific, and environment-dependent parameters. Parallel updates are applied to gene–environment-specific components, while interaction parameters are updated centrally using TMCMC. Figure 2 presents a schematic representation of the proposed framework.

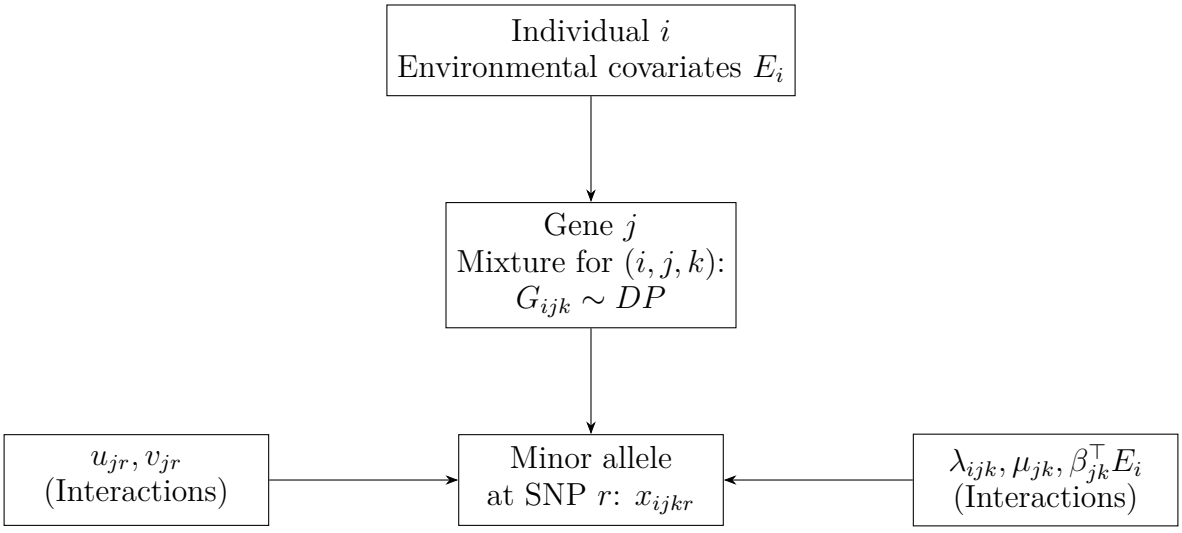

Figure 2: Diagram of the extended Bayesian framework incorporating gene-environment interactions

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

Let E_i denote the set of environmental variables associated with the i-th individual. We model the case-control genotype data, together with environmental information, using our Bayesian semiparametric model.

Let $x_{ijkr} = (x^1_{ijkr}, x^2_{ijkr})$ denote the genotype at the r-th locus of the j-th gene for the i-th individual in the k-th group (case/control), and let $\mathbf{X}_{ijk} = (x_{ijk1}, x_{ijk2}, \ldots, x_{ijkL_j})$ denote the genotype information across all L_j loci of the j-th gene. Let p_{mijkr} denote the minor allele frequency at the r-th locus of the j-th gene for the i-th individual in the k-th group. The minor allele frequency represents the frequency at which the second most common allele occurs in a given population.

We assume the mixture distribution:

$$[\mathbf{X}_{ijk}] = \sum_{m=1}^{M} \pi_{mijk} \prod_{r=1}^{L_j} f(x_{ijkr} \mid p_{mijkr}), \tag{12}$$

where $f(\cdot \mid p_{mijkr})$ denotes the Bernoulli mass function:

$$f(x_{ijkr} \mid p_{mijkr}) = p_{mijkr}^{x^1_{ijkr}+x^2_{ijkr}} (1 - p_{mijkr})^{2-(x^1_{ijkr}+x^2_{ijkr})}, \tag{13}$$

and M is the maximum number of mixture components. The allocation variables z_{ijk} are such that:

$$[z_{ijk} = m] = \pi_{mijk}, \quad m = 1, \ldots, M. \tag{14}$$

We set $\pi_{mijk} = 1/M$ for all (i, j, k) and m, as this fixed weight approach has been shown to yield better performance than Dirichlet priors in learning about the true number of components.

This representation captures the possibility that different individuals, even within the same group and gene, may belong to different sub-populations, influenced by their environmental exposures E_i. This is a substantial extension from the model in Section 3, which did not account for environmental effects.

3.2. Summary of the results of MI data analysis with this new model

We applied the proposed model to the myocardial infarction (MI) dataset previously analyzed in Section 2.4, incorporating sex as an environmental covariate. The resulting inferences were consistent with established findings in the literature. Although gene–gene interactions were not statistically significant, SNP–SNP correlations, quantified via Euclidean distances between case and control groups, provided plausible explanations for discrepancies between our identified disease-predisposing loci (DPLs) and those reported in earlier studies.

Importantly, the Bayesian framework produced interpretable results despite the limited sample size of 200 individuals, underscoring the utility of hierarchical modeling with informative priors and the efficiency of the employed MCMC algorithms.

4. A general model for gene-gene and gene-environment interactions based on hierarchies of Dirichlet processes

As discussed in Section 3, gene–gene interactions alone are insufficient to explain the etiology of most complex diseases. Similarly, examining environmental factors in isolation from genetic variation is inadequate; biomedical evidence underscores the pivotal role of gene–environment interactions in elucidating complex disease mechanisms. Given the absence of a simple, additive relationship between genetic and environmental influences, linear or additive models commonly used to date are inadequate for modeling these interactions.

In Section 3, we introduced a Bayesian semiparametric model for case–control genotype data, employing Dirichlet process-based finite mixtures at the subject level. A hierarchical matrix-normal dependence structure linked these mixtures to capture correlations among genes under environmental influence. However, a potential limitation of this framework arises from its induced covariance structure: for individual i, the relevant gene–gene covariance matrix is $\tilde{\sigma}_{ii}A$, where A is a common gene–gene interaction matrix (in the absence of environmental variables) and $\tilde{\sigma}_{ii} = \sigma_{ii} + \phi$, with σ_{ii} denoting the i-th diagonal element of a symmetric positive-definite matrix unrelated to environmental variables, and $\phi \geq 0$ representing the effect of the environmental covariate E. This formulation assumes that environmental exposures modify gene–gene interactions in an identical manner across all individuals, which may be unrealistic when the magnitude and nature of exposure vary.

To address this limitation, we propose a Bayesian nonparametric framework for modeling joint gene–gene and gene–environment interactions, as developed in Bhattacharya (2019) (see also Bhattacharya and Bhattacharya (2024)). Like the earlier model, individual genotype distributions are represented via Dirichlet process-based finite mixtures; however, in place of the matrix-normal dependence structure, we introduce a hierarchy of Dirich-

let processes that flexibly captures nonparametric dependencies among genes induced by environmental covariates, case–control status, and inter-individual heterogeneity. This hierarchical construction overcomes the restrictive assumptions of the matrix-normal approach described in Section 3. Although conceptually related to the hierarchical Dirichlet process (HDP) of Teh *et al.* (2006), our model introduces an additional level of hierarchy, enhancing flexibility.

For computation, we develop a highly parallelizable MCMC algorithm that integrates modern parallel computing resources with Gibbs sampling, retrospective sampling, and Transformation-based MCMC (TMCMC). The Bayesian hypothesis testing procedures from our earlier framework are extended to this enriched setting.

Letting $s = 1, 2$ denote the two chromosomes, we define $y^s_{ijkr} = 1$ and 0 to indicate the presence and absence, respectively, of the minor allele for the i-th individual in group $k \in \{0, 1\}$ (with $k = 1$ denoting the case group), at the r-th locus of the j-th gene, for $i = 1, \ldots, N_k$; $r = 1, \ldots, L_j$; and $j = 1, \ldots, J$. Let $N = N_0 + N_1$, and let E_i denote a vector of environmental variables associated with individual i.

Again, before describing the components of the model in detail, we first present the schematic diagram in Figure 3.

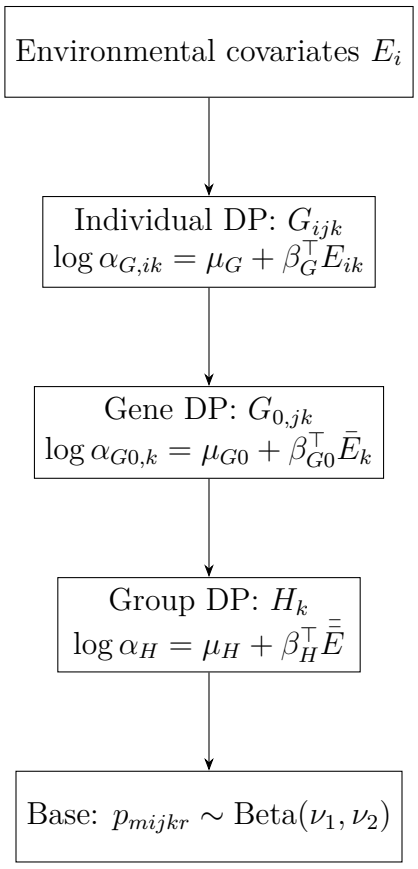

Figure 3: Schematic representation of the hierarchical Dirichlet process (HDP) model for gene-gene and gene-environment interactions

This fully nonparametric framework models dependencies across individuals, genes, and groups through a three-level hierarchy of Dirichlet processes. Environmental covariates influence the precision parameters at each level, allowing flexible, individualized representation of interaction structures. The base distribution is a Beta prior on allele frequencies. This hierarchy enables rich modeling of stratification and interaction while maintaining computational scalability.

4.1. Summary of the MI data analysis with the HDP model

Our analysis of the MI dataset revealed a strong effect of the sex variable, consistent with the findings in Section 3. Our hypothesis tests indicated no significant marginal effects of individual genes, in agreement with Section 3 where only weak marginal effects were observed.

Most notably, even though gene-gene correlations were generally weak, again consistent with Section 3 and Lucas *et al.* (2012), our tests detected that two genes, *AP006216.10* and *C6orf106*, exhibited broad, beneficial interactions with other genes that may help combat the disease. Furthermore, in the only subgroup for which all gene-gene interactions were found to be insignificant was the male cases. Hence, our results lend statistical support to the widely held belief that males may be more susceptible to heart attacks than females.

4.2. Summary and future directions

This work presents a unified Bayesian nonparametric framework for analyzing gene–gene and gene–environment interactions in case–control studies. The proposed approach is designed to accommodate multiple layers of uncertainty, a feature that distinguishes it from many existing methods that prioritize computational feasibility for large-scale datasets. Such differences in objectives necessarily lead to different performance criteria, making direct comparisons with standard approaches inappropriate. Both the simulated and real datasets analyzed here exhibit multiple subpopulations. While methods such as principal component analysis can infer subpopulation structure, most approaches require the number of subpopulations to be fixed a priori, which can lead to misestimation and inflated false positives Bhattacharjee *et al.* (2010). Since genetic interactions may differ across subpopulations, such errors can bias inference. Our method explicitly models this uncertainty, in contrast to De Iorio *et al.* (2015b) and De Iorio *et al.* (2015a), which do not address gene–gene or gene–environment interactions.

Existing approaches generally test only for the presence of interactions without quantifying their strength, whereas our framework enables classification of genes by the magnitude of their interactions. Many standard methods rely on heuristic definitions of main and interaction effects, for example, kernel-based methods Larson and Schaid (2013), Kullback–Leibler divergence Wan *et al.* (2010), entropy-based information gain Li *et al.* (2015), or genotype categories Yi *et al.* (2011), which can yield results sensitive to the chosen definition. In contrast, our framework models interactions using established statistical principles. Furthermore, most current models analyze pairwise SNP–SNP interactions via logistic regression, neglecting genes as functional units and lacking scalability to higher-order interactions. Two-stage approaches such as `BOOST` and Bayesian methods like BEAM or EpiBN operate only at the SNP level and overlook gene-level modeling Niel *et al.* (2015). Our

unified Bayesian approach simultaneously models uncertainties in both gene- and SNP-level interactions within a coherent probabilistic structure. Finally, the simulation datasets used to validate our method were generated under logistic models, which form the basis for most competing approaches. Given that our framework is nonparametric and fundamentally distinct from logistic regression, such simulation settings do not allow for direct performance comparisons.

The proposed methodology addresses several key challenges in genetic association studies, including population stratification, uncertainty in subgroup structure, and the joint modeling of genetic effects at both the SNP and gene levels. The model incorporates complex dependency structures through hierarchical Dirichlet process mixtures, and Bayesian hypothesis testing procedures are introduced to assess interaction significance and identify disease-predisposing loci. Computationally, the framework is highly scalable, leveraging parallelization, Gibbs sampling, and Transformation-based MCMC to efficiently analyze high-dimensional genomic data. Simulation studies and application to a myocardial infarction dataset demonstrated the accuracy, robustness, and interpretability of the approach, yielding results consistent with established findings while also uncovering novel patterns, including sex-specific susceptibility.

The flexibility of the proposed model allows for natural extensions to incorporate additional biological complexities, such as dynamic environmental effects or longitudinal data. Future work may extend the framework to handle time-to-event outcomes and integrate multi-omics data. Overall, this study demonstrates how Dirichlet process-based Bayesian nonparametric methods can advance the analysis of complex diseases by providing a principled, flexible, and computationally efficient alternative to traditional GWAS analyses, thereby contributing to a deeper understanding of the genetic architecture underlying complex traits.

Conflict of interest

The author do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

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

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

Bhattacharya, D. (2019). *Bayesian Nonparametric Approaches to Investigating Gene-Gene and Gene-Environment Interactions in Case-Control Studies*. Doctoral thesis, Indian Statistical Institute. Available at `https://www.researchgate.net/publication/361505764_Bayesian_Nonparametric_Approaches_to_Investigating_Gene-Gene_and_Gene-Environment_Interactions_in_Case-Control_Studies`.

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. (2024). Gene-gene and gene- environment interactions in case-control studies based on hierarchies of dirichlet processes. *Statistics and Applications*, **22**, 327–360.

Bhattacharya, S. (2008). Gibbs sampling based Bayesian analysis of mixtures with unknown number of components. *Sankhya. Series B*, **70**, 133–155.

Bonetta, L. (2010). Protein-protein interactions: Interactome under construction. *Nature*, **468**, 851–854.

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**, 2463–2468.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews*, **10**, 392–404.

De Iorio, M., Elliott, L. T., Favaro, S., Adhikari, K., and Teh, Y. W. (2015a). Modeling population structure under hierarchical Dirichlet process. Available at "arXiv:1503.08278v1".

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

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

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinborough*, **52**, 399–433.

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

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London. Series B*, **143**, 103–113.

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

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

Larson, N. B. and Schaid, D. J. (2013). A kernel regression approach to gene-gene interaction detection for case-control studies. *Genetic Epidemiology*, **37**, 695–703.

Li, J., Huang, D., Guo, M., Liu, X., Wang, C., Teng, Z., Zhang, R., Jiang, Y., Lv, H., and Wang, L. (2015). A gene-based information gain method for detecting gene-gene interactions in case-control studies. *European Journal of Human Genetics*, **23**, 1566–1572.

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.

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

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

Mukherjee, B., Ahn, J., Gruber, S. B., Ghosh, M., and Chatterjee, N. (2010). Bayesian sample size determination for case-control studies of gene-environment interaction. *Biometrics*, **66**, 934–948.

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

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

Mukhopadhyay, S. and Bhattacharya, S. (2021). Bayesian MISE convergence rates of Polya urn-based density estimators: Asymptotic comparisons and choice of prior parameters. *Statistics: a Journal of Theoretical and Applied Statistics*, **55**, 120–151.

Musameh, M., Wang, W., Nelson, C., C.L-Ganella, Debiec, R., Subirana, I., Elosua, R., Balmforth, A., Ball, S., Hall, A., Kathiresan, S., Thompson, J., Lucas, G., Samani, N., and Tomaszewski, M. (2015). Analysis of gene-gene interactions among common variants in candidate cardiovascular genes in coronary artery disease. *Plos One*, **10**, 1–12.

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, **6**, 285.

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

Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Review Genetics*, **9**, 855–867.

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

Sanchez, B., Kang, S., and Mukherjee, B. (2012). A latent variable approach to study of gene-environment interactions in the presence of multiple correlated exposures. *Biometrics*, **68**, 466–476.

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

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.

VanderWeele, T. J. (2009). Sufficient cause interactions and statistical interactions. *Epidemiology*, **20**, 6–13.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., and Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics*, **87**, 325–340.

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

Yi, N. (2010). Statistical analysis of genetic interactions. *Genetics Research*, **92**, 443–459.

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.