Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

DOTTORATO DI RICERCA
IN MATEMATICA, INFORMATICA, STATISTICA

CURRICULUM IN STATISTICA
CICLO XXXIV

Sede amministrativa Università degli Studi di Firenze
Coordinatore Prof. Matteo Focardi

Covariate-dependent Bayesian Models for Heterogeneous Populations

Settore Scientifico Disciplinare SECS-S/01

Dottorando:
Matteo Pedone

Tutore
Prof. Francesco C. Stingo

Co-Tutore
Prof. Raffaele Argiento

Coordinatore
Prof. Matteo Focardi

Anni 2018/2021

Abstract

Covariate-dependent Bayesian Models for Heterogeneous Populations

by

Matteo Pedone

In this thesis, we propose two novel Bayesian models for the analysis of health and genomic data, for which traditional methods are often found to be inefficient or unsuitable. Our approaches are motivated by the emerging field of precision medicine, whose ultimate goal is to select the optimal treatment accounting for patient and disease's variability. The main distinctive mark of statistical methodology in the precision medicine paradigm is to leverage patients' heterogeneity to obtain *subject-specific* inference.

First, motivated by a microbiota study on patients affected by colorectal cancer, we propose a model designed to analyze data that exhibit a hierarchical structure induced by measurements from multiple tissues of the same patient. Our goal is to capture patients' heterogeneity and similarities in terms of effects altering microbiota composition. Building upon the Dirichlet-multinomial model, we propose a flexible regression model, where coefficients are allowed to be smooth functions of the covariates. This results in a subject-specific model where varying coefficients include two-way linear and non-linear interactions as special cases. This allows us to recover associations and interactions patterns that may be specific for each individual rather than estimated at population level.

In the second contribution, we develop a predictive model for the selection of the personalized optimal treatment in oncology, when a predictive signature and a set of prognostic biomarkers are available. Predictive covariates are used to drive a clustering process that results in homogeneous groups of patients. This step is integrated into a prognostic model to predict response to competing treatments for new untreated patients. Finally, a utility-based approach allows us to select the treatment that ensures the larger predicted utility for new patients, based on their genetic profiles. We employed a Bayesian nonparametric model for random partition to build our integrative approach. In particular, we explored the use of the Normalized Generalized Gamma process as cohesion function in a product partition model with covariates. In contrast with existing methods, we jointly estimate model-based clustering and treatment assignment from the data, and hence treatment selection fully accounts for patients' variability.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In this chapter, we will detail the scientific and methodological reasons that motivated the projects collected in the dissertation. In particular, in Section 1.1 we illustrate that the development of a methodology able to account for heterogeneity in the population is crucial in the emerging field of precision medicine. The proposed methods are developed under the paradigm of Bayesian inference. This choice is also discussed. Section 1.2 introduces some of the modeling tools that we will use in the following chapters. In Section 1.3, we finally give an overview of the projects.

1.1 Scientific Background and Motivating Studies

The past decade has witnessed impressive advances in the understanding of molecular mechanisms underlying cancer, leading to some remarkable clinical successes in molecularly targeted cancer therapy (Ke and Shen, 2017). Nonetheless, successes in the development of targeted therapies are restricted to limited cases, as the moderate response rate across an unselected population confirms (Huang et al., 2014). The complexity that characterizes oncological diseases stems from the fact that heterogeneity arises from both patients with the same type of cancer and between cells within one patient. This makes population-based approaches unsuitable for an adequate understanding of oncogenesis and to devise appropriate cancer therapies (Betensky et al., 2002). Thus, the approach to cancer treatment is evolving from the traditional "one-size fits all" toward tailored treatments that account for individual variability in genes, clinical, and environmental features, termed *precision medicine* (De Bono and Ashworth, 2010).

Statistical methodology research in precision medicine is devoted to the development of personalized treatment rules to inform decision-making. Broadly speaking, the main distinctive mark of statistical inference in the precision medicine paradigm is to disregard heterogeneity as a nuisance to inference, but rather to take advantage of it to improve therapeutic strategies (Kosorok and Laber, 2019). This approach can shape the statistical methodology through the different stages of oncological studies. In fact, estimating average effects (that is population-level effects) using linear or constant-effects models for heterogeneous populations may result in biased estimates (Pearl, 2017). As a consequence, neglecting heterogeneity may lead inference to invalid conclusions.

In this sense, precision medicine represents a challenging discipline and specific statistical literature is now emerging to develop adequate methodologies. This

dissertation consists of two projects in which we develop statistical models that aim at providing contributions to this area of research and that are specifically devised for motivating applications.

Motivating study 1. Niccolai et al. (2020) provide a microbiome study conducted on patients affected by colorectal cancer (CRC). For each patient, up to three biological samples have been collected from different districts (tumor, fecal and salivary samples), and dietary habits have been measured along with clinical factors. Due to tumor heterogeneity, even patients with the same type of cancer may feature diverse microbial alteration, determined by their specific genomic profiles. Moreover, clinical covariates and dietary habits may produce alterations in the microbiota composition that either vary across districts or are district-specific.

We propose a hierarchical regression model in Chapter 2 to detect significant associations between covariates and microbial composition. The proposed method accounts for heterogeneity through the use of varying coefficients that include two-way interactions as a special case. Interactions among continuous covariates are modeled with a penalized splines (P-splines) approach for flexible modeling of continuous interactions.

Motivating study 2. The Cancer Genome Atlas (TCGA) provides clinical and level 3 protein expression data for patients affected by lower-grade glioma (LGG). We are interested in personalized optimal treatment selection, that is detecting the therapy ensuring the largest benefit for each patient, integrating predictive and prognostic biomarkers in the decision-making process. However, this poses a challenging statistical problem, since oncological patients should not be always regarded as exchangeable, since each tumor is unique, due to its large heterogeneity.

Our approach consists in identifying, for each treatment, a group of historical patients to which the new untreated patient should be considered exchangeable based on a measure of molecular similarity. Building on this, the treatment that ensures the largest predicted utility is to be considered the optimal one. In Chapter 3, we develop a predictive model for treatment selection that employs a covariate-dependent random partition model to obtain homogeneous groups of patients with respect to predictive biomarkers.

1.1.1 Bayesian paradigm

The Bayesian approach is particularly suited for the statistical challenges which we aim to address in the two projects. It allows to easily incorporate biological assumptions into the model structure, through prior distribution and flexible modeling techniques. The main challenges posed by the motivating examples are:

1. **Heterogeneity.** Biological data are intrinsically heterogeneous. Each patient has unique features and conducting inference at the population level may neglect the inherent patients' diversity and potential latent structure in the data.

2. **Sparsity.** In biomedical applications, sparsity is a fair prior expectation. Although complex, biological *phenomena* are usually explained by sparse represen-

tations. Moreover, sparse solutions improve statistical models' interpretability and inference accuracy.

3. **Nonlinearity.** Complex biological mechanisms suggest nonlinear relationships between outcome and clinical factors. Approaches that fail to account for such nonlinear dependencies may lead to erroneous inference.

In addition, the Bayesian paradigm offers a sound framework for decision-making. The decision-making process relies on *a posteriori* quantities of unknown parameters estimated on available evidence. As a consequence, the decision fully accounts for the uncertainty that characterizes our knowledge, leading to a decision-making process that is reliable and evidence-based.

Finally, MCMC procedures developed for fitting Bayesian models can easily be derived to implement fairly complex models tackling above listed challenges.

1.2 Technical Background

In this section, we briefly introduce some modeling tools that we use in the dissertation to obtain flexible models. What follows has no claim to completeness: the goal is to provide a brief overview of techniques the statistical methods developed in the following chapters build upon. In particular, we will present Varying Coefficient Models (Section 1.2.1), P-Splines (Section 1.2.2), and Product Partition Models with Covariates (1.2.3).

1.2.1 Varying coefficient models

Varying Coefficient Models (VCM) are a useful tool to explore dynamic patterns in the data. In particular, they are a flexible extension of classical linear regression models, of which they retain interpretability. VCMs were firstly proposed by Cleveland et al. (1991). In their original formulation, considering a scalar X and a $P-$dimensional vector of covariates $\boldsymbol{Z}$, the VCMs assumes the form of a multivariate regression function:

$$E(y|X, \boldsymbol{Z}) = \boldsymbol{Z}^\top \boldsymbol{\beta}(X),$$

for unknown functional coefficients $\boldsymbol{\beta}(X) = (\beta_1(X), \dots, \beta_P(X))^\top$.

VCMs and approaches for their estimation received great interest in recent years (see Fan and Zhang (2008) for a comprehensive review). VCMs are of particular interest to us because they allow the coefficients to vary smoothly over the groups stratified by X and hence permits nonlinear interactions between X and $\boldsymbol{Z}$.

Here we focus on the VCM generalization proposed by Hastie and Tibshirani (1993) where, in the framework of GLMs, they defined a class of models that are linear in the regressors, but their coefficients are allowed to change smoothly with the value of other variables termed "effect modifiers".

Considering n observations, indexed by $i = 1, \dots, n$, assume we observe a response vector $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ and P covariates $\boldsymbol{z}_1, \dots, \boldsymbol{z}_P$, where z_{ip} is the value of the $p-$th covariate observed for subject i. Considering a vector of coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)^\top$, in the setting of GLMs we relate the linear predictor $\eta_i = \boldsymbol{z}_i^\top \boldsymbol{\beta}$ to

the mean $\mu_i = E(y_i|\boldsymbol{z}_i)$ via the link function $g(\cdot)$, that is $g(\mu_i) = \eta_i$. In VCMs the linear predictor is defined as

$$\eta_i = \beta_0 + z_{i1}\beta_1(x_{i1}) + z_{i2}\beta_2(x_{i2}) + \cdots + z_{iP}\beta_P(x_{iP}), \tag{1.1}$$

where $x_{i1}, \ldots, x_{iP}$ are the effect modifiers. Basically, they are additional predictors that alter coefficients for $z_{i1}, \ldots, z_{iP}$ through unspecified functions $\beta_1(\cdot), \ldots, \beta_P(\cdot)$.

Effect modifiers can be both continuous or categorical variables. When $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P$ are continuous, then is natural to assume $\beta_p(\boldsymbol{x}_p)$, for $p = 1, \ldots, P$ to be smooth functions. Otherwise, if the effect modifiers are categorical variables such that $x_{ip} \in \{1, \ldots, K\}$, then $\beta_p(x_{ip})$ are step functions of the form $\sum_{k=1}^{K} \beta_{pk}\mathbb{1}_{[x_{ip}=k]}$, where $\mathbb{1}_{[\cdot]}$ is an indicator function. Given parameters $\beta_{p1}, \ldots, \beta_{pK}$, the $p-$th component of the linear predictor is $z_{ip}\beta_p(x_{ip}) = \sum_{k=1}^{K} z_{ip}\beta_{pk}\mathbb{1}_{[x_{ip}=k]}$. For categorical effect modifiers, such a structure leads to the identification of clusters of patients that share same effects.

Note that the unspecified functions $\beta(\cdot)$ can be modeled in a variety of ways and no predetermined form needs to be assumed on this function. This leaves great freedom in its specification and hence a careful construction of $\beta(\cdot)$ permits the pursuit of several modeling objectives.

It is straightforward to see two-way interactions as a special case of (1.1). Depending on the specification adopted for $\beta_1(\cdot), \ldots, \beta_P(\cdot)$, the interactions arising from (1.1) can be linear or non-linear (or both).

Finally, equation (1.1) can be seen as a generalization of several classes of models. In particular, the following models fall under the VCMs class:

GLM if $\beta_p(x_{ip}) = \beta_p$ for $p = 1, \ldots, P$, that is $\beta_p(\cdot)$ is a constant function and the terms are linear in $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_P$;

GAM (Generalized Additive Models) if $\boldsymbol{z}_p = 1$ for $p = 1, \ldots, P$, then $\beta_p(\boldsymbol{x}_p)$ is an unspecified function of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P$;

dGLM (dynamic Generalized Linear Models (West and Harrison, 1989)) if data consist of repeated measurements of $\boldsymbol{y}$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_P$ over S time points $t \in (t_1, \ldots, t_S)$ (time is the effect modifier), then (1.1) may be defined over time as

$$\eta_{it} = \beta_0(t) + z_{i1}(t)\beta_1(t) + \cdots + z_{iP}(t)\beta_P(t).$$

In the context of behavioral sciences, this class of model is also known as time-varying effect models (TVEM) (Tan et al., 2012).

1.2.2 P-splines

In many practical regression situations, common regression models, such as GLMs, may not be appropriate. This may be due, without limitation, to correlation patterns among observations, complex interactions among covariates and large heterogeneity among individuals. Proposing a more general class of models, Generalized Additive Models (GAM), Hastie and Tibshirani (1986) consider a general regression problem where observations $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, on a continuous response $\boldsymbol{y}$ and a vector of continuous covariates $\boldsymbol{x} = (x_1, \ldots, x_P)^\top$ are given. We also assume responses to be independent with predictor

$$\eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \cdots + f_P(x_{iP}), \tag{1.2}$$

for $i = 1, \dots, n,$ and $p = 1, \dots, P$ with common variance across observations.

Assuming the same number of knots for each function, f_p can be approximated by a spline of degree l partitioning the domain of the variable x_p with $r + 1$ equally spaced knots (Eilers and Marx, 1996)

$$x_{p,min} = \xi_{p,0} < \xi_{p,1} < \cdots < \xi_{p,r-1} < \xi_{p,r} = x_{p,max}.$$

This spline can also be represented as a combination of $D = r + l$ B-spline basis functions:

$$f_p(x_p) = \sum_{\rho=1}^{D} \beta_{p\rho} B_{p\rho}(x_p),$$

where $\boldsymbol{\beta}_p = (\beta_{p1}, \dots, \beta_{pD})^\top$ is a vector of unknown regression coefficients to be estimated. The basis functions $B_{p\rho}$ are defined only locally (they are nonzero only on a domain spanned by $2 + l$ knots). Moreover, we define the $n \times D$ design matrices $\boldsymbol{X}_p$, that are constructed such that the element in row i and column ρ is given by $X_p(i, \rho) = B_{p\rho}(x_{ip})$. We can then rewrite the predictor (1.2) in matrix notation as

$$\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{X}_P\boldsymbol{\beta}_P. \tag{1.3}$$

In the frequentist setting, regression coefficients are estimated using standard maximum likelihood algorithms for linear models.

The crucial point of spline regression is the selection of the number of knots and their positioning. In fact, the number of basis functions must be large enough to allow for sufficient flexibility, so that the estimated function provides an adequate fit. Nonetheless, a large number of basis may incur in overfitting, resulting in variable estimates. Eilers and Marx (1996) proposed the P-spline approach to approximate $f_p(x_p)$ using a linear combination of B-spline basis functions. Smoothness is achieved by imposing a roughness penalty based on differences of adjacent B-Spline coefficients; in order to guarantee sufficient smoothness of the fitted curves. This leads to penalized likelihood estimation where the penalized likelihood

$$L = l(y, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_P) - \lambda_1 \sum_{\rho=k+1}^{M} (\Delta^k \beta_{1\rho})^2 - \cdots - \lambda_P \sum_{\rho=k+1}^{M} (\Delta^k \beta_{P\rho})^2 \tag{1.4}$$

is maximized with respect to the unknown regression coefficients $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_P$. Moreover, Δ^k denotes the difference operator of order k. P-splines performances are highly dependent on the choice of λ_ps, usually selected via cross validation or on the ground of a goodness-of-fit criterion. However, for a large number of smoothing parameters, these procedures fail since the effort to compute an optimal solution (if there is any) becomes intractable. To overcome this limitation Lang and Brezger (2004) and Brezger and Lang (2006) developed a Bayesian approach to P-splines. Consistently with the Bayesian paradigm, the unknown parameters $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_P$ are random variables and priors are defined by replacing penalties in (1.4) with first or second order random walk. The general form of the prior for $\boldsymbol{\beta}_p$ is given by

$$p(\boldsymbol{\beta}_p | \tau_p^2) \propto \frac{1}{(\tau_p^2)^{rank(\boldsymbol{K}_p)/2}} \exp\Big(-\frac{1}{2\tau_p^2} \boldsymbol{\beta}' \boldsymbol{K}_p \boldsymbol{\beta}_p \Big),$$

where $\boldsymbol{K}_p$ is an appropriate penalty matrix. Since $\boldsymbol{K}_p$ is rank deficient, the prior is improper. Finally, the variance parameter τ_p^2 is distributed as an Inverse Gamma $p(\tau_p^2) \sim IG(a_p, b_p)$. It controls the trade-off between flexibility and smoothness, hence it corresponds to the (inverse) smoothing parameter in a penalized likelihood approach. The Gaussian prior proposed by Brezger and Lang (2006) results in a conditionally Gaussian prior. Scheipl et al. (2012) show that such prior can always be recast based on i.i.d. priors, through a suitable spectral decomposition. The procedure to obtain proper Gaussian priors proposed by Scheipl et al. (2012) is thoroughly described in Section 2.2.1.

Interestingly, Lang and Brezger (2004) extend their formulation to VCMs. In particular, they adopt a flexible approach based on nonparametric two-dimensional surface fitting to estimate the interaction between two continuous covariates x_p and x_q. The interaction term is modeled by a two-dimensional smooth surface $f_{pq}(x_p, x_q)$ leading to a predictor of the form

$$\eta_i = \cdots + f_p(x_{ip}) + f_{iq}(x_{iq}) + f_{pq}(x_{ip}, x_{iq}).$$

Furthermore, they assume that the unknown surface can be approximated by the tensor product of two one-dimensional B-spline used to model $f_p(x_p)$ and $f_q(x_q)$, respectively.

1.2.3 Covariate informed random partition

Probability models for random partitions are routinely used in Bayesian data analysis to uncover latent groups suspected in the data or to discover groups of homogeneous observations. The groups forming the partition are also referred to as clusters.

Random partition models are employed to obtain a partition of the data such that the non-overlapping groups are as dissimilar as possible and that the observations within the same group are as similar as possible. Similarity within groups (dissimilarity across groups) can be defined with respect to some covariates. In this case, covariates guide the construction of the partition to reveal a latent group structure that corresponds to unobserved heterogeneity. In this sense, the partition depends on available covariates.

Note that the field of dependent random partition is a very active area of research and several useful methods have been developed, see Park and Dunson (2010); Blei and Frazier (2011); Dahl et al. (2017); Paganin et al. (2021). The purpose of this section is not to give a comprehensive nor exhaustive review of the advancements in this field, but rather to provide some background context to covariate dependent partitions used in the following chapters.

In Section 1.2.3 we will present Random Partition Models that implicitly define a distribution on the partitions set, induced by a discrete random measure. Product partition models (Section 1.2.3), instead, explicitly define a distribution on the set of all partitions. Finally in Section 1.2.3 we present product partition models with covariates (PPMx), a prior on the partition that is informed by covariates, that is a model-based clustering approach.

Random Partition Models

Let us start with some notation. Let $[n] = \{1, \ldots, n\}$ denote a set of n subjects. A cluster arrangement $\Pi_n = (S_1, \ldots, S_{C_n})$ is a partition of $[n]$ into a set of groups

$\{S_j\}$ where $j = 1, \ldots, C_n$, with $S_j \cap S_{j'} = \emptyset$ for $j \neq j'$ and $\cup_{j=1}^{C_n} S_j = [n]$. That is Π_n consists of C_n nonempty and mutually exclusive subsets. Also, let $n_j = |S_j|$ denote the size of the $j-$th cluster. Moreover, we denote with $\mathcal{P}_n$ the set of all partitions of $[n]$, whose size is $\mathcal{B}_n$, the $n-$th Bell number. It is often convenient to represent a partition Π_n by cluster membership indicators $e_1, \ldots, e_n$ with $e_i \in [C_n]$ and $e_i = j \iff i \in S_j$, for $i \in [n]$ and $1 \leq j \leq C_n$.

A random probability model is a probability distribution over $\mathcal{P}_n$:

$$\{p(\Pi_n) : \Pi_n \in \mathcal{P}_n\}.$$

A common approach to defining $p(\Pi_n)$ is through discrete random probability measures (RPM). Let us consider a discrete distribution $G(\cdot) = \sum_{h=1}^{\infty} \omega_h \delta_{\psi_h}(\cdot)$ with probability masses ω_h in locations ψ_h with $P(\sum_{h=1}^{\infty} \omega_h = 1) = 1$. A simple Bayesian Nonparametric model is

$$\begin{aligned} \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \mid G &\overset{\text{iid}}{\sim} G \\ G &\sim RPM \end{aligned} \tag{1.5}$$

Equation (1.5) indirectly defines a random partition model. In fact, sampling $\boldsymbol{\theta}_i \mid G \overset{\text{iid}}{\sim} G$ for $i = 1, \ldots, n$ results in ties among $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ due to the discreteness of G. Denoting with $\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_{C_n}^\star$ the unique values, then Π_n can be obtained through cluster membership indicators:

$$e_i = j \iff \boldsymbol{\theta}_i = \boldsymbol{\theta}_j^\star \text{ or equivalently } \boldsymbol{\theta}_i = \boldsymbol{\theta}_{e_i}^\star.$$

Hence, $S_j = \{i \in [n] : \boldsymbol{\theta}_i = \boldsymbol{\theta}_j^\star\}$ defines a random partition of $[n]$.

The partition structure can also be investigated considering the exchangeable partition probability function (eppf). Denoting with $(N_1, \ldots, N_{C_n}) = (n_1, \ldots, n_{C_n})$ the relative frequencies of the C_n distinct values in $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, the eppf is the probability of observing a specific sample $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ with C_n unique values $\boldsymbol{\theta}_j^\star$ with frequencies $\boldsymbol{n} = (n_1, \ldots, n_{C_n})$. Let $N^\star = \cup_{k=1}^{\infty} N_k$ and let $\boldsymbol{n}^{j+}$ denote $\boldsymbol{n}$ with the $j-$th cluster size incremented by 1. Formally the eppf $p(\cdot)$ is a symmetric function $p : N^\star \to [0, 1]$ with

$$p(\boldsymbol{n}) = \sum_{j=1}^{C_n+1} p(\boldsymbol{n}^{j+}) \; \forall \; \boldsymbol{n} \in N^\star$$

and $p(1) = 1$. This condition formalizes coherence across sample sizes. This implies that a random partition model can be expressed as $p(\Pi_n = (S_1, \ldots, S_{C_n})) = P(n_1, \ldots, n_{C_n})$ for any eppf $p(\cdot)$.

The eppf is not always available, but a popular case is the Dirichlet Process (DP) Random Partition.

Example 1

The random partition induced by a DP random probability measure is known as the Pólya Urn:

$$\begin{aligned} G|\kappa, G_0 &\sim DP(\kappa G_0) \\ \boldsymbol{\theta}_i|G &\overset{\text{iid}}{\sim} G, \text{ for } i = 1, \ldots, n, \end{aligned}$$

where κ is the concentration parameter. Assuming that cluster labels are indexed by appearance, the implied prior is of the form

$$p(\Pi_n) = \frac{\kappa^{C_n-1}\prod_{j=1}^{C_n}(n_j-1)!}{\prod_{i=1}^{n}(\kappa+i-1)!}.$$

Product Partition Model

Hartigan (1990) and Barry and Hartigan (1993) proposed the product partition model (PPM), where the clustering is not induced by a discrete random measure, but $p(\Pi_n)$ is explicitly defined over $\mathcal{P}_n$. PPM defines a product partition probability for the random partition:

$$p(\Pi_n) \propto \prod_{j=1}^{C_n} \rho(S_j), \tag{1.6}$$

where $\rho(S_j)$ is a non–negative function, called the cohesion function, that measures how strongly one believes elements in S_j are to co-cluster a priori. The normalization constant in (1.6) is $\sum_{\Pi_n\in\mathcal{P}_n}\prod_{j=1}^{C_n}\rho(S_j)$.

Assume we have a set of responses $\boldsymbol{y} = (y_1,\ldots,y_n)$ and let $\boldsymbol{y}_j^\star = (y_i : i \in S_j)$ denote the responses arranged by cluster. Conditional on a given partition, PPM assumes independence across clusters:

$$p(\boldsymbol{y}|\Pi_n) = \prod_{j=1}^{C_n} p(\boldsymbol{y}_j^\star|\boldsymbol{\theta}_j^\star),$$

where $\boldsymbol{\theta}_j^\star$ are cluster-specific parameters. Exchangeability of y_i across $i \in S_j$ can be leveraged assuming that y_i are independent given $\boldsymbol{\theta}_j^\star$. The posterior distribution $p(\Pi_n|\boldsymbol{y})$ is again of product form, that is PPM is conjugate. The updated cohesion functions are $\rho(S_j)p(\boldsymbol{y}_j^\star)$, where $p(\boldsymbol{y}_j^\star)$ is the marginal sampling model for $y_i, i \in S_j$ given partition Π_n. Finally, the Pólya Urn implied by the DP is a special case of a PPM, with cohesion function $\rho(S_j) = \kappa(n_j-1)!$. Also the eppf of a Gibbs-type prior takes the form of a product partition distribution with cohesion function $\rho(S_j) = (1-\sigma)_{n_j-1}$, where $(a)_k = \Gamma(a+k)/\Gamma(a)$ denotes a rising factorial.

Covariate-Dependent PPMs

When for inferential purposes there is the need to define subpopulations of subjects that are homogeneous with respect to some baseline characteristics, it is useful to add dependence on covariates to the PPM model.

Müller et al. (2011) extended the PPM specifying a probability model for random partitions that favors clusters that are homogeneous in the covariates $\boldsymbol{x}$. That is, using a model $p(\Pi_n|\boldsymbol{x})$ together with a sampling model $p(\boldsymbol{y}|\Pi_n)$. Note that the conditioning of the partition on the covariates is just to stress this dependence and there is no notion of $\boldsymbol{x}$ being random. Let $\boldsymbol{x}_i = (x_{i1},\ldots,x_{iQ})$ be a $Q-$dimensional vector of covariates and let $\boldsymbol{x}_j^\star = (\boldsymbol{x}_i, i \in S_j)$ denote covariates arranged by clusters. The product partition model with covariates (PPMx) introduces a non-negative function $g(\boldsymbol{x}_j^\star)$ that formalizes homogeneity among covariate vectors, that is the

larger values of $g(\boldsymbol{x}_j^\star)$ indicate that subjects in cluster j are judged to be similar. A PPMx model with similarity function $g(\boldsymbol{x}_j^\star), g(\cdot) \geq 0$ is a random partition:

$$p(\Pi_n|\boldsymbol{x}) \propto \prod_{j=1}^{C_n} g(\boldsymbol{x}_j^\star)\rho(S_j),$$

with normalization constant $\sum_{\Pi_n \in \mathcal{P}_n} \prod_{j=1}^{C_n} g(\boldsymbol{x}_j^\star)\rho(S_j)$.

Müller et al. (2011) provide a discussion on the theoretical properties the *similarity* function should satisfy and some guidelines for its choice, but any non-negative function that guarantees an increasing value for close covariate values is suitable (Page and Quintana, 2018).

The default choice proposed by Müller et al. (2011) is to define g as the marginal probability of an auxiliary Bayesian model

$$g(\boldsymbol{x}_j^\star) = \int \prod_{i \in S_j} \prod_{q=1}^{Q} q(x_{iq}|\boldsymbol{\xi}_j^\star) q(\boldsymbol{\xi}_j^\star) \mathrm{d}\boldsymbol{\xi}_j^\star,$$

even if $\boldsymbol{x}_i$ are not considered random.

Choosing $q(x_{iq}|\boldsymbol{\xi}_j^\star)$ and $q(\boldsymbol{\xi}_j^\star)$ as a conjugate pair greatly simplifies analytical evaluation of $g(\boldsymbol{x}_j^\star)$.

1.3 Overview of Projects

This thesis consists of two projects in which we develop flexible statistical models for populations that feature significant heterogeneity. Motivated by the challenges that the emerging field of precision medicine poses, we formulate suitable methods to deliver personalized inference, taking advantage of the Bayesian framework.

In Chapter 2 we propose a model motivated by a microbiota cancer study designed to analyze data that exhibit a hierarchical structure induced by measurements from multiple body tissues. We develop a flexible regression model that accounts for patients' heterogeneity and microbial variability across districts to select significant associations between microbial compositions and covariates. Such flexibility is achieved by defining varying coefficients that include two-way interactions as a special case. We use penalized splines to model continuous interactions and penalize spline coefficients, preventing overfitting. Moreover, continuous interactions can be classified into linear or nonlinear ones by imposing a suitable prior on the smoothing parameter of the splines. Finally, varying coefficients depend on subject-specific parameters so that the resulting model captures dependence structures that vary across patients. This work is co-authored with Amedeo Amedei (University of Florence) and Francesco C. Stingo (University of Florence).

In Chapter 3 we present a predictive model for optimal treatment selection for oncological patients. The motivation for this method comes from an open problem in cancer genomics and personalized medicine. We devise a model to assign untreated patients to the treatment that ensures the largest benefit among the possible therapies. We account for various sources of heterogeneity by integrating prognostic and predictive biomarkers. Namely, predictive markers are useful for the identification of patients that are more likely to benefit from a particular therapy. Homogeneous groups of patients are obtained by constructing a random partition model, whose

cluster production is driven by predictive covariates. In particular, the class of product partition model with covariates (PPMx) induces a partition distribution that encourages patients with close covariates to co-cluster. This strategy is adopted to characterize the extent of benefit offered by each therapy on groups of patients with similar predictive determinants. This work is co-authored with Raffaele Argiento (University of Bergamo) and Francesco C. Stingo (University of Florence).

Chapter 4 concludes this thesis with a brief discussion and indicates directions for future research.

Chapter 2

Subject-specific Dirichlet-multinomial regression for multi-district microbiota data analysis

2.1 Introduction

The human microbiota is the set of microbes that the human body harbors. Its composition is highly diverse due to several host traits such as genotype, physiological status, and lifestyle (Turnbaugh et al., 2007; Sanz et al., 2015). Microbiota's importance in disease and metabolic dysfunction has been increasingly recognized (La Rosa et al., 2012; Li, 2015). Understanding microbiome dynamics can provide precious insights on human health. In fact, changes in the regular microbiome composition have been reported to be correlated with many diseases, including diabetes, obesity, inflammatory bowel disease, autoimmune diseases, and neurodegenerative disease (Vieira et al., 2014; Matsuoka and Kanai, 2015; Tai et al., 2015; Mandrioli et al., 2019). In addition, investigating the genetic diversity of microbial populations and their role is now crucial in non-infectious diseases such as cancer, where specific bacteria have been demonstrated to influence the carcinogenesis' process (Zhang and Sun, 2018), especially in colorectal cancer (CRC) (Russo et al., 2018; Niccolai et al., 2020).

Recent DNA sequencing technologies can be used to evaluate the composition of the microbiome; statistical analysis of these measurements can shed light on interactions between microbiota and host. Microbiota data are usually obtained targeting 16S rRNA. The 16S rRNA gene of the bacteria in the samples is sequenced and the resulting reads are clustered into operational taxonomic units (OTUs). OTUs are defined as a cluster of reads that show 97% similarity. The membership count of sequences in each cluster is considered to measure the abundance of *taxa* in each sample. From this procedure an $n \times J$ OTUs table is obtained, where n is the number of samples read and J is the number of *taxa*. OTUs abundance table can be used to identify which factors regulate microbiota composition and whether any associations may be established between biological or genetic traits and microbiota abundance. Microbiome data are typically challenging as OTUs tables exhibit zero-inflation,

overdispersion, complex correlation structures, and high-dimensionality. Moreover, microbiome data have compositional structure, due to fixed sequencing depth.

The relationship between predictors and microbe abundance is often explored through regression models for multivariate count data. Chen and Li (2013b) proposed a Dirichlet-multinomial (DM) regression to test the association between microbiome composition and covariates and developed a penalized likelihood approach to induce sparsity by imposing a ℓ_1 penalty. Regularization and different penalties are explored in Zhang et al. (2017), where generalized linear models that incorporate various correlation structures among counts have been studied and compared. Bayesian methods have proved to capture model selection uncertainty better than constrained optimization approaches, especially in high-dimensional and highly-correlated settings. Wadsworth et al. (2017) adopted a DM regression framework imposing *spike-and-slab* priors, which led to better significant associations recovery. Ren et al. (2020) proposed a generalized mixed-effects linear model where the marginal prior on each microbial composition is a Dirichlet process and dependence across compositions is induced through a linear composition of individual covariates and latent factors. In a longitudinal setting, considering a negative binomial regression model, Shuler et al. (2018) extended conventional nonlocal priors (Rossell and Telesca, 2017) where the hierarchical model construction fosters *borrowing strength* across OTUs, while Martin et al. (2019) extended DM regression by using random effects to account for correlation between time points in the repeated-measurement setting.

Several methods exploit microbial abundances (compositional covariates) to explain clinical variables, this approach is called *compositional regression*. Lin et al. (2014) formulates a ℓ_1 regularization method for the linear log-contrast model taking OTUs tables as compositional covariates. In this context, Bayesian methods successfully account for the whole phylogenetic tree, using structured prior for joint selection of closely related organisms (Zhang et al., 2020), or focusing on the difference in microbiota composition across groups with graphical models that takes advantage of a phylogenetic scan test (Tang et al., 2018; Mao and Ma, 2020). An interesting approach that unifies multivariate regression for count data and compositional regression is a Bayesian joint model proposed by Koslovsky et al. (2020) that performs variable selection on clinical covariates associated with microbial compositional data that are concurrently used for the prediction of continuous response.

The objective of this project is to develop a hierarchical Bayesian model for the analysis of hierarchically structured microbiota count data. We apply our approach to available measurements of dietary habits and biological samples collected in a study conducted on patients affected by CRC from the Florence metropolitan area in central Italy (Niccolai et al., 2020). The study was designed to assess associations between available covariates and different microbiota districts counts. In fact, for each patient up to three biological samples have been collected (tumor, fecal and salivary samples). The proposed method accommodates both the hierarchical structure of the data and the correlation structure induced by repeated measurements on the same patients.

Cancer is inherently heterogeneous, and the approach to cancer treatment is evolving from standard procedures like chemotherapy to tailored treatments termed *personalized medicine* (De Bono and Ashworth, 2010). Different patients have diverse responses to the same treatment, even for the same type of cancer, due to individual variability in genes, clinical variables, and environmental features. The

proposed approach takes into account patients' heterogeneity even for small sample sizes. This is possible because *subject-specific parameters* are allowed to vary across patients to accommodate for their heterogeneity, while patients with similar profile are encouraged to have close effects. Note that *subject-specific* here does not refer to the subject-level effect as is typical in mixed models, but rather to coefficients that vary as a function of a subject's covariate pattern.

Our approach builds upon the Dirichlet-multinomial regression framework (Wadsworth et al., 2017) and provides three novel features. Firstly, our approach includes subject-specific coefficients that can capture a wide range of heterogeneous effects, *borrowing strength* among patients with similar clinical profiles and close dietary regimes. Consequently, coefficients are allowed to vary at the subject level; the model can identify subgroups of patients characterized by similar biological mechanisms (precision medicine). Secondly, microbial community composition and function is not constant across body districts; the proposed method combines information from multiple body districts and learns which effects change and which ones remain constants. Finally, our method explicitly models two-way interactions. Dealing with interactions is equivalent to assuming that the effects of predictors depend on the value of other covariates; specifically, interactions among predictors are modeled as subject-specific coefficients, hence the proposed model captures dependence structures that vary across patients.

The rest of the Chapter is organized as follows. Section 2.2 contains the model formulation, Section 2.3 introduces prior specification; we briefly outline the posterior computation steps in Section 2.4. In Section 2.5, finite sample performances are evaluated via simulation studies. In Section 2.6, we illustrate and discuss our analysis of the CRC data. Section 2.7 concludes the chapter with a brief discussion.

2.2 Bayesian Hierarchical Subject-specific Dirichlet-multinomial Regression Model

In this section we describe our methodological approach. In Section 2.2.1 we review the Dirichlet-multinomial distribution, and highlight the features that make this model appropriate for the analysis of microbiota data. In Section 2.2.2, we present our novel approach, a Dirichlet-multinomial regression model with subject-specific coefficients.

2.2.1 Dirichlet-multinomial model for microbiota composition

Considering J microbial *taxa*, let $\boldsymbol{Y} = (Y_1, \ldots, Y_J)$ be the random vector of the corresponding counts that follows a multinomial distribution:

$$f_{\boldsymbol{Y}|\boldsymbol{\phi}}(y_1, \ldots, y_J|\boldsymbol{\phi}) = \binom{y^+}{\boldsymbol{y}} \prod_{j=1}^{J} \phi_j^{y_j}$$

where $y^+ = \sum_{j=1}^{J} y_j$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J)$ are taxa proportions defined on the $J-$dimensional simplex

$$\mathcal{S}^{J-1} = \{(\phi_1, \ldots, \phi_J) : \phi_j \geq 0, \forall j, \sum_{j=1}^{J} \phi_j = 1\}.$$

Note that in our notation upper case letters denote random variables, and we refer to their particular realization with the corresponding lower case letter.
Since microbiota compositions are highly heterogeneous we need to let proportions vary across samples. To account for overdispersion a conjugate prior is imposed on the taxa proportions, that is $\boldsymbol{\phi} \sim$ Dirichlet($\boldsymbol{\gamma}$), where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ is a $J-$dimensional vector with generic strictly positive entry $\gamma_j > 0$ (Chen and Li, 2013b). This hierarchical structure leads to the Dirichlet-Multinomial distribution, introduced by Mosimann (1962). In fact, integrating the weights $\boldsymbol{\phi}$ out we have:

$$f_{\boldsymbol{Y}|\boldsymbol{\gamma}}(y_1, \ldots, y_J|\boldsymbol{\gamma}) = \frac{\Gamma(y^+ + 1)\Gamma(\gamma^+)}{\Gamma(y^+ + \gamma^+)} \prod_{j=1}^{J} \frac{\Gamma(y_j + \gamma_j)}{\Gamma(\gamma_j)\Gamma(y_j + 1)}$$

where $\gamma^+ = \sum_{j=1}^{J} \gamma_j$ and $\Gamma(\cdot)$ is the Gamma function. The first two moments of the DM distribution are

$$E(\boldsymbol{Y}) = y^+ \frac{\boldsymbol{\gamma}}{\gamma^+}, \quad cov(\boldsymbol{Y}) = y^+ \frac{\gamma^+ + y^+}{\gamma^+ + 1} \left\{ \text{diag}\left(\frac{\boldsymbol{\gamma}}{\gamma^+}\right) - \left(\frac{\boldsymbol{\gamma}}{\gamma^+}\right)\left(\frac{\boldsymbol{\gamma}}{\gamma^+}\right)^T \right\}.$$

The correlation among counts is negative and is affected by the variance inflating factor $\frac{\gamma^+ + y^+}{\gamma^+ + 1}$ (Wadsworth et al., 2017; Zhang et al., 2017). Note that γ^+ controls the degree of overdispersion; larger values of γ^+ correspond to smaller variances.

2.2.2 Dirichlet-multinomial subject-specific regression model

The DM distribution is more flexible than other models for multivariate count data since it can account for extra variation in the proportions and overdispersion. This implies that a DM distribution is particularly suited for the analysis of ecological count data, and microbiome in particular (Harrison et al., 2020).

The aim of our method is to select associations between covariates and *taxa* abundances. Suppose we have n microbiome samples and J *taxa*. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})$ represent the $J-$dimensional response vector of microbial *taxa* abundance counts, where y_{ij} denotes the observed count of OTU j collected from the $i-$th sample, with $i = 1, \ldots, n$, that is:

$$\boldsymbol{Y}_i \sim \text{ Multinomial}(y_i^+ \mid \boldsymbol{\phi}_i), \tag{2.1}$$

with $y_i^+ = \sum_{j=1}^{J} y_{ij}$ and $\boldsymbol{\phi}_i$ defined on the $J-$dimensional simplex $\mathcal{S}^{J-1}$. To account for overdispersion in the counts we specify a conjugate prior on the taxa probability:

$$\boldsymbol{\phi}_i \sim \text{ Dirichlet}(\boldsymbol{\gamma}_i),$$

with the $J-$dimensional vector $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iJ}), \gamma_{ij} > 0 \; \forall \; j$. Let $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{nj})$ be the $n-$dimensional vector of strictly positive parameters for the DM distribution.

In order to relate covariates to *taxa* abundances, we assume that the parameter γ_{ij} in the DM model depends on the covariates via a log-linear regression model (Chen and Li, 2013b; Wadsworth et al., 2017; Koslovsky et al., 2020):

$$\log(\gamma_{ij}) = g_j(\boldsymbol{u}_i), \tag{2.2}$$

where $g_j(\cdot)$ is a generic function, and $\boldsymbol{u}_i$ is the $1 \times M$ vector of the observed M covariates for sample i. The growth of a bacterial population usually occurs in at exponential rate, when placed in a favorable medium. Hence the log-linear link is an assumption commonly considered as appropriate (Chen and Li, 2013b).

The function $g_j(\cdot)$ captures the effect that each covariate has on the abundances of each patient, and it is usually set to be a linear combination of the covariates. In our framework, we let $g_j(\cdot)$ depend on a set of *subject-specific* coefficients.

Subject-specific coefficients arise when constraints on the parameters need to be relaxed. In certain applications, it is reasonable to let parameters change with covariates. In particular, we let covariates affect the parameters linearly and nonlinearly. We assume that subjects with similar covariate patterns are likely to have similar values of the parameters; this approach allows the model to borrow strength from similar individuals. Moreover, the model includes body district (i.e., tumor, fecal, and salivary) specific coefficients that can additionally vary at the patient level, in an effort to fully account for patient heterogeneity. In summary, we want to investigate whether similar clinical profiles result in similar alterations of the microbiota composition; an approach based on subject-specific associations provides the needed flexibility and avoids the *"one-size-fits-all"* approach of models with only population-level parameters. Specifically, let $\boldsymbol{u}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$ be the observed values of continuous covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{iP})$ and binary factors $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq}, \ldots, z_{iQ})$ for sample i, with $M = P + Q$. Moreover, let $\boldsymbol{x}_p$ and $\boldsymbol{z}_q$ denote n samples of the $p-$th continuous covariate and $q-$th factor, respectively. Given the large number of possible effects, we want the model to select which effects are relevant. In order to achieve this goal, we let $g_j(\cdot)$ depend on varying coefficients and selection mechanisms. Varying coefficients will vary with the data, both in terms of sparsity and magnitude:

$$\log(\gamma_{ij}) = \mu_j + \sum_{m=1}^{M} \beta_{mj}(\boldsymbol{u}_i) u_{im}, \tag{2.3}$$

where the subject-specific coefficient $\beta_{mj}(\boldsymbol{u}_i)$ is an unknown smooth function of $\boldsymbol{u}$ (Ni et al., 2019a). The summation term in equation (2.3) is defined as follows:

$$\sum_{m=1}^{M} \beta_{mj}(\boldsymbol{u}_i) u_{im} = \sum_{p=1}^{P} \beta_{pj}(\boldsymbol{x}_i, \boldsymbol{z}_i) x_{ip} + \sum_{q=1}^{Q} \beta_{qj}(\boldsymbol{z}_i) z_{iq}. \tag{2.4}$$

Note that equation (2.4) includes linear interactions as a special case; moreover, our approach based on varying coefficients (Hastie and Tibshirani, 1993) can include more flexible (nonlinear) interaction terms. In our approach, covariates serve two purposes: on one hand, they enter equation (2.4) as linear predictors with associated linear/main effects, on the other hand, coefficients change smoothly with the values of the remaining covariates, i.e., all remaining covariates can potentially act as *effect modifiers* (Hastie and Tibshirani, 1993). This double role poses identifiability issues: when covariates are either main effects or *effect modifiers*, like in Ni et al. (2019a)

and Ni et al. (2019b), no identifiability issue occurs. In our model, we develop a careful construction of subject-specific coefficients to ensure identifiability in the likelihood. The subject-specific coefficients, as introduced in equation (2.4), depend on generic functions $\{\beta_{pj}(\cdot)\}, \{\beta_{qj}(\cdot)\}$ of the covariates. Specifically, we model these functions as:

$$\beta_{pj}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \theta_{pj} + h\left(\sum_{q=1}^{Q} b_{pqj} z_{iq} + \sum_{k>p}^{P} f_{pkj}(x_{ik}), t_x\right) \tag{2.5a}$$

$$\beta_{qj}(\boldsymbol{z}_i) = \theta_{qj} + h\left(\sum_{l>q}^{Q} b_{qlj} z_{il}, t_z\right). \tag{2.5b}$$

The subject-specific coefficient is a structured additive model defined by two sets of parameters: main effects $\{\theta_{pj}\}, \{\theta_{qj}\}$ and interactions effects $\{b_{pqj}\}, \{b_{qlj}\}, \{f_{pkj}(\cdot)\}$, plus a thresholding function $h(\cdot)$ described at the end of this Section. In equations (2.5a) and (2.5b), $\{b_{pqj}\}$ and $\{b_{qlj}\}$ are parameters that estimate the adjustment provided by discrete covariates $\{\boldsymbol{z}_q\}$ to main effects of continuous and discrete covariates, $\{\theta_{pj}\}$ and $\{\theta_{qj}\}$ respectively. Note that the $\{b_{qlj}\}$ coefficients are included in the model only if $l > q$; this simple constrain ensures model identifiability.

We take a semiparametric approach to model the adjustment provided by continuous covariates to $\{\theta_{pj}\}$ and set $\{f_{pkj}(\cdot)\}$ to be penalized splines. The functions $\{f_{pkj}(\cdot)\}$ can effectively capture nonlinear interactions; in order to make the likelihood identifiable, we include in the model only $\{f_{pkj}(\cdot)\}$ such that $k > p$. Moreover, continuous covariates are effect modifiers of the main effect of other continuous covariates through the term $\sum_{k>p} f_{pkj}(x_{ik})$ in equation (2.5a); continuous covariates do not modify the main effects of binary factors. This construction results in an "asymmetric" definition of $\{\beta_{pj}(\cdot)\}$ and $\{\beta_{qj}(\cdot)\}$, that nevertheless ensures identifiability for model parameters in the likelihood. In fact, including continuous covariates as effect modifiers also for $\{\theta_{qj}\}$ would have resulted in a clear overparameterization of the model. Simulation studies to assess inference invariance to both the ordering of the covariates and to ties imposed to ensure identifiability in the likelihood are reported in Appendix A.1 along with a detailed discussion. The functions $f(\cdot)$ are specified following the penalized spline approach proposed by Scheipl et al. (2012). Specifically, we set $f_{pkj}(\boldsymbol{x}_k) = \tilde{\boldsymbol{x}}_k \boldsymbol{\alpha}_{pkj}$, where $\tilde{\boldsymbol{x}}_k$ represents the design matrix of the spline bases for $\boldsymbol{x}_k$ and $\boldsymbol{\alpha}_{pkj}$ are the corresponding spline coefficients. P-splines can be treated as a Bayesian hierarchical model (Ruppert et al., 2003) assuming $\boldsymbol{\alpha}_{pkj} \sim N(0, s\boldsymbol{K}^-)$ with the singular penalty matrix $\boldsymbol{K}$ constructed from the second-order differences of the adjacent spline coefficients. Since the coefficients that parameterize the constant and linear trends of $\{f_{pkj}(\cdot)\}$ are in the null space of $\boldsymbol{K}$, they are not penalized. Then we take a spectral decomposition of the covariance of $\tilde{\boldsymbol{x}}_k \boldsymbol{\alpha}_{pkj}$:

$$cov(\tilde{\boldsymbol{x}}_k \boldsymbol{\alpha}_{pkj}) = s\tilde{\boldsymbol{x}}_k \boldsymbol{K}^- \tilde{\boldsymbol{x}}_k^T = s[\boldsymbol{u}_k \;\; \boldsymbol{u}_0] \begin{bmatrix} \boldsymbol{d}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} [\boldsymbol{u}_k \;\; \boldsymbol{u}_0]^T,$$

where $\boldsymbol{u}_k$ is the orthonormal matrix of eigenvectors with corresponding positive eigenvalues along the diagonal of matrix $\boldsymbol{d}_k$, while $\boldsymbol{u}_0$ are the eigenvectors associated with the zero eigenvalues. The smooth functions $\{f_{pkj}(\cdot)\}$ can now be re-defined as the sum of nonlinear (penalized) term and a linear (non-penalized) term, that is $f_{pkj}(\boldsymbol{x}_k) = \boldsymbol{x}_k^\star \boldsymbol{\alpha}_{pkj}^\star + \boldsymbol{x}_k \alpha_{pkj}^0$. The penalized term is $\boldsymbol{x}_k^\star \boldsymbol{\alpha}_{pkj}^\star$, with $\boldsymbol{x}_k^\star = \boldsymbol{u}_k \boldsymbol{d}_k^{\frac{1}{2}}$ that is

the orthogonal basis. In particular, $\boldsymbol{\alpha}^{\star}_{pkj}$, the nonlinear effect, is a r_k-dimensional vector, where r_k is the number of eigenvectors and eigenvalues that explain a majority of the variability of $f_{pkj}(\boldsymbol{x}_k)$. In fact, to enhance computational efficiency we retain only the first several eigenvectors and eigenvalues that explain 99% of the variability of $f_{pkj}(\boldsymbol{x}_k)$, obtaining substantial reduction in dimensionality. Finally, α^{0}_{pkj} is the coefficient associated with the linear term $\boldsymbol{x}_k$. The varying-coefficients for the continuous covariates are as follows:

$$\beta_{pj}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \theta_{pj} + h\left(\sum_{q=1}^{Q} b_{pqj} z_{iq} + \sum_{k>p} (\boldsymbol{x}^{\star}_{ik} \boldsymbol{\alpha}^{\star}_{pkj} + x_{ik} \alpha^{0}_{pkj}), t_x \right). \tag{2.6}$$

The intercepts are merged into the global term θ_{pj}. With respect to standard approaches for P-splines, the procedure proposed by Scheipl et al. (2012) results in separate coefficients for linear and nonlinear effects. We can then assign specific selection priors to α^{0}_{pkj} and $\boldsymbol{\alpha}^{\star}_{pkj}$; consequently, interactions among continuous covariates can have linear or nonlinear forms, or be excluded from the model, as detailed in Section 2.3.

We finally comment on the selection mechanisms. The subject-specific coefficients feature a thresholding function $h(\cdot)$; this function $h(\vartheta, t) = \vartheta \mathbb{1}_{[|\vartheta|>t]}$ depends both on its argument ϑ, which in our case is a combination of covariates value and regression parameters, and the random thresholding parameter t, and it is discussed in Section 2.3 along with variable selection and shrinkage priors that jointly induce model sparsity.

2.3 Prior Distributions

In this section, we discuss the prior choice for all parameters and detail the thresholding and selection mechanisms. Hyperparameter setting and sensitivity analysis are presented and discussed in Section 2.5.2 and Appendix A.2, A.3.

2.3.1 Main effects

The main effects $\{\theta_{pj}\}$ and $\{\theta_{qj}\}$ represent the associations between the multivariate response and the continuous and binary variables, respectively. Note that main effects do not directly produce subject-specific effects, that are only driven by the interaction parameters. We implement an approach based on *spike-and-slab* priors, i.e., a two-component mixture prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1997). We introduce J latent $P-$dimensional indicator vectors $\boldsymbol{\xi}_j = (\xi_{1j}, \ldots, \xi_{Pj})$ such that:

$$\theta_{pj} \mid \xi_{pj}, \tau_j^2 \sim \xi_{pj} N(\theta_0, \tau_j^2) + (1 - \xi_{pj}) \delta_0(\theta_{pj}),$$

for $p = 1, \ldots, P,\ j = 1, \ldots, J$, where $\delta_0(\theta_{pj})$ is a Dirac delta function at 0. If $\xi_{pj} = 1$, the $p-$th covariate affects the abundance of the $j-$th *taxon*, and if $\xi_{pj} = 0$ otherwise. We assume independent Bernoulli priors for the latent vectors:

$$p(\boldsymbol{\xi}_j \mid \boldsymbol{\omega}_j) = \prod_{p=1}^{P} \omega_{pj}^{\xi_{pj}} (1 - \omega_{pj})^{(1-\xi_{pj})}$$

where $\omega_{pj} \sim \text{Beta}(a_\omega, b_\omega)$; this prior has been shown to provide an automatic adjustment for multiplicity (Scott and Berger, 2010) and it is equivalent to placing a Beta mixed Binomial on ξ_{pj}. Prior distributions on the $\{\theta_{qj}\}$ are analogously defined.

2.3.2 Interactions parameters

Accounting for potential relationships among predictors could avoid inaccurate estimates and erroneous best subset selection; excluding interactions is equivalent to assuming that there is a lack of synergistic or antagonist effects, which is often a questionable assumption (Chipman, 1996; Gustafson, 2000). We will discuss the priors associated with the function $f(\cdot)$ later and we now focus on the prior placed on $\{b_{pqj}\}$, $\{b_{qlj}\}$. These coefficients represent interaction terms that involve discrete covariates. Whereas the main effects previously discussed are population-level parameters (i.e., they are constant across subjects), the inclusion of these terms into the model results in subject-level coefficients.

In order to build a computationally efficient approach, we opt for a horseshoe prior (Carvalho et al., 2009, 2010), an absolutely continuous mixture distribution that belongs to the class of global-local scale mixtures of normals. The prior for $\{b_{pqj}\}$ can be summarized as follows:

$$\begin{aligned} b_{pqj} &\sim N(0, \lambda^2_{pqj}\zeta^2_{qj}) \\ \lambda_{pqj} &\sim C^+(0,1), \\ \zeta_{qj} &\sim C^+(1/n,1), \end{aligned}$$

where C^+ denotes a half-Cauchy distribution, $\{\lambda_{pqj}\}$ are local shrinkage parameters, and $\{\zeta_{qj}\}$ are global shrinkage parameters. All coefficients will be nonzero, nonetheless, only associations supported by the data will have large values, due to the heavy tails of the prior. The interaction terms among discrete variables $\{b_{qlj}\}$ follow this same prior construction.

The decomposition and reparameterization of $\{f_{pkj}(\cdot)\}$ -detailed in equation (2.5a)- defines linear effects $\{\alpha^0_{pkj}\}$ and nonlinear effects $\{\boldsymbol{\alpha}^\star_{pkj}\}$. In our model, we assume that the relationship between covariates and counts is sparse, that is, only a small number of association is non-zero. To define a suitable prior we impose a parameter-expanded normal-mixture-of-inverse-gamma (peNMIG) prior on both $\{\alpha^0_{pkj}\}$ and $\{\boldsymbol{\alpha}^\star_{pkj}\}$. This prior is particularly suited for the simultaneous selection, or exclusion, of vectors of coefficients, such as coefficients associated with spline basis functions or with the levels of random intercepts (Gelman et al., 2008; Scheipl et al., 2012). We expand $\boldsymbol{\alpha}^\star_{pkj} = \eta_{pkj}\tilde{\boldsymbol{\psi}}_{pkj}$ to be a product of a scalar η_{pkj} and a vector $\tilde{\boldsymbol{\psi}}_{pkj}$ with the same size as $\boldsymbol{\alpha}^\star_{pkj}$ (that is r_k). This technique enables us to select jointly the batch of coefficients for each penalized term. The scalar η_{pkj} is assigned a spike-and-slab prior, while $\tilde{\boldsymbol{\psi}}_{pkj}$ distributes η_{pkj} across the entries of $\boldsymbol{\alpha}^\star_{pkj}$. This construction induces on $\boldsymbol{\alpha}^\star_{pkj}$ a parameter-expanded normal-mixture-of-inverse-gamma (peNMIG) prior (Scheipl et al., 2012). The hierarchical structure of these prior distributions is summarized as follows:

$$
\begin{aligned}
\eta_{pkj} \mid \tilde{\xi}_{pkj}, \tilde{\tau}^2_{pkj} &\sim N(0, \tilde{s}^2_{pkj}), \text{ with } \tilde{s}^2_{pkj} = \tilde{\xi}_{pkj}\tilde{\tau}^2_{pkj} \\
\tilde{\xi}_{pkj} \mid \tilde{\omega}_{pj} &\sim \tilde{\omega}_{pj}\delta_1(\tilde{\xi}_{pkj}) + (1-\tilde{\omega}_{pj})\delta_{v_0}(\tilde{\xi}_{pkj}) \\
\tilde{\tau}^2_{pkj} &\sim \text{Inverse-gamma}(a_{\tilde{\tau}}, b_{\tilde{\tau}}) \\
\tilde{\omega}_{pj} &\sim \text{Beta}(a_{\tilde{\omega}}, b_{\tilde{\omega}}).
\end{aligned}
$$

The hyperparameter v_0 is set to a small positive constant; this choice was proposed by George and McCulloch (1993) and has the advantage, with respect to the setting $v_0 = 0$, that transdimensional MCMC are avoided. Entries of the vector $\tilde{\boldsymbol{\psi}}_{pkj}$ are identically and independently distributed as $N(m_{pkj}, 1)$ with prior mean $m_{pkj} \sim 0.5\delta_1(m_{pkj}) + 0.5\delta_{-1}(m_{pkj})$. This discrete mixture implies $E[\tilde{\boldsymbol{\psi}}_{pkj} \mid m_{pkj}] = \pm 1$, hence the gain is twofold: i) since $\mid \tilde{\boldsymbol{\psi}}_{pkj} \mid = 1$, the scale and then the interpretation of $\boldsymbol{\alpha}^{\star}_{pkj}$ is preserved; ii) very little mass is given to values close to zero a priori, an approach that helps the model to identify non-zero effects. A Beta hyperprior is assumed for $\{\tilde{\omega}_{pj}\}$, which automatically adjusts for multiplicity. Through this hierarchical prior we easily select a batch of coefficients through a single latent indicator $\tilde{\xi}_{pkj}$. If $\tilde{\xi}_{pkj} = 1$, then the vector $\boldsymbol{\alpha}^{\star}_{pkj}$ is included in the model as a nonlinear effect. If $\tilde{\xi}_{pkj} = v_0$ then $p-$th continuous predictor's interactions with other continuous covariates will have a negligible effect. The parameter α^0_{pkj} follows the same prior as $\boldsymbol{\alpha}^{\star}_{pkj}$. In this way, we can explore the potential interactions among continuous covariates in several directions, namely whether the interaction exists, whether the interaction term is linear, or whether it exhibits nonlinearity.

The inclusion of interaction effects may depend on the inclusion of the corresponding main effects. A common assumption made on two-way interactions is *strong heredity*, which states that an interaction term can be included only if both its main effects are in the model. *Weak heredity* is a lesser stringent assumption and requires only one of the main effects to be included in the model to admit the interaction term (Griffin et al., 2017). Depending on the analysis of interest, these additional assumptions can be made part of the proposed model.

The need for different sparsity-inducing priors is due to the diverse inferential goal we want to achieve through each model component. We chose a *spike-and-slab* priors for main effects' coefficients in order to obtain truly sparse solutions. This is possible because the latent indicators $\{\xi_{pj}, \xi_{qj}\}$ allow to sharply include or exclude the effect of a covariate on a given taxon. This feature (that is not shared by horseshoe priors) turns out to be particularly useful as the inclusion of the interaction effect between two covariates is conditioned on the inclusion of their main effects, according to a given *hereditary assumption* (Griffin et al., 2017). We closely follow the approach developed by Scheipl et al. (2012) for $\{f_{pkj}(\cdot)\}$. In fact, this hierarchical prior allows us to simultaneously select the batch of coefficients associated with the spline bases obtained from each continuous covariate that act as effect modifier for the varying coefficient for continuous covariates. Adopting a *spike-and-slab* prior for interaction terms involving discrete covariates would have led to a more unified strategy. Nonetheless, the computational cost would have been excessive, as it would have required the specification of a latent indicator parameter of inclusion for each interaction term associated with each *taxon*. The horseshoe prior represented a more computationally efficient solution. On the other hand, we could have adopted the horseshoe prior even for main effects, as in Griffin et al. (2017), but it would have been rather complex, if not unfeasible, to impose any given hereditary assumption.

2.3.3 Thresholding mechanism

An unconstrained estimation of $\{\beta_{mj}(\boldsymbol{u})\}$ may result in many effects with negligible magnitudes. In fact both the horseshoe and the peNMIG prior do not set coefficients exactly to zero. We do not expect variables to have a little or negligible effect on each patient; we then include a soft thresholding mechanism that operates at the subject level, so that small effects are set to zero. The threshold parameters t_x, t_z are latent variables with a straightforward interpretation as minimum effect size parameters. Note that for threshold parameters to be identifiable, it is sufficient that a single interaction in equations (2.5a, 2.5b) exceeds them. A key feature of this thresholding mechanism is the randomness in both the argument and the threshold parameter (i.e., both are random variables); consequently, the thresholding mechanism accounts for both the magnitude and the variability of the effects. Since we have no prior knowledge of the true minimum effect size, we chose a noninformative prior on the threshold parameters: $t_x, t_z \sim \text{Unif}(0, b_t)$. Finally, note that since covariates enter the first argument of the thresholding function, consequently the resulting coefficients are subject-specific. The sparsity induced by the thresholding function depends on the covariates that act as effect modifiers, hence the varying coefficient will be able to consider as negligible dissimilar effects for each patient. Hypothetically, threshold functions and horseshoe priors could be replaced by *spike-and-slab*-like priors. This modeling approach would have required $60,610$ latent indicators instead of the 330 needed by the proposed model to obtain subject-specific coefficients.

2.3.4 Intercepts

Samples obtained from different districts (tumor, fecal and salivary samples) of the same patient are not independent. To deal with correlation due to repeated measurements within the same subject, Martin et al. (2019) proposed an extension of the DM regression model that includes a normally distributed random effect. In order to account for the correlation structure existing among counts of samples collected from different districts of the same patient we introduce a random intercept $\iota_{s(i)}$ in the linear predictor, where the subscript $s(i)$ denotes the patient from whom the $i-$th sample is taken. Conditional on this random effect, counts are independent (Martin et al., 2019). The variance of this random effect captures the correlation between the counts of the same subject across districts (Gelman et al., 2006):

$$\iota_{s(i)} \sim N(0, \sigma^2_\iota), \sigma^2_\iota \sim \text{Inverse-gamma}(a_{\sigma^2_\iota}, b_{\sigma^2_\iota}).$$

Note that the random intercept $\iota_{s(i)}$ is univariate, hence it is shared by all taxa. In fact, as an alternative, it could be defined as a $J-$dimensional vector of random effects. Since the variance of the random effect σ^2_ι captures the correlation structure among the samples of the same subject, rather than extra-heterogeneity, we avoided a category-specific random term to escape computational burden, since a multivariate random vector would have required a $J \times J$ covariance matrix. We complete the model specifying the global intercept term μ_j, which is *taxon*-specific and corresponds to the log baseline parameter for the *taxon* j. We assume the intercept terms μ_j follow a $N(0, \sigma^2)$; we set σ^2 to a large value to make this prior weakly informative.

We can now define the linear predictor introduced in equation (2.2):

$$\log(\gamma_{ij}) = \mu_j + \iota_{s(i)} + \sum_{m=1}^{M} \beta_{mj}(\boldsymbol{u}_i) u_{im}. \tag{2.7}$$

The telescopic structure of the linear predictor is apparent here. In fact, recalling equations (2.5a, 2.5b), while the global intercept μ_j is the *"coarser"* effect on the $j-$th category, main effects of the covariates are population-level parameters that model associations between covariates and *taxa*. Finally interactions among covariates represent the finer level, and are modeled with subject-level effects. A thresholding function is employed to ensure truly sparse solutions; finally, a random intercept accounts for the correlation among counts obtained from the same individual.

2.4 Posterior Computation

We implement a Markov Chain Monte Carlo (MCMC) algorithm to obtain the posterior distribution of the parameters of interest. To construct an efficient algorithm and improve computational feasibility we adopt the data augmentation approach implemented in Wadsworth et al. (2017) and detailed in Koslovsky et al. (2020); this approach is based on the representation of the Dirichlet distribution via independent latent Gamma random variables, and greatly facilitates the sampling procedure. In the following we outline the implemented Metropolis-Hastings within Gibbs algorithm:

1. **Jointly update** $\{(\theta_{pj}, \xi_{pj})\}$ **and** $\{(\theta_{qj}, \xi_{qj})\}$ with the two-step scheme proposed in Savitsky et al. (2011) by Metropolis-Hastings with adaptive proposal;
2. **Update** $\{b_{pqj}\}$ **and** $\{b_{qlj}\}$ by Metropolis-Hastings with adaptive proposal; hyperparameters are updated through slice sampler (Neal, 2003), as suggested in Polson et al. (2014);
3. **Update** $\{\alpha^0_{pkj}\}$ **and** $\{\boldsymbol{\alpha}^\star_{pkj}\}$ by Metropolis with random walk proposal. Accordingly to Scheipl et al. (2012) and Ni et al. (2019a) we updated the parameters involved in the peNMIG hierarchical structure by Gibbs;
4. **Update** t_x, t_z by Metropolis with random walk proposal;
5. **Update** $\{\mu_j\}$ by Metropolis with random walk proposal;
6. **Update** $\{\iota_{s(i)}\}$ by Metropolis with random walk proposal.

A detailed description of this algorithm can be found in Appendix A.4. The algorithm is initialized at a random point in the parameter space and then it is repeatedly used to generate draws from the posterior distribution. After burn-in and thinning, inference is carried out on the remaining samples. Model selection for main effects and for varying coefficients of the continuous covariates can be based on the marginal posterior probability of inclusion (MPPI); a MPPI can be computed by taking the average of sampled values of the corresponding inclusion indicator. We evaluate the inclusion of interaction terms conditional on the inclusion of corresponding main effects (see also Stingo et al., 2011). According to the strong or weak heredity

assumption, an interaction term is not included if one or both the correspondent main effects are not in the model. For coefficients representing interactions of discrete covariates, MPPI cannot be computed, since their prior is an absolutely continuous mixture distribution. We follow the procedure proposed by van der Pas et al. (2017) based on credible sets, and include in the model the interaction parameters $\{b_{pqj}\}$ and $\{b_{qlj}\}$ if their marginal credible interval does not include zero. For example, under the strong heredity assumption, the parameter $\{b_{qlj}\}$ is selected if:

$$0 \notin \left[Pr\left(b_{qlj} \leq q_1 \mid \{\boldsymbol{z}_q\}, \theta_{qj}, \theta_{lj}, t_z, \mathbb{1}_{[\xi_{qj}\xi_{lj}=1]} \right), Pr\left(b_{qlj} \leq q_3 \mid \{\boldsymbol{z}_q\}, \theta_{qj}, \theta_{lj}, t_z, \mathbb{1}_{[\xi_{qj}\xi_{lj}=1]} \right) \right],$$

where q_1, q_3 are two arbitrary quantiles. An analogous condition is specified for $\{b_{pqj}\}$. It should be pointed out that this procedure may result in the selection of interaction terms that do not respect the strong heredity assumption.

2.5 Simulation Studies

We carry out a comparative study on simulated data to evaluate the performances of our method on finite samples. We compare the proposed subject-specific DM approach (SSDM) with three other methods specifically developed for the analysis of microbiome data or multivariate count data, namely Dirichlet-multinomial Bayesian Variable Selection (DMBVS) (Wadsworth et al., 2017), the sparse group ℓ_1 penalized likelihood procedure for variable selection for the DM (pen-CL) (Chen and Li, 2013b) and the penalized likelihood approach by Zhang et al. (2017) (pen-Z).

2.5.1 Generating mechanism

The simulation scenarios considered in our studies closely follow the ones presented in Chen and Li (2013b). We simulate n samples with P continuous covariates, Q binary factors, and J bacterial *taxa* to emulate the motivating data set. The covariates and the factors are generated from a multivariate normal distribution with mean 0 and covariance matrix $\Sigma_{ij} = \rho^{|i-j|}$. Binary factors are generated by first drawing random samples from the same normal distribution used to generate continuous covariates, and then setting to 1 all the positive values and to 0 the negative ones. The global intercept is sampled from a Uniform distribution such that $\mu_j \sim \mathrm{U}(-2.3, 2.3)$ (Chen and Li, 2013b), so that the base *taxa* abundances can differ up to 100 folds, while the random intercept is drawn from a standard Normal distribution $\iota_{s(i)} \sim N(0, 1)$. 5% of the associations between covariates and *taxa* are selected and equally spaced over the interval $[0.25, 1.5]$ with alternate signs. Note that in our approach two-way interactions are explicitly modeled; this additional flexibility results in an increased complexity of the model. In these simulation studies, interaction effects are generated according to the *heredity assumption* made. Half of the possible two-way interactions are randomly selected and their value are sampled from $[0.25, 1.5]$. Linear and nonlinear interactions among continuous covariates are generated by emulating the structure of the varying coefficient $\beta_{pj}(\boldsymbol{x}_i, \boldsymbol{z}_i)$. Following the strategy adopted for other linear interactions, we set 25% of the interaction coefficients between continuous covariates ($\{\alpha^0_{pkj}\}$ in equation (2.6)) to non-zero values. Specifically we set $\alpha^0_{pkj} = \theta_{pj}\theta_{kj}$. We then added to the linear term $x_{ik}\alpha^0_{pkj}$ the function $\tilde{f}_{pkj}(x_{ik}) = 1.5(x^2_{ik} - 1)$,

hence the nonlinear component is not simulated from our model. We then generate counts using the DM model for a given overdispersion parameter $\theta_0 \in [0, 1]$, where $\theta_0 = 1/(1 + \gamma_+)$. The parameter θ_0 controls overdispersion in the samples: a large value will lead to severe overdispersion, while values close to zero make the generating mechanism similar to the Multinomial distribution. The linear predictor γ_{ij} can be then set accordingly to equation (2.3). For $i = 1, \dots, n$ we draw samples from a Multinomial distribution, $\boldsymbol{y}_i \sim \text{Multinomial}(N_i, \boldsymbol{\pi}^\star)$, where the row sum N_i is a sample from a discrete Uniform distribution $N_i \sim \text{DiscreteUniform}(\upsilon, 2\upsilon), \upsilon = 1000$ and $\boldsymbol{\pi}^\star = (\pi^\star_{i1}, \dots, \pi^\star_{iJ}) \sim \text{Dirichlet}(\boldsymbol{\gamma}^\star)$. The vector $\boldsymbol{\gamma}^\star = (\gamma^\star_1, \dots, \gamma^\star_J)$ is obtained from the simulated data and the fixed coefficient setting $\gamma^\star_j = \frac{\gamma_j}{\gamma_+} \frac{1-\theta_0}{\theta_0}$, for $j = 1, \dots, J$. In order to simulate the latent random threshold, subject-specific effects with an absolute value smaller than 0.5 are set to 0.

2.5.2 Hyperparameter settings

Hyperparameters a_ω, b_ω are set to induce a weakly informative Beta prior. Specifically, we set $a_\omega + b_\omega = 2$, and the prior expected mean $m = a_\omega/(a_\omega + b_\omega)$ to a small value $m = 0.025$ which corresponds to a 2.5% prior probability of inclusion of the main effect (Wadsworth et al., 2017). The parameters $\{\tau^2_j\}$ are set to 10, a large value that induces a vague prior on the regression coefficients $\{\theta_{pj}\}$ and $\{\theta_{qj}\}$ (Chipman et al., 2001). Hyperparameters of the peNMIG priors, for both linear and the nonlinear effects, are set following the guidelines provided by Scheipl et al. (2012): $(a_{\tilde{\omega}}, b_{\tilde{\omega}}) = (1, 1)$, $v_0 = 0.00025$, $(a_{\tilde{\tau}}, b_{\tilde{\tau}}) = (5, 25)$. Threshold parameters are assumed to be uniformly distributed in the unit interval. Finally, the hyperparameters for the variance of the subject-specific random intercept are set to $a_{\sigma^2_\iota} = b_{\sigma^2_\iota} = 1$, and the variance for the global intercept is set to a large value $\sigma_{\mu_j} = 10$ (Wadsworth et al., 2017). For more details on hyperparameter settings please refer to Appendix A.2.

2.5.3 Simulation scenarios and results

We start from a reference scenario that mimics our case study and evaluate the competing methods on an array of scenarios of increased complexity. Complexity is defined in terms of the number of covariates or *taxa* considered, by the interaction pattern, or by the level of overdispersion. Note that among the competing methods only the proposed SSDM includes subject-specific coefficients; in order to perform a fair comparison, methods are evaluated only based on inference on population-level parameters. Subject-specific parameters recovery is still evaluated for our model. For the same reason, no random intercept is used to generate the data in these scenarios; we study the proposed method on data with repeated measurements for each sample in Appendix A.5. The *reference scenario* (scenario a) in Table 2.1) is defined by the following setting: $n = 100$, $P = 10$, $Q = 5$, $J = 10$, $\rho = 0.4$. Moreover, overdispersion is set at a low level, $\theta_0 = 0.01$, and *strong heredity* is assumed. A list of all scenarios is given in Table 2.1. In particular, scenarios b) and c) are defined by an increased number of covariates and dimensions of the response, respectively. In scenario d), response variables are generated given the covariates observed in the case study, while in scenarios e) and f) data are generated with different interactions assumptions: in the first one no interactions among covariates are assumed, while in the latter the less stringent *weak heredity* assumption is considered. Microbiome data usually

Table 2.1: List of simulation scenarios and their characteristics. Scenario a) is regarded as *reference scenario*.

Scenario	n	P	Q	J	θ_0	Interaction Assumption
a)	100	10	5	10	0.01	strong heredity
b)	200	10	10	10	0.01	strong heredity
c)	200	5	5	20	0.01	strong heredity
d)	100	5	5	11	0.01	strong heredity
e)	100	5	5	10	0.01	no interactions
f)	100	5	5	10	0.01	weak heredity
g)	200	5	5	10	0.1	strong heredity
h)	200	10	10	10	0.1	strong heredity

exhibit both overdispersion and zero-inflation, while DM model can explicitly account only for overdispersion. In order to test our method under scenarios resembling our case of study, in scenarios g) and h) we generate counts under settings that exhibit large overdispersion ($\theta_0 = 0.1$). In these simulation scenarios, zero-inflation goes from 17.2% to 45.95% in settings with low overdispersion (a-f), and from 44.25% to 53.55% in settings with large overdispersion (g-h).

Results for the selection of subject-specific and population-level parameters are reported in Table 2.2 and Table 2.3, respectively. Comparisons are made in terms of true positive rates (TPR), false positive rate (FPR), and Matthew's Correlation Coefficient (MCC) a measure of overall selection accuracy, that takes into account true and false positives and negatives (TP, FP, TN, FN, respectively),

$$TPR = \frac{TP}{FN + TP}, \quad FPR = \frac{TN}{FP + TN},$$
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

For SSDM and DMBVS, covariates are selected if they belong to the median probability model (Barbieri et al., 2004). Reported values are averaged over 100 replicates, with standard errors in parentheses. DMBVS results in the highest TPRs in scenarios a) and c) (see Table 2.3). Both DMBVS and pen-CL tend to include more false associations, resulting in larger FPR and smaller MCC compared to SSDM. This discrepancy is

Table 2.2: Selection of subject-specific parameters: mean across 100 replicated datasets (standard errors are in parentheses). We evaluate the performances of the proposed approach SSDM in terms of TPR, FPR and MCC.

	TPR		FPR		MCC	
a)	0.6446	(0.0090)	0.1282	(0.0089)	0.3986	(0.0110)
b)	0.7795	(0.0168)	0.1206	(0.0087)	0.4020	(0.0150)
c)	0.7142	(0.0113)	0.0673	(0.0058)	0.4553	(0.0179)
d)	0.7345	(0.0430)	0.1134	(0.0093)	0.4173	(0.0290)
e)	-	-	0.0579	(0.0059)	-	-
f)	0.7332	(0.0272)	0.3330	(0.0281)	0.2908	(0.0263)
g)	0.5713	(0.0098)	0.0490	(0.0030)	0.4630	(0.0122)
h)	0.6046	(0.0298)	0.0561	(0.0036)	0.4285	(0.0247)

clearer in scenario b), where, given the larger number of covariates, the interaction terms play a bigger role. In scenario d), the covariates observed in the case study are used to generate the response variables; the proposed method SSDM, which has been designed explicitly for these data structures, outperforms all other methods in terms of MCC. In scenarios e-f) performance are evaluated for contexts that show different assumption on the generating mechanism of the interactions' pattern. While our model outperforms other methods in case of no interactions, it fails to account for the weak heredity assumption, despite being the only method that is not agnostic with respect to heredity assumption.

In scenarios g-h) performances are evaluated on scenarios that show overdispersion levels closer to the one in CRC data. In scenario g) DMBVS has the better recovery of associations, despite the larger overdispersion. In scenario h), the larger number of covariates leads all methods to poorer performances, even though SSDM's loss is the less severe of all.

In order to evaluate both the flexibility and the robustness of our method, we ran further simulations. Firstly, following Chen and Li (2013b), we evaluated the sensitivity of the proposed approach to model misspecification by analyzing data generated from the linear growth model instead of the exponential growth model. To evaluate the robustness of our approach for an increasing number of taxa, we tested our method in scenarios where $J > n$. Even if it is not the case in our case study, oftentimes, when microbial counts are collected at a lower level of the phylogenetic tree, the number of *taxa* can exceed the number of samples. To assess our approach's ability to capture flexible relationships between outcomes and covariates, we matched it with a nonparametric kernel-based method developed for compositional data (Tsagris et al., 2021). Details on these additional three simulation studies and results are reported in Appendix A.5. Finally, we perform a sensitivity analysis with respect to the global intercept variance, the variance hyperparameters of the *spike-and-slab* prior, hyperparameters on the Beta priors for main effects, and the upper bound for the threshold parameter. Results are presented in Appendix A.3; we find little or negligible sensitivity.

2.6 Case Study

The motivating study involves patients affected by either adenocarcinoma or diverticulosis, a condition that may eventually lead to tumor development. A total of 36 adult patients were enrolled in the study, between the ages of 36 and 85 years old; approximately a quarter of enrolled patients are female. Clinical and behavioral covariates measured by protocol consist of 5 continuous covariates and 5 binary factors: frequency (weekly basis) of dietary intake for *Vegetables* and *Meat*, frequency of *Physical Activity* (weekly basis), *Body mass index (BMI)*, *Age*, *Gender* (Male category used a baseline), daily use of *Mouthwash* (binary variable), presence of *Adenocarcinoma* (binary variable). Continuous variables were standardized. For each patient, up to 3 samples were collected (for a total of 100 records), specifically tumor, fecal and salivary samples. Information regarding the district of each sample is encoded in two binary variables (*Stool* and *Saliva*), using tumoral tissue as baseline. Microbiota measurements were obtained as follows: DNA from biological samples was measured through spectrophotometer. Samples underwent total genomic DNA extraction and 16S rDNA sequencing. In particular, amplification of bacterial DNA

Table 2.3: Model selection performance: mean across 100 replicated datasets (standard errors are in parentheses). In each scenario and for each index the best performance is in bold. Evaluation is carried out on population parameters only.

a) $n = 100, P = 10, Q = 5, J = 10$							b) $n = 200, P = 10, Q = 10, J = 10$					
	TPR		FPR		MCC		TPR		FPR		MCC	
SSDM	0.5308	(0.0125)	0.0123	(0.0006)	**0.5833**	(0.0124)	0.5639	(0.0235)	0.0081	(0.0006)	**0.4227**	(0.0170)
DMBVS	**0.6186**	(0.0129)	0.0295	(0.0008)	0.5351	(0.0103)	0.5745	(0.0388)	0.1011	(0.0200)	0.1965	(0.0287)
reg-Z	0.1363	(0.0064)	**0.0049**	(0.0003)	0.3304	(0.0104)	0.1442	(0.0095)	**0.0012**	(0.0002)	0.2755	(0.0168)
pen-CL	0.5104	(0.0309)	0.1535	(0.0125)	0.2476	(0.0076)	**0.6808**	(0.0387)	0.1267	(0.0139)	0.2045	(0.0113)
c) $n = 200, P = 5, Q = 5, J = 20$							d) $n = 100, P = 5, Q = 5, J = 11$					
	TPR		FPR		MCC		TPR		FPR		MCC	
SSDM	0.6990	(0.0145)	0.0113	(0.0009)	**0.5038**	(0.0170)	0.6551	(0.0172)	0.0056	(0.0020)	**0.6150**	(0.0388)
DMBVS	**0.7696**	(0.0148)	0.0222	(0.0013)	0.4612	(0.0123)	**0.7017**	(0.0480)	0.1132	(0.0429)	0.3804	(0.0622)
reg-Z	0.2194	(0.0090)	**0.0014**	(0.0001)	0.3740	(0.0135)	0.2915	(0.0262)	**0.0055**	(0.0011)	0.4038	(0.0278)
pen-CL	0.5337	(0.0464)	0.1127	(0.0147)	0.1446	(0.0094)	0.4606	(0.0523)	0.0683	(0.0143)	0.2349	(0.0119)
e) $n = 100, P = 10, Q = 5, J = 10$, no interactions							f) $n = 100, P = 5, Q = 5, J = 10$, weak heredity					
	TPR		FPR		MCC		TPR		FPR		MCC	
SSDM	**1.0000**	(0.0000)	0.0689	(0.0003)	**0.7800**	(0.0079)	0.5888	(0.0170)	0.0478	(0.0068)	0.4094	(0.0248)
DMBVS	0.8962	(0.0202)	0.0328	(0.0063)	0.7115	(0.0330)	**0.7668**	(0.0414)	0.1023	(0.0340)	0.4425	(0.0485)
reg-Z	0.4111	(0.0095)	**0.0018**	(0.0002)	0.5669	(0.0105)	0.3660	(0.0329)	**0.0094**	(0.0018)	**0.4533**	(0.0356)
pen-CL	0.8388	(0.0192)	0.1081	(0.0074)	0.5850	(0.0070)	0.5540	(0.0668)	0.0858	(0.0105)	0.2811	(0.0281)
g) $n = 200, P = 5, Q = 5, J = 10, \theta_0 = 0.1$							h) $n = 200, P = 10, Q = 10, J = 10, \theta_0 = 0.1$					
	TPR		FPR		MCC		TPR		FPR		MCC	
SSDM	0.5294	(0.0147)	0.0043	(0.0003)	0.5716	(0.0157)	0.4713	(0.0329)	0.0030	(0.0005)	**0.4579**	(0.0420)
DMBVS	**0.6652**	(0.0141)	0.0085	(0.0004)	**0.5900**	(0.0116)	**0.5596**	(0.0462)	0.0319	(0.0250)	0.4220	(0.0572)
reg-Z	0.1491	(0.0075)	**0.0012**	(0.0001)	0.3079	(0.0119)	0.0904	(0.0108)	**0.0009**	(0.0003)	0.2162	(0.0181)
pen-CL	0.4601	(0.0217)	0.0916	(0.0060)	0.1550	(0.0060)	0.4604	(0.0535)	0.0565	(0.0078)	0.1515	(0.0183)

was performed by PCR, targeting V3 region, encoding 16S rDNA. This region is particularly suited for microbiota assessment because it is phylogenetically well conserved, and allows the investigators to clearly identify bacteria present in the biological samples. All measurements analyzed were collected at baseline. We focus on estimating associations between *phylum*-level counts and covariates. Among the 28 *phyla*, there were 2 with missing names and 16 whose count was nonzero in less than 3% of the samples. To preserve compositionality the latter 18 *phyla* were aggregated in a *Residual* category. These *phyla* have up to 83% of zeros (four have more than 50% of zeros). Overdispersion is estimated to be large: $\theta_0 = 0.11$. A detailed discussion on overdispersion and zero-inflation in the case study data is reported in Appendix A.6.1.

The goal of our analysis is to identify clinical and dietary covariates that influence microbiota composition in the three districts of interest.

2.6.1 Inferring associations between taxonomic abundances and covariates

We ran the MCMC algorithm for 300,000 iterations, with a burn-in period of 150,000 iterations; chains were thinned, and we kept every 10−th sampled value. Our code takes nearly 31 min to run on an Intel Core i7-9750 2.60 GHz processor. Hyperparameters are set following the same specifications detailed in Section 2.5.2. Nonetheless, the large overdispersion and the weak signal in the data required a careful setting of τ_j, a_ω, and b_ω. Specifically, we set $\tau_j = 1, a_\omega = 0.25$ and $b_\omega = 1.75$; this setting allows the model to select main effects of smaller absolute values. Comparisons with other models are not performed since we could not obtain a stable and coherent selection over several replications. It should be noted that our approach has been specifically designed for the analysis of these data. Presumably, other methods fail because they do not take into account the hierarchical structure of the data and cannot deal with the interactions without any heredity assumption.

The posterior mean of main effects of covariates on microbiota relative abundances and their MPPI are reported in Table A.11; Table A.12 reports interaction effects in *Firmicutes* and *Bacteroidetes*. It should be made clear that posterior means are to be interpreted on the log scale. The exponentiation of a given posterior mean can be interpreted as the (average) change of the proportion of a *taxon* that corresponds to a unit increase of a continuous covariate or to a change in category with respect to baseline for discrete covariates. The MPPI of a coefficient can be interpreted as the degree of support provided by the data to the effect of a covariate on a given *taxon*.

An increase in meat consumption is associated with a larger *Bacteroidetes* and *Firmicutes* relative abundance. In fact, the posterior mean of these associations is 2.25 and 1.41, respectively, and these effects are strongly supported by the data (MPPI of 0.86 and 0.66). For example, a unit increase - that means eating meat all days instead of never- leads to a 2.25 times larger relative abundance of *Bacteroidetes*. Similarly, higher vegetable intake leads to a larger relative abundance of the *Bacteroidetes* phylum (posterior mean is 1.62 and MPPI= 0.69). Also *Gender*, *Adenocarcinoma*, *Stool* and *Saliva* show associations supported by the data. Several interaction effects among these covariates were also identified. *Stool* specimens show a larger relative abundance of *Bacteroidetes* than tumoral tissue specimens. This effect is even stronger for those patients affected by adenocarcinoma: in fact, the interaction term

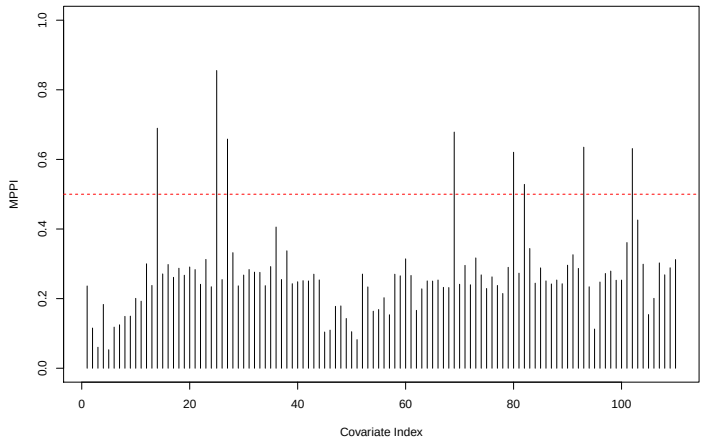
between *Stool* and *Adenocarcinoma* in *Bacteroidetes* has a posterior mean of 2.49 (with a credible interval that ranges from 1.63 to 3.04). The effect of *Gender* shows that for female patients there is an increase in the relative abundance of *Bacteroidetes* with respect to male patients: the posterior mean value for this association is 0.68 (MPPI= 0.68). Even if both *Gender* and *Stool* are included as main effects in the *a-posteriori* model, they do not show any synergistic or antagonistic effect, with respect to the *Bacteroidetes taxon*. They have a strong synergetic effect on *Firmicutes* relative abundance instead, with respect to whom only *Stool* shows a significant effect. This means that *Firmicutes*' relative abundance is larger in stool samples than in tumoral tissues and this effect happens with an increased magnitude for female patients. Another example of *weak heredity* in the *a-posteriori* model is the interactions with *Adenocarcinoma* reported in the *Bacteroidetes taxon*. In fact, patients with adenocarcinoma have a positive effect on the relative abundance in *Bacteroidetes* (posterior mean of 1.26). Even if both the effects of *Saliva* and *Mouthwash* are not included in the model as main effects with respect to the *Bacteroidetes taxon*, they exhibit significant interactions with *Adenocarcinoma*. Finally, the interaction term between *Mouthwash* and *Saliva*, which exhibits a strong magnitude (2.34, CI 1.96, 2.56), proves the peculiar flexibility of our approach. In fact, it is able to catch the varying effect of *Mouthwash* across groups, which results to be significant only in salivary samples. This implies that *Firmicutes* have a larger relative abundance in salivary samples rather than in tumor tissues and this effect is enhanced in patients that use mouthwash. Interactions found to be significant are among binary factors. In this case inference on personalized features leads to the characterization of subgroups, determined by the combination of binary variables. Interaction patterns arising in the case study are reported and discussed in Appendix A.6.2.

2.6.2 Biological findings

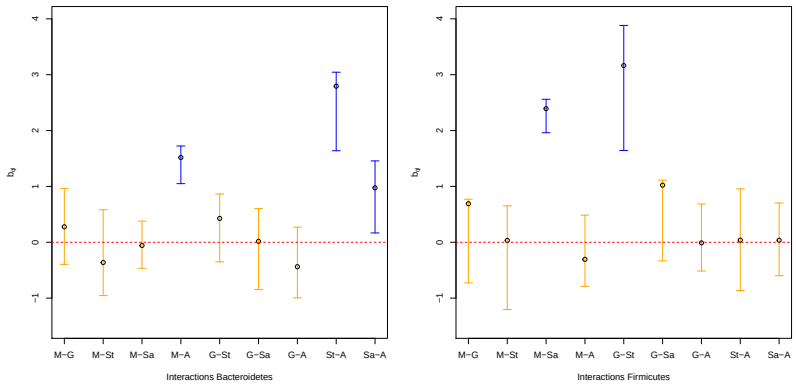
Investigating the biological aspect of the associations selected by our model, we found some relationships to be relevant and interesting for the microbiota-CRC dynamics. Associations between meat and vegetable consumption and *Bacteroidetes* and *Firmicutes* are supported and confirmed by several studies. *Bacteroides* (one of the most representative *genera* of the *Bacteroidetes phylum*) and *Firmicutes* relative abundance increase is led by protein-based dietary styles (Hentges et al., 1977; Zhu et al., 2015, 2016).

The association with microbiota's source has been reported in the literature, too. Relative abundance of *Bacteroidetes* and *Firmicutes* results to be higher in stool samples (Jenkins et al., 2018). Moreover, the significant association between *Saliva* and *Firmicutes* confirms that the oral cave shows a higher concentration of this *phylum* with respect to tissue (Xun et al., 2018). In addition, the interaction between *Saliva* and *Mouthwash* indicates that the use of mouthwash significantly increases the abundance of *Firmicutes* (Bescos et al., 2020).

We observed an interaction between adenocarcinoma and stool samples, leading to a higher concentration of *Bacteroidetes*. According to Jahani-Sherafat et al. (2018), more prevalent gut microbiota variations in the fecal and tissue samples of CRC patients were registered in *Fusobacterium*, *Porphyromonas*, *Bacteroidetes* and *Prevotella*, showing an increased *Bacteroides* percentual in fecal sample. Even if it's confirmed that this *phylum* is present both in CRC patients and in diverticulosis



(a) Plot of marginal probabilities of inclusion for main effects on microbiota *taxa*. Red dashed line represents the median model threshold (0.50).



(b) 90% marginal credible intervals for interaction among discrete covariates. Circles represent the median value. Covariates are abbreviated: "M" is Mouthwash, "G" is Gender, "St" is Stool, "Sa" is Saliva and "A" is Adenocarcinoma.

Figure 2.1: Marginal probabilities of inclusion for main effects and 90% marginal credible intervals for interaction among discrete covariates.

patients, this effect is still ambiguous. Finally, the relative abundance of *Bacteroidetes* results higher in female patients with respect to male, *ceteris paribus*. Haro et al. (2016) observed that the abundance of *Bacteroidetes* was lower in men than in women, showing interaction with BMI in male patients, meanwhile for women it remained unchanged regardless of BMI.

2.7 Discussion

We have proposed a subject-specific Bayesian Dirichlet-multinomial regression model that identifies complex association patterns in microbiota data analysis. Our approach based on varying coefficients builds upon the work of Ni et al. (2019a); in this project, we allow covariates to have a double role, as both predictors and effect modifiers. Consequently, varying coefficients result in two-way interactions that may be specific

for each patient, rather than quantified and estimated at a population level. This is consistent with the pursuit of individual characterization of the disease proper to the precision-medicine paradigm (Kosorok and Laber, 2019). Motivated by a microbiota CRC study, the proposed model is designed to analyze data that exhibit a hierarchical structure induced by measurements from multiple districts; overall, the proposed approach captures patients' heterogeneity and similarities in terms of effects affecting microbiota composition. Our analysis of the CRC data reveals interesting links between specific *phyla* and available covariates, which are confirmed by the existing literature.

Extensions of our model are possible. Microbiota is known to change over time (Faith et al., 2013); accounting for longitudinal measurements would offer precious insights into the dynamics of the microbiota abundance with respect to CRC progression. Repeated measurements over time would induce a correlation structure within each subject. Correlation among repeated measurements could be easily accounted for, as we have done for multi-district microbiota counts. Moreover, in future research, we will explore model formulations alternative to the Dirichlet-multinomial distribution. In fact, the typical DM formulation does not naturally capture all kinds of dependencies; covariances in the DM model are inherently negative, and this model may lead to poor performances when multivariate count data exhibit positive correlations. Alternative modeling approaches may be better suited for the analysis of microbiota data that exhibit positive correlations.

Chapter 3

Bayesian Nonparametric Predictive Model for Personalized Treatment Selection in Cancer Genomics

3.1 Introduction

Cancer is a complex and dynamic process characterized by heterogeneous cellular mutations, both in patients' genomes and among cancer cells within the same tumor (Bedard et al., 2013). Patients with close clinical cancer phenotypes may show diverse responses to treatment, reflecting their inherent heterogeneity. A therapeutic strategy for a particular diagnosis may be effective on average, but its effectiveness may vary across subpopulations. For instance, Trastuzumab -a monoclonal antibody used to treat breast cancer- is a very effective agent for HER2-positive tumors only (Slamon et al., 2001). Moreover, clinical benefits from Trastuzumab increase for patients with significant HER2 overexpression (ICH3+) (Slamon et al., 2001; Marty et al., 2005). Precision medicine's mission is to tailor treatment to individual patient characteristics leveraging various sources of heterogeneity. It ultimately aims to exploit current understandings of the biological mechanisms of diseases to devise therapeutic strategies best suited to the patients' genetic and clinical features (Simon, 2010). There is an increasing interest in discovering individualized treatment rules (ITRs) for patients who have heterogeneous responses to treatment, e.g., when the treatment effect varies across groups of patients. An ITR is a decision rule that assigns the patient to the treatment given patient/disease characteristics (Ma et al., 2015). The optimal ITR is the one that maximizes the population mean outcome. Statistical methodology research in precision medicine is devoted to developing personalized treatment rules to inform decision-making. The distinctive mark of statistical inference under the precision medicine paradigm is to disregard heterogeneity as a nuisance to inference, but rather, to take advantage of it to improve therapeutic strategies (Kosorok and Laber, 2019). In this paper, we restrict our focus on treatment selection at a single decision point. Conventional methods are based on semi- and non-parametric procedures to identify subgroups of patients more likely to benefit from a treatment leveraging few baseline markers (Bonetti and Gelber, 2000; Song and Pepe, 2004). When planned in prespecified analysis, the

subgroup approach can provide valuable information. Nonetheless, defining groups with respect to a few markers may result in an inadequate stratification. Moreover, it is customary to conduct multiple subgroups analysis, either in post hoc studies or obtaining subgroups as all the possible combinations of baseline characteristics and endpoints. In multiple subgroups analysis, the probability of false-positive findings can be substantial (Lagakos et al., 2006). A more systematic approach to derive ITR accommodating patients heterogeneity is through covariate adjustment. Kang et al. (2014) proposed a method for treatment selection that requires modeling the treatment effect as a function of markers. As an alternative, covariate adjustment can be performed through parametric or non-parametric methods to construct weighted estimators of the treatment rule (Zhang et al., 2012; Zhao et al., 2012). Another common strategy for choosing the optimal treatment is the identification of predictive biomarkers, which are features that determine the extent of benefit offered by a particular therapeutic strategy (the HER2 gene). Predictive associations can be identified by conducting inference on gene-wise generalized linear models including interaction terms between gene expressions and targeted treatment (Werft et al., 2012). Nonetheless, the high-dimensional nature of genomic features is an obstacle to predictive biomarkers' identification. Penalized estimation proves to be a viable solution (Krämer et al., 2009; Breheny and Huang, 2011; Lu et al., 2013), as it provides a set of nonzero coefficients for treatment-biomarkers interactions, selecting only those relevant for the optimal treatment. Covariate adjustment and penalized approaches suffer from some limitations shared with subgroups analysis. Indeed, it is of paramount importance to adjust for the appropriate covariates. When multiple features are available, *post hoc* selection of covariates for adjustment leads to biased estimates of the treatment effect, especially if the sample size is moderate (Pocock et al., 2002). On the other hand, selecting predictive associations deals with thousands of potentially predictive biomarkers simultaneously, which makes the control of false discovery rate essential in multiple testing scenarios (Werft et al., 2012). Moreover, for these methods, the correct definition of treatment-by-markers interactions is crucial and relies on sensitive assumptions, which are difficult to specify in the clinical practice and may be limited by generalized linear models (Ma et al., 2016).

In order to overcome these limitations, Ma et al. (2016, 2018, 2019) have established a predictive model for personalized treatment utility based on a heuristic measure of interpatient molecular similarity, obtained using an unsupervised clustering approach. Given a genomic signature and a set of prognostic markers, they constructed a predictive framework that integrates predictive and prognostic determinants. For a new, untreated patient, the model provides a probabilistic basis to predict personalized treatment utility offered by each competing treatment. This framework establishes two significant improvements over existing methods. Firstly, the common assumption of statistical exchangeability among patients is relaxed. Since each tumor is unique, patients are considered partially exchangeable only to the extent to which their tumors are molecularly similar. Moreover, this approach utilizes complementary sources of information for treatment selection, integrating both predictive and prognostic characteristics of a patient's disease.

In this chapter, we propose a Bayesian predictive model for personalized treatment selection. This method does not address biomarker discovery, but rather, it assumes that a genomic signature and a set of prognostic markers are available, known

from previous studies or earlier experiments. As in Ma et al. (2019), we leverage prognostic determinants to measure how likely is a patient to reach a given clinical response. Predictive biomarkers drive patients' clustering within each one of the competing treatments. This approach characterizes the extent of benefit offered by each therapeutic strategy on groups of patients with close profiles in predictive determinants.

The proposed method generalizes the modeling approach by Ma et al. (2019), constructing an integrative framework for clustering and prediction rather than a two-step procedure. We jointly estimate model-based clustering and treatment assignment from the data, making treatment selection fully account for patients heterogeneity.

In particular, we adopted the product partition model with covariates (PPMx) (Müller et al., 2011) to induce clusters of observations that are more homogeneous in terms of predictive covariates. The class of PPMx models incorporates covariates information into the prior for the random partition. The resulting partitions are only partially exchangeable, and patients with similar covariates are *a priori* more likely to be clustered together. In this Chapter, we use the Normalized Generalized Gamma (NGG) process as the *cohesion function* of a PPMx model. This process overcomes the *rich-get-richer* property of the Dirichlet process. Despite being well studied in the Bayesian nonparametric literature as a prior inducing a Gibbs-type random partition (Lijoi et al., 2007; De Blasi et al., 2013; Favaro et al., 2013) and Argiento et al. (2010) proving the feasibility of Normalized Generalized Gamma mixture model in addressing real problems, it is still not a commonly used process. To the best of our knowledge, it is the first time the NGG process is employed as the *cohesion function* in a PPMx model.

This chapter is organized as follows. Section 3.2 contains the model adopted to integrate prognostic and predictive biomarkers. We devise the covariate-dependent distribution for the random partition in Section 3.3. Section 3.4 introduces prior specification; we detail the posterior computation steps in Section 3.5. In Section 3.6 we elaborate a strategy for treatment selection that is evaluated with a simulation study in Section 3.7. In Section 3.8 we use our method for the analysis of publicly available data on brain cancer and we present some preliminary results. A brief discussion and some prospectives for future works conclude the Chapter in Section 3.9.

3.2 Bayesian Integrative Model

Our approach is motivated by an open problem in cancer genomics and personalized medicine. Given a sample of n patients assigned to T different treatments for whom predictive and prognostic biomarkers are measured and given a discrete set of ordered response levels of the clinical outcome, we build a model able to leverage complementary sources of information.

In this section, we describe the model devised to integrate prognostic and predictive biomarkers.

Let $a = 1, \dots, T$ indices candidate therapies to which $n = \sum_{a=1}^{T} n^a$ patients are assigned, where n^a denotes the number of patients treated with therapy a. For each patient, the response to treatment is evaluated with an ordinal valued assessment. A common choice to characterize varying levels of treatment response is to evaluate it

in terms of the extent of residual disease after a given clinically relevant post-therapy follow-up duration. Let y_i^a be the random variable of the $i-$th patient's response to treatment a among K possible levels of increasing treatment benefit, where $y_i^a = k$ for $i = 1, \dots, n^a$ and $k = 1, \dots, K$. In addition, let $\boldsymbol{\pi}_i^a = (\pi_{i1}^a, \dots, \pi_{iK}^a)$ denote the vector such that π_{ik}^a is the probability of observing outcome k for the $i-$th patient under treatment a, that is

$$P(y_i^a = k \mid \boldsymbol{\pi}_i^a) = \pi_{ik}^a,$$

hence we assume that $y_i^a \mid \boldsymbol{\pi}_i^a \overset{\text{ind}}{\sim} \text{Multinomial}(1, \boldsymbol{\pi}_i^a)$.

For $a = 1, \dots, T$, we consider a training dataset of n^a patients, $(y_i^a, \boldsymbol{z}_i^a, \boldsymbol{x}_i^a)$ where $\boldsymbol{z}_i^a$ and $\boldsymbol{x}_i^a$ are a $P-$dimensional and $Q-$dimensional vector of prognostic and predictive features, respectively.

To quantify the effectiveness of each competing therapeutic strategy for patients with close genetic profiles, we adopted a random partition model depending on predictive markers. We denote with $\Pi_{n^a}^a = \{S_1^a, \dots, S_{C_{n^a}^a}^a\}$ the treatment-specific partition of the indices $\{1, \dots, n^a\}$, where $C_{n^a}^a$ is the number of clusters among patients treated with therapy a and $n_j^a = |S_j^a|$ is the cardinality of cluster j, for $j = 1, \dots, C_{n^a}^a$. Since we assume the partition of the units to be a random parameter, the partition and the number of clusters ($\Pi_{n^a}^a$ and $C_{n^a}^a$) depend on the number of observations, n^a. Following a common convention in the Bayesian nonparametric community, we identify cluster-specific quantities using the superscript "$\star$". For example, when considering the $j-$th cluster for treatment a, the response vector is $\boldsymbol{y}_j^{a\star} = \{y_i^a : i \in S_j^a\}$ while $\boldsymbol{x}_j^{a\star} = \{\boldsymbol{x}_i^a : i \in S_j^a\}$ is the partitioned covariate matrix. To maintain the presentation of the model self-contained, we will not give any detail regarding $\Pi_{n^a}^a$ here. In Section 3.3, we will present the random partition model and how we make it covariate-dependent.

We complete the Multinomial model for treatment response using a conjugate prior for $\boldsymbol{\pi}_i^a$, in particular for $a = 1, \dots, T$ we assume the following hierarchical model:

$$\begin{aligned} y_i^a | \boldsymbol{\pi}_i^a &\overset{\text{ind}}{\sim} \text{Multinomial}(1, \boldsymbol{\pi}_i^a) \\ \boldsymbol{\pi}_1^a, \dots, \boldsymbol{\pi}_{n_a}^a \mid \boldsymbol{\eta}_1^{a\star}, \dots, \boldsymbol{\eta}_{C_{n^a}^a}^{a\star}, \Pi_{n^a}^a, \boldsymbol{\beta} &\sim \prod_{j=1}^{C_{n^a}^a} \prod_{i \in S_j^a} \text{Dirichlet}(\boldsymbol{\pi}_i^a; \boldsymbol{\gamma}_i^a(\boldsymbol{\eta}_j^{a\star}, \boldsymbol{\beta})), \end{aligned} \tag{3.1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$ is a $P \times K$ matrix of regression parameters shared across levels of response and individuals.

The K-dimensional vectors $\boldsymbol{\eta}_1^{a\star}, \dots, \boldsymbol{\eta}_{C_{n^a}^a}^{a\star}$ are cluster-specific parameters, that is, $\boldsymbol{\eta}_j^{a\star}$ is a parameter shared by all the individuals in cluster S_j^a. We want to allow the response probabilities to change from treatment to treatment even for subjects with similar predictive markers. We then enforce $\eta_{jk}^{a\star}$ to be specific for each treatment a. Potentially, we could have clustered $\eta_{jk}^{a\star}$ across treatments, but this independence assumption avoids the model to induce a partition that implies the same response probability for subjects with close predictive determinants that have received different treatments. The joint law of $(\Pi_{n^a}^a, \boldsymbol{\eta}_j^{a\star})$ is assigned hierarchically. We will show how predictive markers drive the clustering process in Section 3.3 and detail their prior distributions in Section 3.4.

Finally, $\boldsymbol{\gamma}_i^a(\boldsymbol{\eta}_j^{a\star}, \boldsymbol{\beta}) = (\gamma_{i1}^a(\eta_{j1}^{a\star}, \boldsymbol{\beta}_1), \ldots, \gamma_{iK}^a(\eta_{jK}^{a\star}, \boldsymbol{\beta}_K))$, is a vector of log-linear functions on the prognostic markers and cluster-specific parameters defined as follows:

$$\log(\gamma_{ik}^a(\eta_{jk}^{a\star}, \boldsymbol{\beta}_k)) = \eta_{jk}^{a\star} + \beta_{1k} z_{i1}^a + \cdots + \beta_{Pk} z_{iP}^a. \tag{3.2}$$

It is apparent from the structure of the linear predictor the strategy we have adopted to integrate multiple sources of information. Predictive determinants enter equation (3.2) only through the cluster (and treatment) specific parameters $\eta_{jk}^{a\star}$. This construction results in a random intercept that accounts for the heterogeneity among patients arising from their genetic profiles. Prognostic markers enter (3.2) as linear predictors. The associated coefficients $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ are defined across treatments since the prognostic determinants impact the likelihood of achieving a given therapeutic response regardless of the treatment.

3.3 Bayesian Nonparametric Covariate Driven Clustering

The joint evaluation of prognostic and predictive covariates guides the optimal treatment selection. Since predictive markers identify patients likely to benefit from a particular therapy, they drive patients' clustering within each treatment. In this way, we may quantify the extent of benefit offered by a specific therapeutic strategy on groups of patients characterized by close profiles in predictive determinants. The Bayesian framework naturally handles model-based clustering assuming as a random parameter of the model the partition of the sample subjects. In this section, we present a prior distribution for the random partition that is informed by covariates (predictive markers). In Section 3.3.1 we present the product partition model (Hartigan, 1990), which is extended to a more general class in Section 3.3.2. Following Müller et al. (2011), in Section 3.3.3, we make the distribution depend on covariates. The resulting product partition distribution with covariates represents a prior distribution for the random partition with the specific feature that patients with close covariates are encouraged to co-cluster.

3.3.1 Product partition distribution

In this section, we build our prior for the random partitions $\Pi_{n^a}^a$. Since these parameters are assumed independent across therapies, for the sake of notation, we will suppress the a superscript. Moreover, we will first consider a prior free from covariate information, then discuss how to modify the prior to include this information.

We consider the partition of the sample subjects as a random parameter of the model (Hartigan, 1990). Let us denote with $\Pi_n := \{S_1, \ldots, S_{C_n}\}$ the partition of the data label set $\{1, \ldots, n\}$ into C_n subsets S_j, for $j = 1, \ldots, C_n$ and with $n_j = |S_j|$ being the cardinality of cluster j. The product partition distribution is a probability mass function for random partitions that features the following product structure:

$$p(\Pi_n) \propto \prod_{j=1}^{C_n} \rho(S_j). \tag{3.3}$$

The function ρ is referred to as cohesion (Hartigan, 1990). It measures the compactness of the elements in S_j and describes how likely the elements of S_j are thought to be grouped together *a priori*. If $\rho(S_j)$ is only a function of $n_j = |S_j|$, then the resulting model for Π_n is invariant under permutations of the labels of the set of integers $\{1, \dots, n\}$. Under this assumption, the resulting model for Π_n falls in the class of exchangeable random partitions induced by Species Sampling models (see Pitman (1996)). In this framework, $p(\Pi_n)$ is denoted as *exchangeable partition probability function* (eppf). The connection between product partition models and exchangeable random partitions has been deeply investigated since the seminal paper by Quintana and Iglesias (2003). In the latter paper, the authors observed as the cohesion $\rho(S_j) = \kappa(n_j - 1)!$, for $\kappa > 0$, introduced by Hartigan (1990) yields to product partition distribution coinciding with the eppf induced by a Dirichlet Process (DP).

Despite being computationally very convenient, this *cohesion* function features the "rich-gets-richer" property of the DP. Indeed, this means that the resulting product partition distribution assigns a high probability to a small number of large clusters. As a consequence, when new data are considered, they are more likely to join already large clusters. To overcome this issue, exploiting the connection between product partition and species sampling models, we propose here a generalization of (3.3), adopting the *cohesion* function induced by the wider class of Normalized Generalized Gamma process.

3.3.2 NGG-induced cohesion

In this paper we consider a product partition model with

$$\rho(n_j) = (1 - \sigma_j)_{n_j - 1}, \tag{3.4}$$

where $(1 - \sigma)_{n_j - 1}$ are rising factorials, defined as $(a)_n = a(a + 1) \dots (a + n - 1)$, with $(a)_0 = 1$. It is possible to show that the resulting probability mass function coincides with the eppf induced by the Normalized Generalized Gamma process Brix (1999). In Lijoi et al. (2007), the Normalized Generalized Gamma (NGG) process has been adopted to overcome the rich-get-richer issue in the context of Bayesian nonparametric mixture models. Moreover, several methodological properties of the NGG are presented. In particular, the authors show that the NGG induces a random partition among the observations whose probability mass function (e. g. the eppf) has the following analytical expression

$$p(\Pi_n) = V_{n,C_n} \prod_{j=1}^{C_n} \rho(n_j) = V_{n,C_n} \prod_{j=1}^{C_n} (1 - \sigma)_{n_j - 1}, \ \sigma \in (0, 1], \tag{3.5}$$

where the normalizing constant has the following integral representation:

$$V_{n,C_n} = \frac{\omega^{C_n}}{\Gamma(n)} \int_0^\infty u^{n-1} \exp\left\{ -(\omega/\sigma)[(\kappa + u)^\sigma - \kappa^\sigma] \right\} (\kappa + u)^{-n + \sigma C_n} \mathrm{d}u. \tag{3.6}$$

The law of an NGG process is assigned by the parameters $(\kappa, \sigma, \omega, \mu_0)$, where μ_0 is referred to as base distribution and is a nonatomic probability measure, while $0 \leq \sigma \leq 1, \omega, n \geq 0$. We mention that this parameterization is not unique, since the

same distribution can be obtained from (κ, σ, ω) and $(s^\sigma \kappa, \sigma, \omega/s)$ for any $s > 0$, due to the scaling property (Pitman, 2003). Following Argiento et al. (2010), we take $\omega = 1$, hence the NGG will depend only on (κ, σ). The NGGP includes as special cases relevant nonparametric priors. For $\omega = 1$ and $\sigma \to 0$, the DP is recovered. Moreover, if $\omega = 0$ the normalized stable process (Kingman, 1975) is obtained. Finally, for $\sigma = \frac{1}{2}$ the NGGP is denoted by normalized inverse Gaussian process (Lijoi et al., 2005; Argiento et al., 2009).

Predictive distribution of a sample from NGG process

The knowledge of the predictive distributions is useful to characterize the NGG predictive mechanism and to understand how the available information about partition associated with the sample $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ is leveraged. Denoting with $\tilde{i}$ a new observation and defining $\tilde{\boldsymbol{x}} = \boldsymbol{x}_{\tilde{i}}$, $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}_{\tilde{i}}$, from (3.3), we obtain the predictive distribution:

$$p(\tilde{\boldsymbol{\eta}} \in \cdot | \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n) = w_n^0 \mu_0(\cdot) + w_n^1 \sum_{j=1}^{C_n} (n_j - \sigma) \delta_{\boldsymbol{\eta}_j^\star}, \tag{3.7}$$

where

$$w_n^0 = \frac{V_{n+1, C_n+1}}{V_{n, C_n}}, \text{ and } w_n^1 = \frac{V_{n+1, C_n}}{V_{n, C_n}}.$$

The predictive distribution is a linear combination of the prior guess μ_0 and a weighted empirical distribution that depends on the parameter σ. Since a closed-form is available for V_{n,C_n}, explicit expressions for w_n^0 and w_n^1 can be derived (see Lijoi et al. (2007); Argiento et al. (2010)).

Equation (3.7) induces a two-step mechanism for mass allocation among a new cluster and previously observed ones. Given a sample $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$, first, the mass is allocated between a new value $\boldsymbol{\eta}_{C_n+1}^\star$ and the set of observed values $\{\boldsymbol{\eta}_1^\star, \ldots, \boldsymbol{\eta}_{C_n}^\star\}$. The second step follows conditionally: if $\boldsymbol{\eta}_{n+1}$ is a new value it is drawn from μ_0, otherwise, the probability to be assigned to previously observed clusters depends on the frequencies and σ. The role played by σ in the weighting of the empirical measure has been thoroughly studied in Lijoi et al. (2007). The growth of the number of distinct components for $NGG(\kappa, \sigma)$ for n samples is of the order n^σ, hence for $\sigma \to 1$ a larger number of distinct clusters is generated. Moreover, Lijoi et al. (2007) observed that in the second step of the predictive mechanism, σ drives a reinforcement mechanism that favors clusters with higher frequencies, reducing the number of singletons, thus counteracting the *rich-get-richer* behavior of the DP.

The predictive properties of the NGG process make it more suitable than the DP since it is possible to obtain a large number of clusters still penalizing those with lower frequencies that are not supported by empirical evidence in the data.

3.3.3 Choice of the similarity function

In this section, we allow the product partition distribution to depend on covariates. We obtain a non-exchangeable prior for the random partition that encourages two subjects to co-cluster when they have close covariates. We follow Müller et al. (2011), where the prior on the random partition is defined by perturbing the cohesion function of a product partition distribution via a similarity function g that induces the desired dependence on covariates. The *similarity* function g is a non-negative

function that depends on the covariates associated with subjects in each cluster. Let x_i denote the covariate for the $i-$th unit, while $\boldsymbol{x}_j^\star = (x_i, i \in S_j)$ represent the covariate arranged by cluster.

The product partition distribution with covariates is

$$p(\Pi_n) \propto V_{n,C_n} \prod_{j=1}^{C_n} \rho(n_j) g(\boldsymbol{x}_j^\star), \tag{3.8}$$

where $V_{n,c}$ is defined in (3.6) and $\rho(n_j)$ is of the form (3.4).

The *similarity* formalizes the homogeneity with respect to covariate values of subjects clustered together. Thus, the more the covariates are judged to be close, the larger the value of g gets. Müller et al. (2011) discuss the theoretical properties the *similarity* function should satisfy and some guidelines for its choice, but any non-negative function that guarantees an increasing value for close covariate values is suitable (Page and Quintana, 2018).

The default choice, proposed by Müller et al. (2011), is to define g as the marginal probability of an auxiliary Bayesian model:

$$g(\boldsymbol{x}_j^\star) = \int \prod_{i \in S_j} q(x_i|\boldsymbol{\xi}_j^\star) q(\boldsymbol{\xi}_j^\star) \mathrm{d}\boldsymbol{\xi}_j^\star, \tag{3.9}$$

even if x_i are not considered random. We define $g(\emptyset) = 1$. The auxiliary similarity in (3.9) is particularly convenient because it is symmetric in its argument (the auxiliary model does not depend on the units' ordering). We included the covariates in our random partition distribution relaxing the exchangeability condition and obtaining a non-exchangeable distribution for Π_n. The function $p(\Pi_n)$ still has the marginal invariance property, being an exchangeable partition probability function:

$$p(n_1, \ldots, n_{C_n}) = \sum_{j=1}^{C_n} p(\ldots, n_j + 1, \ldots) + p(n_1, \ldots, n_{C_n}, 1).$$

Moreover, the analytical evaluation of equation (3.9) is computationally really convenient. Note that when a $Q-$dimensional vector of covariates is available $g(\boldsymbol{x}_j^\star) = \prod_{q=1}^{Q} g(\boldsymbol{x}_{jq}^\star)$.

Quintana et al. (2015) propose a variation of (3.9), defining $g(\boldsymbol{x}_j^\star)$ as the posterior predictive distribution of $\boldsymbol{x}_j^\star$ in cluster S_j:

$$g(\boldsymbol{x}_j^\star) = \int \prod_{i \in S_j} q(x_i|\boldsymbol{\xi}_j^\star) q(\boldsymbol{\xi}_j^\star|\boldsymbol{x}_j^\star) \mathrm{d}\boldsymbol{\xi}_j^\star, \tag{3.10}$$

with $q(\boldsymbol{\xi}_j^\star|\boldsymbol{x}_j^\star) \propto \prod_{i \in S_j} q(x_i|\boldsymbol{\xi}_j^\star) q(\boldsymbol{\xi}_j^\star)$. Since the covariates are used twice, this function is called "Double Dipper". In our preliminary studies we found this similarity to be the most effective. The model in (3.10) is completed with $q(\cdot|\boldsymbol{\xi}_j^\star) = N(\cdot|m_j^\star, v_j^\star)$ where $N(\cdot|m, v)$ is a Gaussian density with mean m and variance v. Assuming $v_j^\star$ to be unknown, $\boldsymbol{\xi}_j^\star = (m_j^\star, v_j^\star)$ and $q(\boldsymbol{\xi}_j^\star) = q(m_j^\star, v_j^\star) = NIG(m_j^\star, v_j^\star|m_0, k_0, v_0, n_0)$ is the Normal-Inverse-Gamma density function. Under such a setting, clusters do not share the variance parameter. Following Page and Quintana (2018) the set of parameters

for the Normal-Inverse-Gamma density is $(m_0 = 0, k_0 = 1.0, v_0 = 1.0, n_0 = 2)$. Since the parameter v_0 has proven to be the most influential, we assessed the sensitivity of the results to its specification in Appendix B.1.

Approaches based on covariate-dependent random partition perform well if the clustering is not completely driven by covariates. As the number of covariates increases, similarity functions tend to overwhelm the information provided by the response, completely driving the clustering process. To counteract this behavior, Page and Quintana (2018) explore some approaches to temper the impact that covariates have on partitions. As an alternative to variable selection or to reducing the dimensionality of the covariate space through the use of sufficient statistics, Page and Quintana (2018) propose to calibrate the influence of covariates on clustering. To this end, we use $\tilde{g}(\boldsymbol{x}_j^\star) = g(\boldsymbol{x}_j^\star)^{1/\sqrt{Q}}$, a small variation of the *coarsened similarity* function by Page and Quintana (2018).

3.4 Priors

In this section, we discuss the prior distributions for $\Pi_{n^a}^a$, $\boldsymbol{\eta}_j^{a\star}$ and $\boldsymbol{\beta}_k$. Both $\Pi_{n^a}^a$ and $\boldsymbol{\eta}_j^{a\star}$ are cluster-specific quantities and their law is jointly assigned. Since the same priors are assumed independently for all treatments, we will omit the superscript a throughout the section. We complete the random partition model in (3.8) defining a prior on $\boldsymbol{\eta}_j^\star$ that induces independence across clusters and conditional independence within clusters. We include cluster-specific parameters $\boldsymbol{\zeta}_j^\star$: $p(\boldsymbol{\eta}_j^\star|\Pi_n) = \prod_{j=1}^{C_n} p_j(\boldsymbol{\eta}_j^\star)$, where $p_j(\boldsymbol{\eta}_j^\star) = \int \prod_{i\in S_j} p_j(\boldsymbol{\eta}_i|\boldsymbol{\zeta}_j^\star)\mathrm{d}p_0(\boldsymbol{\zeta}_j^\star)$ and p_0 is a prior for cluster-specific parameters $\boldsymbol{\zeta}_j^\star$. Introducing cluster membership indicators $e_i \in \{1, \ldots, C_n\}$ with $e_i = j$ if $i \in S_j$, the joint model for $(\Pi_n, \boldsymbol{\eta}_j^\star)$ can be hierarchically written as

$$\begin{aligned} \boldsymbol{\eta}_i|\boldsymbol{\zeta}^\star, e_i &\overset{\text{ind}}{\sim} p(\boldsymbol{\zeta}_{e_i}^\star) \text{ for } i = 1, \ldots, n \\ \boldsymbol{\zeta}_j^\star &\overset{\text{iid}}{\sim} p_0 \text{ for } j = 1, \ldots, C_n \\ p(\Pi_n) &= V_{n,c} \prod_{j=1}^{C_n} \rho(n_j) g(\boldsymbol{x}_j^\star) \end{aligned} \tag{3.11}$$

considering that $\boldsymbol{\zeta}_i = \boldsymbol{\zeta}_{e_i}^\star$. In particular, $\boldsymbol{\zeta}_j^\star = (\boldsymbol{\theta}_j^\star, \boldsymbol{\Sigma}_j^\star)$ and we assume that $\boldsymbol{\eta}_i|\boldsymbol{\theta}^\star, \boldsymbol{\Sigma}^\star, e_i \overset{\text{iid}}{\sim} N_K(\boldsymbol{\theta}_{e_i}^\star, \boldsymbol{\Sigma}_{e_i}^\star)$, where N_K denotes a $K-$dimensional multivariate normal density. We complete this prior specification assuming $\boldsymbol{\theta}_j^\star|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0 \overset{\text{iid}}{\sim} N_K(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ and $\boldsymbol{\Sigma}_j^\star|\nu_0, \boldsymbol{S}_0 \overset{\text{iid}}{\sim} IW(\nu_0, \boldsymbol{S}_0^{-1})$, where IW is an Inverse-Wishart distribution. In particular, $\boldsymbol{\mu}_0$ is a $K-$dimensional vector of 0, $\nu_0 = K + 2$, and $\boldsymbol{\Lambda}_0$ and $\boldsymbol{\Sigma}_0$ are two $K \times K$ diagonal matrices with elements on the diagonal being equal to 10 and 1.0, respectively. Elicitation for the latter two parameters is discussed and motivated in the sensitivity study reported in Appendix B.1.

The priors for the parameters $\boldsymbol{\beta}_k$ are also assumed to be independent and to enhance predictive performance we use shrinkage priors. We adopt the horseshoe prior (Carvalho et al., 2009, 2010), an absolutely continuous mixture distribution that belongs to the class of global-local scale mixtures of normals:

$$\begin{aligned} \beta_{pk} &\overset{\text{iid}}{\sim} N(0, \lambda_{pk}^2 \tau_k^2) \\ \lambda_{pk}, \tau_k &\overset{\text{iid}}{\sim} HC(0, 1), \end{aligned}$$

where HC denotes a half-Cauchy distribution, $\{\lambda_{pk}\}$ are local shrinkage parameters, and $\{\tau_k\}$ are global shrinkage parameters. All coefficients will be nonzero, nonetheless only those supported by the data will have large values, due to the heavy tails of the prior.

3.4.1 Induced prior distribution of the number of clusters

We argue that using the NGGP-induced cohesion rather than the one induced by the DP yields a more efficient prior mass allocation over different partitions, overcoming the *rich-get-richer* feature of the DP. This is due to the reinforcement mechanism induced by σ parameter that takes place in the predictive distribution of the NGGP (3.7). In the following example, we compare the distributions on the number of clusters C_n induced by the PPMx under different combinations of the cohesion and the similarity functions.

Example 2

We consider a mixture of three different 5−variate normal distributions such that

$$p(\boldsymbol{\eta}) = \sum_{j=1}^{3} \phi_j N_5(\boldsymbol{\theta}_j, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\phi} = (0.2, 0.5, 0.3)^\top$, $\boldsymbol{\theta}_j$ are 5−dimensional vectors such that $\boldsymbol{\theta}_1 = -2.1 \cdot \mathbf{1}$, $\boldsymbol{\theta}_2 = 0 \cdot \mathbf{1}$ and $\boldsymbol{\theta}_3 = 2.3 \cdot \mathbf{1}$, where $\mathbf{1}$ is the all-ones vector in $\mathbb{R}^5$. Finally $\boldsymbol{\Sigma}$ is 5×5 diagonal covariance matrix such that $\boldsymbol{\Sigma} = diag(0.5, 0.5, 0.5, 0.5, 0.5)$. We generate 50 values from this mixture and we use the data to compare 5 different model. We consider $DP(\kappa = 19.2333)$, $NGG(\kappa = 0.7353, \sigma = 0.7353)$ and three different PPMx to compare DP and NGG cohesions and evaluate calibrated similarities:

- PPMx with $DP(19.2333)$ cohesion and coarsened similarity,
- PPMx with $NGG(0.7353, 0.7353)$ cohesion and coarsened similarity,
- PPMx with $NGG(0.7353, 0.7353)$ cohesion and non calibrated similarity.

Note that the DP and the NGGP can be considered special cases of the PPMx distibution using $\rho(S_j) = \kappa(n_j - 1)!$ and equation (3.4) respectively as cohesion function and similarity $g \equiv 1$, that is no covariates in the prior.

Parameter elicitation for the DP and NGG process priors centers the distributions on the same expected number of clusters. This elicitation of the parameters of nonparametric priors is such that $E(C_{50}) = 25$, that is we are comparing the models under misspecification since the true number of components of the mixture is 3. The corresponding distributions are based on 10000 iterations adopting the MCMC procedure described in Section B.2.2 (considering only the steps for $(\Pi_n, \boldsymbol{\eta}^\star)$) and are displayed in Figure 3.1.

As expected, when moving from the DP to the NGGP, the distribution of C_{50} becomes flatter, exhibiting a larger variability. In fact, due to the reinforcement mechanism induced by the σ parameter, the NGGP prior gives *a-priori* support to a larger number of clusters, still penalizing the number of singleton partitions (see Table 3.1). This is particularly convenient when there is uncertainty about the true number of clusters. The PPMx models include information carried by covariates in

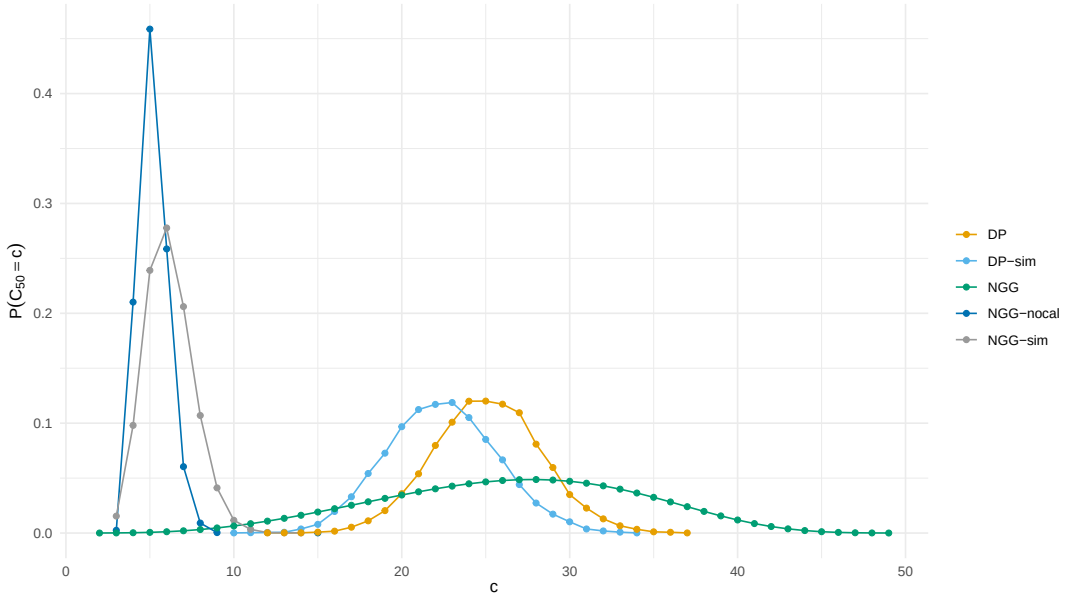


Figure 3.1: Prior distributions on the number of clusters corresponding to the Dirichlet Process (DP), PPMx with DP cohesion and coarsened similarity (DP-sim), normalized generalized gamma process (NGG), PPMx with NGG cohesion and non calibrated similarity (NGG-nocal), PPMx with NGG cohesion and coarsened similarity (NGG-sim).

the prior distribution hence should be able to counteract the misspecification in the prior elicitation. Nonetheless, when the DP cohesion is adopted the PPMx still does not differ much from the DP, exhibiting a distribution that gives support to a number of clusters that is much larger than the true one. This is due to the *rich-get-richer* phenomenon, since a large portion of clusters (53%) are singletons, as displayed in Table 3.1.

Table 3.1: Average number of clusters and proportion of singletons corresponding to the Dirichlet Process (DP), PPMx with DP cohesion and coarsened similarity (DP-sim), normalized generalized gamma process (NGG), PPMx with NGG cohesion and non calibrated similarity (NGG-nocal), PPMx with NGG cohesion and coarsened similarity (NGG-sim).

	DP	DP-sim	NGG	NGG-nocal	NGG-sim
Av. # clusters	25.09	22.38	26.25	5.19	6.13
% singletons	56%	53%	17%	6%	12%

Finally, let us focus on the PPMx models with NGGP as cohesion function. We considered both the case of calibrated and uncalibrated similarity. The implied distributions on the number of clusters give strong support to a moderate number of clusters, either way. That is that the NGGP effectively embeds information carried by the covariates, counteracting the prior misspecification. Nonetheless, when the similarity is not calibrated, the PPMx implies a highly peaked distribution of C_{50}. This is consistent with what is discussed in Page and Quintana (2018), where they draw attention to the risk of similarities completely driving the clustering process. Since this phenomenon is more pronounced for a larger number of covariates (as in our case study), following their results we judge the calibrated similarity to be better suited for our model. □

3.5 Posterior Inference

We implement a MCMC procedure to simulate a Markov Chain whose equilibrium distribution is the posterior distribution of the parameters of interest:

$$\begin{aligned} H(\boldsymbol{\eta}^{\star}, \boldsymbol{\Pi}, \boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\pi}) &= \prod_{a=1}^{T} H^a(\boldsymbol{\eta}^{a\star}, \Pi^a, \boldsymbol{\beta}; \boldsymbol{y}^a, \boldsymbol{x}^a, \boldsymbol{\pi}^a), \\ H^a(\boldsymbol{\eta}^{a\star}, \Pi^a, \boldsymbol{\beta}; \boldsymbol{y}^a, \boldsymbol{x}^a, \boldsymbol{\pi}^a) &= \prod_{i=1}^{n_a} \pi^a_{iy^a_i} \prod_{j=1}^{C^a_n} \prod_{i \in S^a_j} f_{\boldsymbol{\pi}|\boldsymbol{\gamma}}(\boldsymbol{\pi}^a_i | \boldsymbol{\gamma}^a_i(\boldsymbol{\eta}^{a\star}_j, \boldsymbol{\beta})), \end{aligned} \tag{3.12}$$

where y^a_i is the observed response for patient i who received treatment a and $f_{\boldsymbol{\pi}|\boldsymbol{\gamma}}$ denotes the Dirichlet density function. To construct an efficient algorithm and improve computational feasibility we adopt a data augmentation approach (Argiento et al., 2016) to represent the Dirichlet distribution as independent latent Gamma random variables. In particular, we reparameterize equation (3.1) letting $\pi^a_{ik} = d^a_{ik}/D^a_i$, where $D^a_i = \sum_{k=1}^{K} d^a_{ik}$ and assume that $d^a_{ik} \sim Gamma(\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k), 1)$. Refer to Appendix B.2.1 for more details. This greatly facilitates the sampling procedure. The core part of the algorithm is the update of cluster membership. The computation associated with (3.11) is based on Neal (2000)'s Algorithm 8 with Reuse (Favaro et al., 2013). Conditional on the updated cluster labels, all the remaining parameters are easily updated with Gibbs sampler or Metropolis-Hastings steps. In the following we outline the implemented Metropolis-Hastings within Gibbs algorithm:

1. **Update $\boldsymbol{\Pi}$** by Neal (2000)'s Algorithm 8 and the Reuse step proposed in Favaro et al. (2013).

2. **Update $\boldsymbol{\eta}^{\star}$** by Metropolis-Hastings. Hyperparameters are updated through Gibbs steps from their respective full conditional distributions.

3. **Update $\boldsymbol{\beta}$** by Metropolis-Hastings. Hyperparameters are updated through slice sampler (Neal, 2003), as suggested in Polson et al. (2014).

4. **Update $\boldsymbol{d}$** by Gibbs from its full conditional distributions.

Further details can be found in Appendix B.2.2.

3.6 Treatment Selection

To perform treatment selection for a new untreated patient $\tilde{i}$, we need, in the first place, to predict the treatment outcome under each competing scenario $\tilde{y}^a = y^a_{\tilde{i}}$, for $a = 1, \ldots, T$. The posterior predictive distribution arises as a natural choice to perform this task. Given the observed responses for the n^a patients previously treated with therapy a, that is $\boldsymbol{y}^a$, the predictive probability of response level k under treatment a is

$$p(\tilde{y}^a = k \mid \boldsymbol{y}^a, \boldsymbol{z}^a, \boldsymbol{x}^a, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}), \tag{3.13}$$

where $\tilde{\boldsymbol{z}} = \boldsymbol{z}_{\tilde{i}}$ and $\tilde{\boldsymbol{x}} = \boldsymbol{x}_{\tilde{i}}$ denote the P and Q dimensional vectors containing prognostic and predictive markers for the new patient. It is easy to show that

equation (3.13) depends on the posterior predictive distribution of cluster-specific parameters. In Section 3.6.1 we show that for PPMx models posterior predictive distributions are readily available.

Equation (3.13) can be directly employed in treatment selection for binary outcomes. Assuming that outcome category 1 represents a larger utility than category 0, the optimal treatment would be the one ensuring the largest $p(\tilde{y}^a = 1 \mid \cdot)$ for $a = 1, \ldots, T$. Following Ma et al. (2016) in Section 3.6.2, we adopt a utility approach for the selection of competing treatments with multinomial outcomes.

3.6.1 PPMx posterior predictive distribution

To obtain the posterior predictive distribution in (3.13) we need to first assign the untreated patient $\tilde{i}$ to one of the $C^a_{n^a}$ existing cluster or to a new one and then obtain the posterior predictive distribution for $\tilde{\boldsymbol{\eta}}^a = \boldsymbol{\eta}^a_{\tilde{i}}$:

$$p(\tilde{\boldsymbol{\eta}}^a|\tilde{\boldsymbol{x}}, \boldsymbol{y}^a, \boldsymbol{x}^a) = \int p(\tilde{\boldsymbol{\eta}}^a|\tilde{\boldsymbol{x}}, \Pi^a_{n^a+1}, \boldsymbol{y}^a, \boldsymbol{x}^a) \mathrm{d}p(\Pi^a_{n^a+1}|\tilde{\boldsymbol{x}}, \boldsymbol{y}^a, \boldsymbol{x}^a),$$

where $\Pi^a_{n^a+1} = (\Pi^a_{n^a} \cup \tilde{e})$. Müller et al. (2011) show that covariate dependent predictive distributions are readily available from PPMx model. In particular, the prior for $\Pi^a_{n^a+1}$ can be written as

$$P(\Pi^a_{n^a+1}|\Pi^a_{n^a},) = P(\tilde{e}^a = j|\Pi^a_{n^a}) \propto \begin{cases} \dfrac{\rho(S^a_j \cup \{\tilde{i}\})g(\boldsymbol{x}^{a\star}_j \cup \{\tilde{\boldsymbol{x}}^a\})}{\rho(S^a_j)g(\boldsymbol{x}^{a\star}_j)} & \text{for } j = 1, \ldots, C^a_{n^a} \\ \rho(\{\tilde{i}\})g(\{\tilde{\boldsymbol{x}}^a\}) & \text{for } j = C^a + 1. \end{cases} \tag{3.14}$$

Samples for the posterior predictive distribution from $p(\tilde{\boldsymbol{\eta}}^a|\tilde{\boldsymbol{x}}^a, \boldsymbol{y}^a, \boldsymbol{x}^a)$ can be obtained on top of posterior simulation for $\Pi^a_n \sim p(\Pi^a_n|\boldsymbol{x}, \boldsymbol{\eta})$, that are collected within the MCMC (Page and Quintana, 2015).

3.6.2 Predictive utility

To facilitate treatment selection for multinomial ordinal outcomes, we adopt utility weights. In clinical oncology, response categories are ordinal and consider changes in tumor size and/or distant migration after the treatment. We establish utility weights that turn a multinomial setting into a one-dimensional selection criterion considering the relative importance of each level of the ordinal response. Let $\boldsymbol{\omega}$ be a $K-$dimensional vector denoting the utility assigned to tumor response levels. We can then compute the median predictive utility for patient $\tilde{i}$ as:

$$\varphi^a(\tilde{i}) = \sum_{k=1}^{K} \omega_k \mu_{1/2}(y_{\tilde{i}} = k \mid \boldsymbol{y}^a, \boldsymbol{z}^a, \boldsymbol{x}^a, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{x}}). \tag{3.15}$$

The $\tilde{i}-$th patient will be assigned to the therapy ensuring the largest predictive utility, that can be considered to be optimal among the competing treatments.

The predicted optimal treatment for patient i is easily obtained from (3.15):

$$A(\tilde{i}) = \arg\max_a \left\{\varphi^a(\tilde{i})\right\}. \tag{3.16}$$

3.7 Simulation Study

We carry out a comparative study on simulated data to evaluate the performances of our method. We compare the proposed integrative method for personalized treatment selection (treat-ppmx) with the two-stage Bayesian predictive method proposed by Ma et al. (2019) using three different clustering procedures, namely K-means (km-bp), Partitioning Around Medioids (pam-bp) and Hierarchical Clustering (hc-bp).

The approach proposed by Ma et al. (2019) employs the Consensus Clustering method (Monti et al., 2003) that determines clustering for a specified number of clusters C. Since the number of groups in a sample is not known in practice, C is to be selected using a LOOCV for each simulated patient. Note that this is not true for our approach and that our method does not need any effort for the specification of the number of clusters.

In Section 3.7.1 we describe the data generating process. Section 3.7.2 gives details on the construction of the measures adopted for the performance comparison of the methods. Finally, in Section 3.7.3 we present the results of the simulation study and provide a brief discussion.

3.7.1 Generating mechanism

The simulated scenarios considered in our studies are constructed on the strategy devised in Ma et al. (2016, 2019). That is, we do not simulate from our model. In order to emulate the correlation structure that characterizes sequencing data, prognostic and predictive covariates are obtained from a real leukemia dataset. The data available from Golub et al. (1999) provide gene expression levels from 5000 genes, collected across 38 patients, of which 11 were diagnosed with acute myelogenous leukemia and the remaining with acute lymphoblastic leukemia. To obtain scenarios with a larger sample size, Ma et al. (2016, 2019) devised a procedure to expand the dataset, yielding $n = 152$ patients (38×4) with $Q = 90$ predictive and $P = 2$ prognostic biomarkers. This procedure is presented in detail in the Supplementary material of Ma et al. (2019).

The patients are assigned to $T = 2$ competing treatments and 3 levels of the ordinal-valued response variable are considered.

Generating treatment response

Since the observed treatment endpoints were unavailable, the treatment response is generated using two continuation-ratio logistic functions. The first one characterizes the effect of the predictive markers

$$r_{1k}^{a}(\boldsymbol{x}_i) = \left(\frac{P^a(y = k|\boldsymbol{x}_i)}{p^a(y < k|\boldsymbol{x}_i)}\right) = \alpha_{1k}^{a} + \boldsymbol{\beta}_{1k}^{a}\psi_2(\boldsymbol{x}_i), \text{ for } i = 1, \ldots, n, \tag{3.17}$$

where $\psi_2(\cdot)$ is a one-dimensional function of the first two principal components, used to summarize the information carried by predictive markers. Response-level probabilities for prognostic features are defined through the second continuation-ratio logistic function:

$$r_{2k}^{a}(\boldsymbol{z}_i) = \left(\frac{P^a(y = k|\boldsymbol{z}_i)}{p^a(y < k|\boldsymbol{z}_i)}\right) = \alpha_{2k}^{a} + \boldsymbol{\beta}_{2k}\boldsymbol{z}_i, \text{ for } i = 1, \ldots, n. \tag{3.18}$$

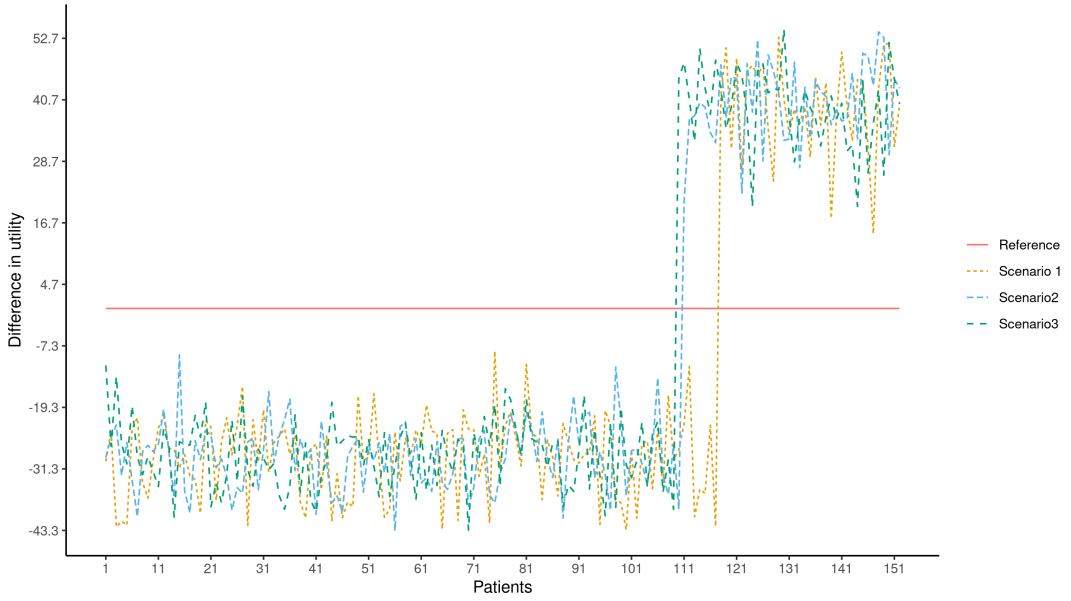


Figure 3.2: Differences in the true mean treatment utilities (the mean utility of treatment 2 minus the mean utility of treatment 1) by patients for Scenarios 1-3. Positive values indicate that treatment 2 is more beneficial and vice versa.

The coefficients $\alpha^a_{1k}, \alpha^a_{2k}, \boldsymbol{\beta}_{2k}, \boldsymbol{\beta}^a_{1k}$ were set to values that could produce realistic response rates, see Ma et al. (2019) for more details. Probability for each level of the ordinal response variable was generated as the pointwise product of (3.17) and (3.18) for each patient.

Finally, the optimal treatment for each simulated patient is determined as the inner product between the ordinal response probability ad the response level utility weight $\boldsymbol{\omega}$. In particular, we set $\boldsymbol{\omega} = (0, 40, 100)^\top$ to make the ordinal response reflect the clinical importance of each level Ma et al. (2016).

3.7.2 Performance evaluation

Prediction performances are compared in terms of the following metrics:

1. MOT: Misassigned to the Optimal Treatment;
2. $\%\Delta MTU_\ell$: Relative Gain in Treatment Utility;
3. NPC: Correctly Predicted Outcome.

The true optimal treatment is available since the treatment specific mean utilities are the pointwise product between equations (3.17) and (3.18). The true optimal treatment is:

$$A(i) = \arg\max_a \left\{ \sum_{k=1}^{K} \omega_k \left(r^a_{1k}(\boldsymbol{x}_i) r^a_{2k}(\boldsymbol{z}_i) \right) \right\}, \ i = 1, \ldots, n.$$

It follows that

$$MOT = n - \sum_{i=1}^{n} \mathbb{1}_{[A(i) - A(\tilde{i}) = 0]},$$

where $A(\tilde{i})$ is the predicted optimal treatment for patient i, defined in (3.16). MOT represents a first measure to compare the methods. Its interpretation is straightforward and lower values are associated with better selection rules.

Nonetheless, the extent to which a treatment is beneficial for each patient is heterogeneous and the improvement offered by a therapy varies from patient to patient. To account for this heterogeneity, performances should be evaluated also considering the relative utility gain. The relative gain in Treatment Utility, $\%\Delta MTU_\ell$ (Ma et al., 2016) allows to measure the overall benefit ensured by a treatment selection rule ℓ, in the case of $T = 2$ competing treatments. Denoting with $MTU_a(i)$ the mean treatment utility of treatment a for patient i, we can obtain the differential treatment utility as $\Delta MTU(i) = MTU_1(i) - MTU_2(i)$. Considering the true optimal treatment $A(i)$ and denoting with $t_\ell(i)$ the treatment recommended by selection rule ℓ, we can construct the indicator function $\delta_{t_\ell(i)}(A(i))$ that is defined as:

$$\delta_{t_\ell(i)}(A(i)) = \begin{cases} 1 & \text{if } t_\ell(i) = A(i) \\ -1 & \text{if } t_\ell(i) \neq A(i). \end{cases}$$

The sum of the true gains achieved by the selection rule ℓ is $\Delta MTU_\ell = \sum_{i=1}^{n} \delta_{t_\ell}(i)(A(i))|\Delta MTU(i)|$. The maximum possible gain in mean treatment utility varies in each simulation scenario. To make performance comparable also across scenarios, we consider the proportion of the maximum possible gain in total mean treatment utility attained by selection rule ℓ, that is:

$$\%\Delta MTU_\ell = \Delta MTU_\ell / \Delta MTU_{opt},$$

where ΔMTU_{opt} is the maximum possible total MTU, achieved when all patients are assigned to their optimal treatment. Finally, $\%\Delta MTU_\ell$ is bounded above by 1, when it always recommends the optimal treatment and $\%\Delta MTU_\ell = -1$ when it fails to select the optimal therapy for all the patients.

The last metric employed for performance comparison is NPC, that counts the number of patients for which the outcome was correctly predicted.

3.7.3 Simulation scenarios and results

We designed the simulated scenarios starting from a reference scenario and evaluated the methods on two arrays of scenarios of increasing complexity. The reference scenario (Scenario 1) is constructed plainly following the generating mechanism described in Section 3.7.1 using only 10 predictive biomarkers. The same applies to Scenarios 2a and 2b, which are defined by an increased number of predictive covariates (25 and 50, respectively). In order to evaluate the methods under a noisy framework, Scenarios 3a and 3b match the dimensions of Scenarios 2a and 2b, but only 10 predictive covariates were effectively used to generate the response variable, that is 15 and 40 variables, standard normal distributed, were added to the reference scenario, respectively. The methods are evaluated on this first array of scenarios using a LOOCV strategy. Note that, since the approaches proposed by Ma et al. (2019) are based on Consensus Clustering a nested LOOCV is needed: the inner loop is used to select the number of clusters, while the outer one is used to perform the prediction.

In order to emulate the large heterogeneity that sequencing data feature, we designed a second group of scenarios. We obtain a training and a testing set using

the generating mechanism described in Section 3.7.1, but the predictive covariates employed to generate the response differ in the training and in the testing set, in the sense that they overlap only to some extent. Obviously, for these scenarios a train and test strategy is employed, that is the model is fitted to the patients that belong to the training set, while the test set is used to judge the predictive performance of the models. The second array of scenarios is obtained as follows. Scenarios 4a and 4b present no difference in the generative mechanism between train and test set and they consider 25 and 50 predictive biomarkers, respectively. That is, they perfectly match Scenarios 2a and 2b. Nonetheless, since in the first pair of scenarios the methods are compared using a LOOCV strategy, the results are not comparable. The pairs of scenarios (5a, 5b) and (6a, 6b), match (4a, 4b) in terms of the number of predictive covariates, but predictive markers employed to generate the response in training and testing set overlap at 90% and 80%, respectively. Scenarios with 25 covariates are labeled with "a", while those with 50 covariates are labeled with "b". Table 3.2 summarizes the list of scenarios and their characteristics.

Table 3.2: List of simulation scenarios and their characteristics. The first array of scenarios (on the left hand side) is analysed with a LOOCV strategy. It varies in terms of number of predictive covariates and also considers covariates not employed to generate the response variable. The second array of scenarios (on the right hand side) is analysed with a train and test set strategy. It varies in terms of number of predictive covariates and in the extent to which predictive covariates overlap in the data generating mechanism of the train and the test set.

LOOCV			Train and Test		
Scenario	# covariates	# noisy variables	Scenario	# covariates	% overlap
1)	10	0	4a)	25	100
2a)	25	0	4b)	50	100
2b)	50	0	5a)	25	90
3a)	25	15	5b)	50	90
3b)	50	40	6a)	25	80
			6b)	50	80

We set $(\kappa, \sigma) = (1, 0.01)$. We assumed $\boldsymbol{\Lambda}_0$ to be a diagonal matrix with all elements on the diagonal equal to 10 and also assumed an identity matrix for $\boldsymbol{S}_0$. Finally we set $v_0 = 1$. Further details and the results of the sensitivity study on these parameters are provided in Appendix B.1. We ran the algorithm for $52,000$ iterations, with a burn-in period of $12,000$ iterations; chains were thinned and we kept every $10-$th sampled value. Comparisons are made in terms of MOT, $\%\Delta MTU_\ell$ and NPC. Reported values are averaged over 30 replicates, with standard deviations in parentheses.

Table 3.4: Prediction performances for Scenarios 1, 2a, 2b: mean across 30 replicated datasets (standard deviations are in parentheses). In each scenario and for each index the best performance is in bold.

	Scenario 1			Scenario 2a			Scenario 2b		
	MOT	$\%\Delta MTU_\ell$	NPC	MOT	$\%\Delta MTU_\ell$	NPC	MOT	$\%\Delta MTU_\ell$	NPC
pam-bp	19.7333 (4.1000)	0.7212 (0.0730)	75.6667 (8.3184)	15.2333 (3.0000)	0.8038 (0.0535)	77.2667 (6.2031)	16.0667 (6.6329)	0.7722 (0.1152)	77.0000 (7.4879)
km-bp	18.0000 (4.1606)	0.7516 (0.0742)	77.3667 (8.2816)	**8.7667** (5.6000)	**0.8792** (0.0863)	77.1333 (7.6777)	**9.7333** (8.4000)	**0.8604** (0.1399)	80.2000 (8.1385)
hc-bp	33.8333 (2.8172)	0.5531 (0.0736)	73.9667 (7.2468)	18.8667 (5.4944)	0.7308 (0.0985)	76.5667 (6.0326)	17.3333 (4.8090)	0.7967 (0.0885)	76.8333 (6.3305)
treat-ppmx	**13.5333** (3.3706)	**0.8341** (0.0449)	**81.8000** (7.3738)	14.3000 (8.8635)	0.8492 (0.0820)	**82.0333** (6.1839)	17.9000 (8.9070)	0.8312 (0.0878)	**84.7333** (6.4430)

Table 3.3: Prediction performances for Scenarios 3a, 3b: mean across 30 replicated datasets (standard deviations are in parentheses). In each scenario and for each index the best performance is in bold.

	Scenario 3a			Scenario 3b		
	MOT	$\%\Delta MTU_\ell$	NPC	MOT	$\%\Delta MTU_\ell$	NPC
pam-bp	19.1667 (3.4000)	0.7319 (0.0602)	75.9000 (7.8580)	30.4333 (7.2000)	0.5353 (0.1010)	72.5333 (7.0746)
km-bp	**16.8333** (2.8000)	**0.7637** (0.0442)	76.8667 (8.2993)	**19.3667** (2.5000)	**0.7266** (0.0347)	78.3667 (6.9406)
hc-bp	22.7333 (3.7040)	0.6881 (0.0569)	75.5000 (7.4498)	26.1667 (5.6000)	0.6354 (0.0633)	73.3333 (7.5992)
treat-ppmx	19.7333 (6.5859)	0.7396 (0.0768)	**82.2333** (5.7156)	27.0667 (12.4815)	0.6449 (0.1181)	**79.4000** (7.6906)

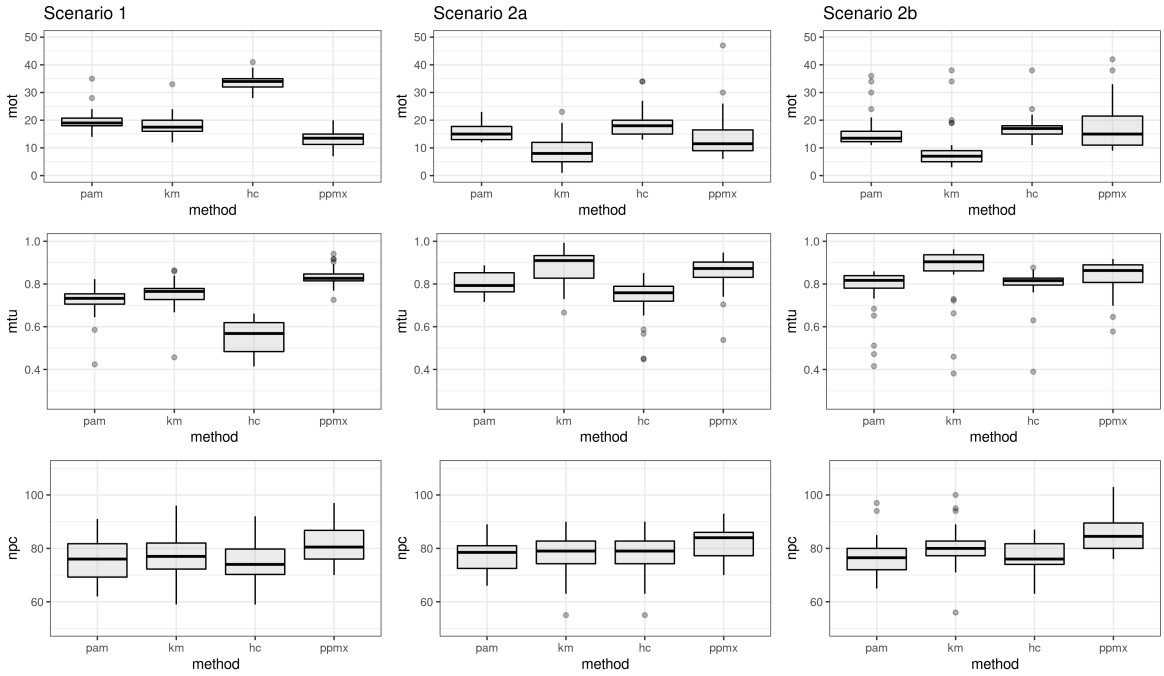


Figure 3.3: Prediction performances for Scenarios 1, 2a, 2b: boxplots display the distributions of values obtained for the summary measures.

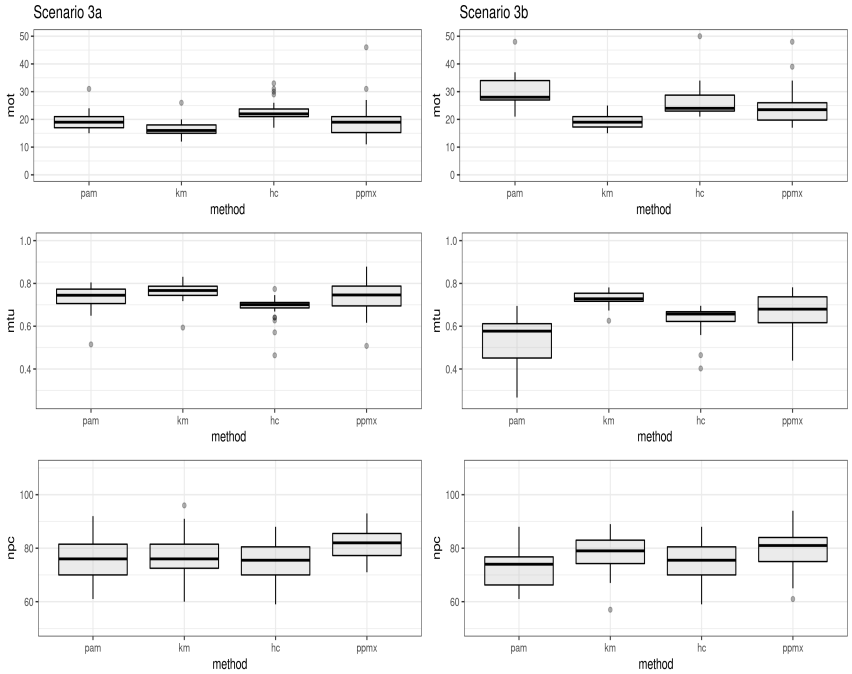


Figure 3.4: Prediction performances for Scenarios 3a, 3b: boxplots display the distributions of values obtained for the summary measures.

Based on the results provided by this first array of scenarios we can say that treat-ppmx outperforms competing two-stage models when the number of predictive biomarkers available is small (Scenario 1, reported in Table 3.4 and in Figure 3.3). This can be attributed to the capability of the covariate-dependent random partition to reach significant clustering arrangements even in the presence of a small amount of information, that is the case in which heuristic methods fail.

For an increasing number of covariates (Scenarios 2a and 2b, Table 3.4 and Figure 3.3), km-bp yields the best performances in terms of MOT and $\%\Delta MTU_\ell$. Nonetheless, among the methods based on heuristic clustering, km-bp is the only one that outperforms our method. Hc-bp and treat-ppmx present close values for the MOT in Scenario 2b.

Finally, our method consistently outperforms all the others in terms of NPC. This is probably due to the integrated prediction mechanism, able to fully account for the uncertainty in the clustering; note that, for the proposed method, optimal treatment misassignment often pertains to patients with similar utility across treatments.

Scenarios 3a and 3b (Table 3.3 and Figure 3.4) consider an additional factor of difficulty, that is the presence of noisy covariates. This yields all methods to poorer performances, especially in Scenario 3b which features a larger number of predictive markers. The same consideration carried out with respect to Scenario 2a and 2b are valid here, since the pattern of the best results is the same.

Table 3.5 and Figure 3.5 report the performances of the competing methods on Scenarios 4a, 4b, 5a, 5b, 6a, 6b. When a moderate amount of predictive covariates is considered, our method outperforms the competitors or attains performances that are close to km-bp. km-bp is the best performing approach among the two-stages methods. For an increasing number of covariates, the discrepancy between our method and km-bp widens, with the latter attaining the best results (in terms of MOT and $\%\Delta MTU_\ell$). The pairs of scenarios of this second array are sorted in

Table 3.5: Prediction performances for Scenarios 4a, 4b, 5a, 5b: mean across 30 replicated datasets (standard deviations are in parentheses). In each scenario and for each index the best performance is in bold.

	Scenario 4a			Scenario 4b		
	MOT	$\%\Delta MTU_\ell$	NPC	MOT	$\%\Delta MTU_\ell$	NPC
pam-bp	6.1667 (3.4450)	0.5838 (0.2078)	14.7333 (2.5452)	5.7667 (3.1697)	0.5561 (0.2034)	14.2000 (2.1560)
km-bp	**4.5333** (2.9912)	**0.6749** (0.1988)	**15.3000** (2.6017)	**4.3667** (2.0083)	**0.6919** (0.1400)	13.4333 (2.2389)
hc-bp	6.8000 (1.9191)	0.5727 (0.1386)	13.9000 (2.3686)	6.5667 (1.9061)	0.5580 (0.1786)	12.7333 (1.8742)
treat-ppmx	5.8000 (3.8721)	0.6687 (0.2590)	15.0333 (2.3851)	7.4667 (3.7207)	0.5938 (0.2462)	**15.3333** (2.3829)
	Scenario 5a			Scenario 5b		
	MOT	$\%\Delta MTU_\ell$	NPC	MOT	$\%\Delta MTU_\ell$	NPC
pam-bp	6.1333 (3.2982)	0.5773 (0.2060)	9.1333 (2.1930)	5.8000 (3.3052)	0.5578 (0.2216)	13.7333 (2.2581)
km-bp	**4.5333** (3.0596)	0.6840 (0.1946)	10.7333 (1.6802)	**5.0333** (2.1413)	**0.6438** (0.1452)	13.2000 (1.9896)
hc-bp	7.3667 (2.0254)	0.5165 (0.1737)	11.6333 (2.7353)	6.8667 (1.9250)	0.5270 (0.1782)	12.7333 (1.7604)
treat-ppmx	4.6000 (2.2984)	**0.7411** (0.1384)	**14.8667** (2.7883)	7.9000 (5.3906)	0.5478 (0.3671)	**15.2000** (2.2804)
	Scenario 6a			Scenario 6b		
	MOT	$\%\Delta MTU_\ell$	NPC	MOT	$\%\Delta MTU_\ell$	NPC
pam-bp	5.9333 (3.5324)	0.5991 (0.2236)	15.1333 (2.3154)	5.4333 (2.7503)	0.5749 (0.2053)	12.3667 (2.2047)
km-bp	5.1667 (3.1303)	0.6411 (0.2061)	15.0667 (2.1645)	**4.8333** (2.9721)	0.6445 (0.2261)	11.7667 (2.0288)
hc-bp	6.8000 (1.8080)	0.5706 (0.1400)	14.9667 (2.4422)	7.1000 (2.1870)	0.5044 (0.2028)	12.1000 (1.9360)
treat-ppmx	**4.1000** (3.8983)	**0.7583** (0.2500)	**15.5333** (2.2397)	5.9333 (3.5809)	**0.6735** (0.2367)	**14.6333** (2.6325)

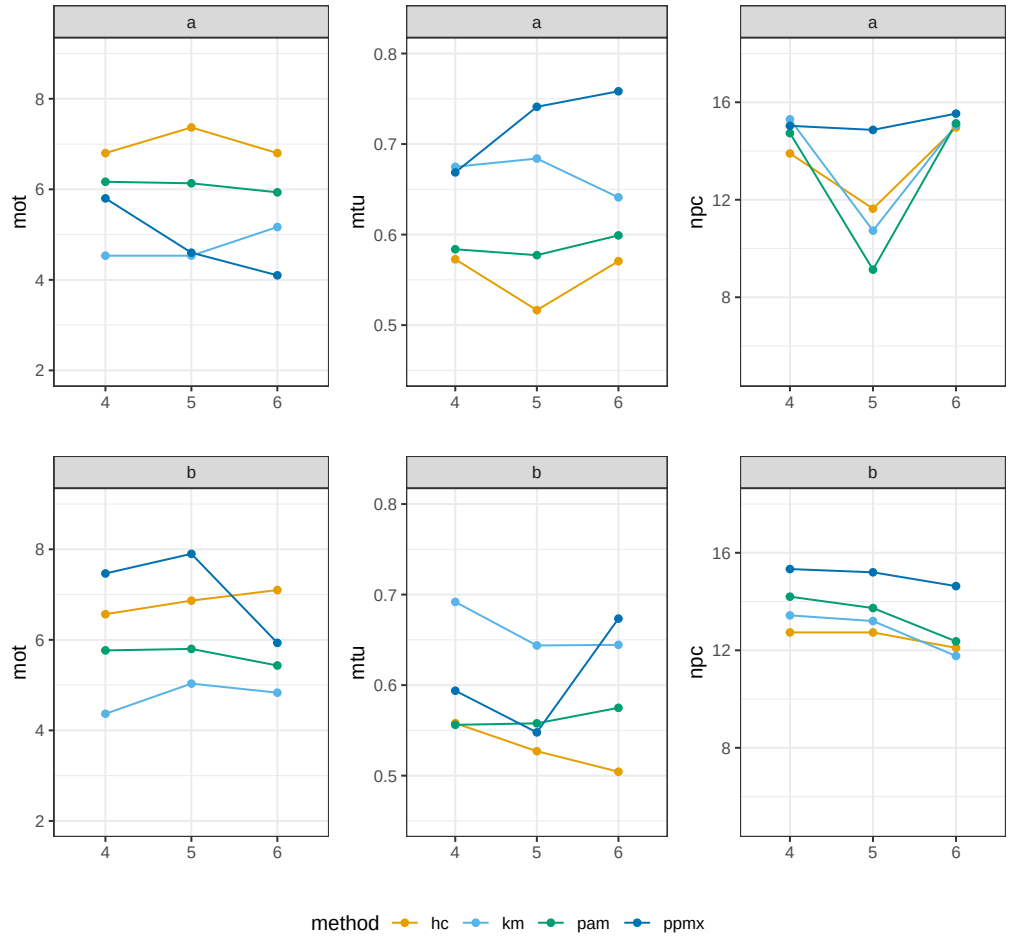


Figure 3.5: Line plots for summary measures in Scenarios (4a, 5a, 6a) and (4b, 5b, 6b). The first row reports MOT, $\%\Delta MTU_\ell$, NPC of the scenarios with 25 covariates (those with label "a"). The second row reports MOT, $\%\Delta MTU_\ell$, NPC of the scenarios with 50 covariates (those with label "b"). A single plot displays results, with respect to a single summary measure, that each competing method attained in three scenarios with common dimension. The scenarios are reported on the "x" axis and are ordered for increasing level of heterogeneity (that is decreasing level of overlap).

order of increasing level of heterogeneity. In fact, Scenarios (4a, 4b) feature a perfect overlap between the set of predictive markers used to generate the response in the train and the test set, while in Scenarios (5a, 5b) and (6a, 6b) this overlap is reduced to 90% and 80%, respectively. It is interesting to observe that our method is the only approach that performed better as the level of heterogeneity increases. It is again imputable to the fact that prediction and clustering are jointly performed in an integrative framework.

Our simulation study suggests that when a limited amount of predictive biomarkers are available, treat-ppmx should be preferred over two-stage methods. However, as the number of covariates grows, methods based on heuristic clustering algorithms should be adopted. In particular, from our results, km-bp obtains the best performances in terms of MOT and $\%\Delta MTU_\ell$. Nonetheless, the advantage offered by two-stage methods diminishes for large heterogeneity. In that case, regardless of the number of predictive covariates, the treat-ppmx should be preferred over all the competing approaches considered.

3.8 Case Study of Lower-grade Glioma

3.8.1 Lower-grade glioma

Glioma is the most frequent brain tumor: it makes up approximately 30% of all brain and central nervous system tumors and 80% of all malignant brain tumors (Goodenberger and Jenkins, 2012). Gliomas are classified as grades I to IV based on histological criteria established by the World Health Organization (WHO). Grade I tumors are generally circumscribed benign tumors with favorable prognosis, while grades II-IV comprise more aggressive tumors (diffuse gliomas). Grade II and grade III gliomas are usually referred to as low-grade glioma (LGG), which may eventually progress to grade IV, high-grade glioma.

Most LGG patients undergo resection and then receive radiotherapy and/or chemotherapy. Nonetheless, these standard procedures have proved to be largely inadequate (Claus et al., 2015). LGG exhibits significant molecular heterogeneity (Weiler and Wick, 2012; Cohen and Colman, 2015; Olar and Sulman, 2015), and many research efforts are now devoted to developing precision medicine for patients diagnosed with LGG (Ius et al., 2018; Taghizadeh et al., 2019; White et al., 2020).

Our goal is to leverage omics covariates to select the personalized optimal treatment.

3.8.2 TCGA data

We apply our method to the dataset analyzed in Ma et al. (2019), where clinical data and protein expression of patients affected by lower-grade glioma are collected from the TCGA data portal (now available at https://portal.gdc.cancer.gov/). Publicly available data underwent an accurate preprocessing, thoroughly documented in Ma et al. (2019) and briefly summarized in this section. The resulting LGG dataset considers patients that received standard and advanced treatments. A treatment qualifies as advanced if it includes targeted therapies or radiotherapy. Each group consists of 79 patients balanced in the covariates to account for potential selection bias. Tumor response is formulated in three ordinal levels: progressive disease (PD), partial response/stable disease (PS), and complete response (CR). Utility weights for treatment selection for ordinal outcomes are elicited, namely $\boldsymbol{\omega} = (0, 40, 100)$ to make the ordinal response reflect the clinical importance of each level (Ma et al., 2016). Finally, (Ma et al., 2019) selected 23 predictive features and 2 prognostic markers (ACVRL1-R-C and HSP70-R-C) through univariate association analyses.

3.8.3 Empirical summary measure

TCGA data do not provide the true optimal treatment, hence only the NPC measure, among those discussed in Section 3.7.2, can be used. However, we employed another summary to evaluate the relative increase in the population response rate attributable to a proposed treatment allocation method when compared with random allocation. This empirical summary measure (ESM) is proposed in Song and Pepe (2004) and its empirical efficacy is discussed in Kang et al. (2014); Ma et al. (2016) and Ma et al. (2019).

Let $Y = \{0, 1\}$ be the binary outcome variable. We define the treatment contrast as $\Delta(\boldsymbol{X}, \boldsymbol{Z}) = P(Y = 1|A = 2, \boldsymbol{X}, \boldsymbol{Z}) - P(Y = 1|A = 1, \boldsymbol{X}, \boldsymbol{Z})$, where $A = \{1, 2\}$

denote the non-targeted and targeted treatment, respectively. Indicating with $P(Y = 1|A)$ the probability of being a respondent under a randomized treatment assignment, we obtain the relative increase in the population response rate under a personalized treatment selection rule as:

$$\begin{aligned} ESM = &\{P(Y = 1|A = 2, \Delta(\boldsymbol{X}, \boldsymbol{Z}) > 0) \times P(\Delta(\boldsymbol{X}, \boldsymbol{Z}) > 0) + \\ &P(Y = 1|A = 1, \Delta(\boldsymbol{X}, \boldsymbol{Z}) < 0) \times P(\Delta(\boldsymbol{X}, \boldsymbol{Z}) < 0)\} - P(Y = 1|A). \end{aligned}$$

The overall response rate under a randomized treatment assignment $P(Y = 1|A)$ can be estimated as the sample proportion of respondents. The clinical benefit that is attributable to the proposed method is defined as the response rate for patients assigned by the proposed method ($P(\Delta(\boldsymbol{X}, \boldsymbol{Z}) > 0)$ vs $P(\Delta(\boldsymbol{X}, \boldsymbol{Z}) < 0)$) to the treatment actually received ($A = 2$ or $A = 1$). Defining n^1 and n^2 as the number of patients that received treatment $A = 1$ and $A = 2$, respectively, the weights $P(\Delta(\boldsymbol{X}, \boldsymbol{Z}) < 0)$ and $P(\Delta(\boldsymbol{X}, \boldsymbol{Z}) > 0)$ can be estimated as n^1/n and n^2/n, respectively. ESM measures the gain in the clinical benefit obtained under a particular treatment selection rule. Note that we based this summary measure on only two response categories, respondents (CR) and non-respondents (PD + PS); whereas, we used all three levels of the ordinal outcome in the data analysis and to implement personalized treatment selection.

3.8.4 Preliminary results

In this section, we applied the proposed method to the LGG dataset alongside the approach proposed by Ma et al. (2019). Table 3.6 reports NPC and ESM summary measures computed from assignments obtained by adopting a 10-fold cross-validation strategy.

To robustify inferences with respect to κ, σ, we define a prior distribution on (κ, σ) to overcome a critical trade-off that occurs when κ and σ are fixed. In particular, smaller values of σ give strong support to a moderate number of clusters, which is typically the case, but at the cost of losing the reinforcement mechanism. We assume a prior distribution on $(\kappa, \sigma) \in (0, \infty) \times (0, 1)$ to let the data choose the appropriate reinforcement rate (Lijoi et al., 2007). Namely, we adopted a discrete prior on a grid 5×5 in $(0, 20) \times (0, 1)$, which assigns uniform probability to all the combinations of (κ, σ).

We ran the algorithm for $52,000$ iterations, with a burn-in period of $12,000$ iterations; chains were thinned and we kept every $10-$th sampled value. The proposed treat-ppmx outperforms competing methods, both in terms of NPC and ESM. These results are consistent with those obtained in our simulations studies, especially in scenarios featuring significant heterogeneity and a moderate number of predictive covariates (Scenarios 2a and 3a). In particular, treat-ppmx attains an ESM of 0.0713, representing a 20% increase in the response rate, compared to randomized assignment with a response rate of $58/158 \simeq 0.361$. ESM for km-bp is 0.0495 reflecting an increase of 14%.

Table 3.6: LGG data 10-fold cross validation.

	NPC	ESM
pam-bp	67	0.0296
km-bp	57	0.0495
hc-bp	53	0.0275
treat-ppmx	**71**	**0.0713**

Figure 3.6 reports the heatmaps of the estimated posterior probabilities of co-clustering for the two treatments. Patients show pronounced heterogeneity, particularly those assigned to Treatment 1. The absence of a sharp separation between clusters in Treatment 1 (left pane) demonstrates a significant uncertainty in the clustering. Patients assigned to Treatment 2 (right pane) form most separate clusters, but the low probability of co-clustering still indicates a large variability in the clusters' production.

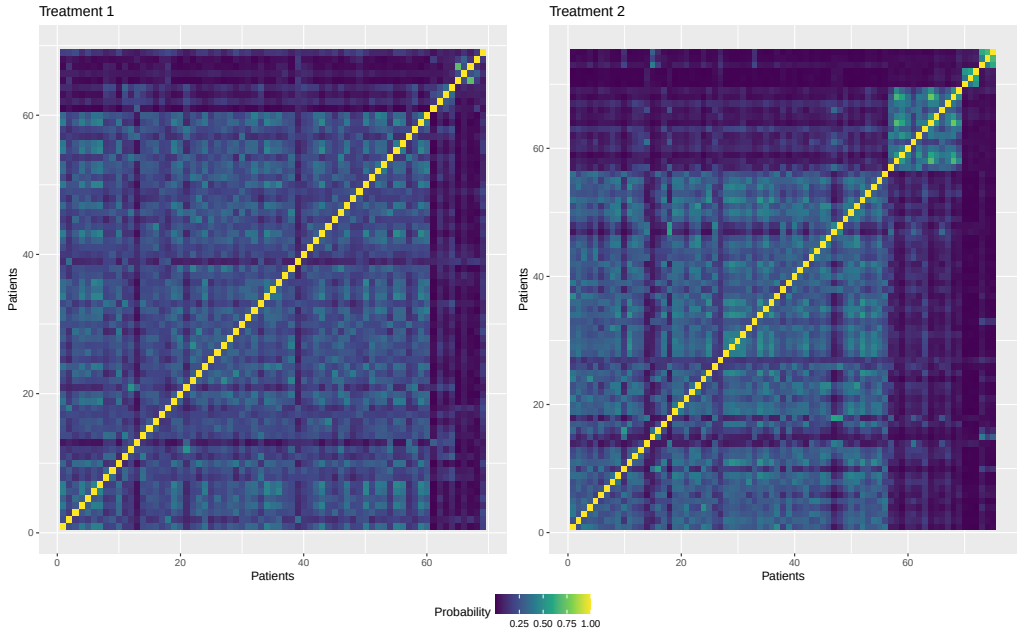


Figure 3.6: Heatmaps of the estimated posterior probabilities of co-clustering for Treatment 1 (left) and Treatment 2 (right) in fold 10.

This should not be interpreted as a bad model fit. In fact, the main goal of the proposed model is prediction. The random partition model is leveraged mainly for a density estimation purpose, rather than clustering. The random intercepts $\boldsymbol{\eta}^a$ may feature several modes since they embed uncertainty in the patients that arises from their predictive determinants.

3.9 Discussion

We proposed a novel Bayesian model aimed at the selection of the personalized optimal treatment in oncology when a *predictive signature* and a set of prognostic biomarkers are available. A well-established strategy to integrate prognostic and predictive covariates is to cluster patients into groups that are homogeneous with

respect to their predictive markers, within each treatment. In a subsequent step, this clustering arrangement is used to adjust the baseline probability of response to treatment obtained by prognostic factors. The advantage that the proposed approach has over existing models is that in our method model-based clustering and treatment assignment are jointly estimated from the data, that is, treatment selection fully accounts for patients heterogeneity.

We employed a Bayesian nonparametric model for random partition to build our integrative approach. In particular, we explored the use of the NGG process as cohesion function in a product partition model with covariates. The resulting predictive model proved to perform adequately well in the simulation study we conducted. Treat-ppmx outperforms competing methods in scenarios characterized by a moderate number of covariates. When the *predictive signature* is large, the two-stage approach may yield to better results, as long as heterogeneity is negligible. Moreover, a careful choice of the heuristic clustering method is still needed.

Our approach does not provide clearly separated posterior clusters in the analysis of the LGG dataset. This phenomenon is probably due to the significant heterogeneity that patients exhibit. Nonetheless, we are evaluating alternative modeling choices to improve clustering. In particular, we are exploring different similarity functions.

Extensions of the proposed model are possible. A delicate issue in models for random partition is the centering of the process, that is the choice of the base measure μ_0. This problem is often addressed with an empirical Bayes approach (e.g. Lijoi et al. (2007)), that is the prior distribution is estimated from the data. Since in our case the random partition distribution acts as prior on a random effect, rather than on data, this approach is not feasible. Nonetheless, it would be interesting to consider adaptive algorithms to improve the exploration of the parameter space and obtain better cluster-specific estimates. Note that the flexibility-computational cost trade-off needs to be evaluated.

Finally, a natural step would be to rephrase our strategy under the causal inference framework. In fact, precision medicine is committed to making the clinical decision-making process evidence-based and the causal approach provides the methodology for a better understanding of the actual effects of specific treatments on specific patients, that is the causal inference of individualized treatment effects.

Chapter 4

Final Remarks

In this thesis, we have presented two flexible regression models to analyze complex genomic and health data exhibiting large heterogeneity. The proposed methods are developed under the Bayesian framework to fully account for the uncertainty in the data generating process, parameter estimation, and model selection. Quantifying all sources of uncertainty is crucial because several models may explain the data equally well and hence point estimators are often not adequate. For each method, we developed state-of-the-art samplers for conducting posterior inference. To obtain efficient implementations, MCMC algorithms are written in R and C++ through the use of the Rcpp and RcppArmadillo libraries.

In Chapter 2, we propose a version of the Dirichlet-multinomial regression where coefficients are allowed to vary smoothly with the covariates. The resulting varying coefficients are carefully constructed to obtain linear and nonlinear interactions as special cases. Motivated by a recent CRC study, we investigate the effect of clinical factors and diet-related covariates on the microbiota compositions at the *phylum* level; for the patients enrolled in this study, microbiota abundance counts are collected from three different districts, namely tumor, fecal and salivary samples. We develop a high-dimensional Bayesian hierarchical model that exploits subject-specific regression coefficients to simultaneously borrow strength across districts and include complex interactions between diet and clinical factors if supported by the data. The proposed method identifies relevant associations through model selection priors and thresholding mechanisms.

We are currently evaluating alternative modeling strategies to extend the proposed method beyond the Dirichlet-multinomial regression framework. In fact, Dirichlet-multinomial regression only admits negative correlations among counts and alternative modeling approaches may be better suited for the analysis of microbiota data that exhibit positive correlations. Moreover, DM restricts the analysis of microbiota at a single taxonomy level, while accounting for the whole microbiome phylogenetic tree may provide precious insights, since it represents evolutionary relations among *taxa*.

In Chapter 3 we develop a predictive model to select the optimal treatment for oncological patients leveraging prognostic and predictive biomarkers. We first explore the use of NGGP as cohesion function in a model for dependent random partition. NGGP shows better theoretical properties and empirical performances with respect to the commonly used DP, due to its *reinforcement mechanism*. The resulting PPMx is employed to obtain clusters of observations that are more homogeneous with respect to predictive biomarkers, building partitions that are only partially exchangeable.

Model based clustering is jointly estimated with the effects of prognostic factors. An utility approach is finally employed to select the treatment that ensures the largest benefit for new untreated patients. This strategy results in an integrative predictive model, where prediction is able to fully account for patients' variability.

We are investigating further extensions of the PPMx. In particular, alternative modeling choices for the similarity could enable us to include a larger number of predictive markers. Finally, we developed this approach to leverage well-established molecular features for personalized treatment selection, rather than to pursue prognostic and predictive biomarkers discovery. In the light of this, it would be interesting to rephrase our strategy under the causal inference framework, with the goal of a better understanding of the actual effects of specific treatments on specific patients, that is the causal inference of individualized treatment effects.

Appendix A

Supplementary Material for Chapter 2

A.1 Identifiability

In this section we will first motivate the need and the rationale for linear constraints imposed in the varying coefficients (Section A.1.1). We will then explore and discuss whether these constraints have negative impact on inference. In particular, in Section A.1.2 we will assess through simple examples and some simulation studies whether inference is affected by a particular ordering of the covariates.

A.1.1 Linear constraints

Let's denote a general Bayesian model by the likelihood $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$ and prior $p(\boldsymbol{\theta})$ and then assume $\boldsymbol{\theta} = (\theta_1, \theta_2)$. According to Dawid (1979)'s definition of Bayesian identifiability, if $p(\theta_2 \mid \theta_1, \boldsymbol{y}) = p(\theta_2 \mid \theta_1)$, then θ_2 is not identifiable. Moreover, as $p(\theta_2 \mid \theta_1, \boldsymbol{y}) \propto \mathcal{L}(\theta_1, \theta_2; \boldsymbol{y}) p(\theta_2 \mid \theta_1) p(\theta_1)$, then θ_2 is unidentifiable if and only if $\mathcal{L}(\theta_1, \theta_2; \boldsymbol{y})$ is free of θ_2. Consequently, the lack of identifiability does not depend on prior specification (Gelfand and Sahu, 1999). Dawid's definition of Bayesian nonidentifiability is equivalent to a lack of identifiability in the likelihood. We want to investigate whether a lack of identifiability in the likelihood is present. Consider the hierarchical Dirichlet-multinomial framework proposed in Chapter 2. In the following illustrative setting we consider only linear effects and we do not assume any thresholding function. Specifically,

$$\begin{aligned} \log(\gamma_{ij}) &= \mu_j + \beta_j^x(z_i) x_i + \beta_j^z(x_i) z_i \\ \beta_j^x(z_i) &= \theta_x + b_{xz} z_i \\ \beta_j^z(x_i) &= \theta_z + b_{zx} x_i, \end{aligned} \tag{A.1}$$

where $x_i, z_i \in \mathbb{R}$. If $x_i, z_i \neq 0$

$$\begin{aligned} \log(\gamma_{ij}) &= \mu_j + \theta_x x_i + b_{xz} z_i x_i + \theta_z z_i + b_{zx} x_i z_i \\ &= \mu_j + \theta_x x_i + \theta_z z_i + (b_{xz} + b_{zx}) x_i z_i \\ &= \mu_j + \theta_x x_i + \theta_z z_i + b^\star x_i z_i \end{aligned} \tag{A.2}$$

where $b^\star = b_{xz} + b_{zx}$, hence there are infinite possible values for b_{xz}, b_{zx} whose sum is equal to $b^\star$. Formally, parameters b_{xz} and b_{zx} are not identifiable since the mapping

$(b_{xz}, b_{zx}) \to \mathcal{L}(\boldsymbol{y} \mid b_{xz}, b_{zx}, \cdot)$ is not injective. Even if this may not be a real issue for Bayesian inference, nonetheless in practice lack of identifiability may induce drift in the MCMC to extreme values in the overparametrized space, even if they remain stable in the lower dimensional subspace identified by the likelihood (Gelfand and Sahu, 1999; Xie and Carlin, 2006). We chose to work with an identifiable model; we achieve identifiability by imposing simple linear constraints, namely we set $b_{zx} = 0$ (in the same spirit of Zanella and Roberts (2020), among others). In the framework constructed in Chapter 2, b_{zx} corresponds to the set of terms $\{b_{qlj}\}_{q>l}$ that model interactions among binary factors generated from the use of discrete covariates as *effect modifiers* for main effects of binary factors. Moreover note that an analogous issue arise among terms that model interactions among continuous covariates. See Section A.1.2 for more details.

A.1.2 Effects on inference of linear constraints

The linear constraints imposed ensure model identifiability. In the following section we are going to address whether inference is affected by:

1. a switch in the arguments of the varying coefficients for continuous and binary variables, for a simplified model (linear interactions and no thresholding function);
2. a permutation in covariate indices, for a simplified model (linear interactions and no thresholding function);
3. a permutation in covariate indices, for the complete model.

Switching arguments

Assuming only linear interactions (i.e. $f_{pkj}(\boldsymbol{x}_k) = b_{pkj}\boldsymbol{x}_k$) and the absence of any thresholding function ($h(\cdot,\cdot)$) we will show that a switch in the arguments of the varying coefficients for continuous and binary variables results in the same model and hence does not affect the inference.

Under these two assumptions equations (2.5a) and (2.5b) reduce to the following:

$$\beta_{pj}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \theta_{pj} + \sum_{q=1}^{Q} b_{pqj} z_{iq} + \sum_{k>p} b_{pkj} x_{ik} \tag{A.3a}$$

$$\beta_{qj}(\boldsymbol{z}_i) = \theta_{qj} + \sum_{l>q} b_{qlj} z_{il}. \tag{A.3b}$$

Considering the "predictor" role of covariates, that makes interactions apparent, we obtain the following expressions:

$$\sum_{p=1}^{P} \beta_{pj}(\boldsymbol{x}_i, \boldsymbol{z}_i) x_{ip} = \sum_{p=1}^{P} \left[\theta_{pj} x_{ip} + \sum_{q=1}^{Q} b_{pqj}(z_{iq} x_{ip}) + \sum_{k>p} b_{pkj}(x_{ik} x_{ip}) \right] \tag{A.4a}$$

$$\sum_{q=1}^{Q} \beta_{qj}(\boldsymbol{z}_i) z_{iq} = \sum_{q=1}^{Q} \left[\theta_{qj} z_{iq} + \sum_{l>q} b_{qlj}(z_{il} z_{iq}) \right]. \tag{A.4b}$$

Restating equations (A.3) switching the arguments of two varying coefficients we obtain:

$$\beta_{pj}(\boldsymbol{x}_i) = \theta_{pj} + \sum_{k>p} b_{pkj} x_{ik} \tag{A.5a}$$

$$\beta_{qj}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \theta_{qj} + \sum_{l>q}^{Q} b_{qlj} z_{il} + \sum_{p=1}^{P} b_{pqj} x_{ip}. \tag{A.5b}$$

Including covariates as predictors we can see that the interaction pattern and the model that we obtain is the same of equations (A.4):

$$\sum_{p=1}^{P} \beta_{pj}(\boldsymbol{x}_i) x_{ip} = \sum_{p=1}^{P} \left[\theta_{pj} x_{ip} + \sum_{k>p} b_{pkj}(x_{ik} x_{ip}) \right]$$

$$\sum_{q=1}^{Q} \beta_{qj}(\boldsymbol{x}_i, \boldsymbol{z}_i) z_{iq} = \sum_{q=1}^{Q} \left[\theta_{qj} z_{iq} + \sum_{l>q} b_{qlj}(z_{il} z_{iq}) + \sum_{p=1}^{P} b_{pqj}(x_{ip} z_{iq}) \right].$$

We conclude that a switch in the arguments of the varying coefficients for continuous and binary variable does not affect inference, since results in the same model.

Permutation in covariate indices (linear interactions)

Let us assume that only linear interactions are included in the model, and no thresholding function. We are going to show that a permutation of covariates indices results in a linear predictor with the same terms (main effects or interactions) and coefficients.

In the following, we focus on coefficients that capture the interaction effects among binary factors $\{b_{qlj}\}$. It should be pointed out that, assuming only linear interactions, the same results hold also for $\{b_{pkj}\}$, the coefficients for linear interactions among continuous covariates in equation (2.5a). If we do not impose constraints on the coefficients while accounting for the *predictor* role of the covariates, the varying coefficient $\beta(\boldsymbol{z}_i)$ (equation (2.5b)) becomes[1]

$$\sum_{q=1}^{Q} \beta_q(\boldsymbol{z}_i) z_{iq} = \sum_{q=1}^{Q} \left[\theta_q z_{iq} + \sum_{l=1 \wedge l \neq q}^{Q} b_{ql}(z_{il} z_{iq}) \right].$$

Note that the constraint $l \neq q$ excludes the elements on the main diagonal of the $Q \times Q$ symmetric matrix $\boldsymbol{b}$. Since every interaction will now have two coefficients to capture its effect on the response, it is clear that this model is overparametrized and hence unidentifiable, as motivated in section A.1.1 of this Appendix. In order to ensure identifiability we impose a constraint on the indices. Considering $\{b_{ql}\}_{l>q}$ is equivalent to consider $\boldsymbol{b}$ a strictly upper triangular matrix. This simple constraints ensures identifiability in the likelihood and the interaction pattern that is built is invariant to any permutation in the indices.

Given a permutation σ of Q elements

$$\sigma : \{1, \dots, Q\} \to \{1, \dots, Q\},$$

[1] we dropped the subscript j for ease of notation

represented in a two line form by

$$\begin{pmatrix} 1 & 2 & \dots & Q \\ \sigma(1) & \sigma(2) & \dots & \sigma(Q) \end{pmatrix},$$

we will show that, if applied to our upper triangular matrix of interaction terms $\boldsymbol{b}$, does not affect inference, since a permutation is a bijective function of a set into itself. Considering the transformation that maps the set of strictly upper triangular $Q \times Q$ matrices into itself $T_\sigma : \mathcal{T}(Q \times Q) \to \mathcal{T}(Q \times Q)$ and obtaining a matrix $\bar{\boldsymbol{b}}$ from a permutation σ of the indices of an original matrix $\boldsymbol{b}$, such that

$$\bar{\boldsymbol{b}} = T_\sigma(\boldsymbol{b}),$$

we have that

$$\bar{b}_{ql} = \begin{cases} 0 & \text{if } q \leq l \\ b_{\min\{\sigma_{(q)},\sigma_{(l)}\}\max\{\sigma_{(q)},\sigma_{(l)}\}} & \text{if } q > l. \end{cases}$$

We can argue that no matter what permutation of the indices is taken, the interaction pattern is the same and hence it is not affect by the ordering of the covariates. Moreover, since the permutation is bijective, applying σ^{-1} to the permuted indices, the original order can always be recovered.

Example 3

A simple example of that can be taken considering a generic matrix $\boldsymbol{b} \in \mathcal{T}(5 \times 5)$

$$\boldsymbol{b} = \begin{pmatrix} 0 & b_{12} & b_{13} & b_{14} & b_{15} \\ 0 & 0 & b_{23} & b_{24} & b_{25} \\ 0 & 0 & 0 & b_{34} & b_{35} \\ 0 & 0 & 0 & 0 & b_{45} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and an arbitrary permutation $\sigma : \{1, 2, 3, 4, 5\} \to \{5, 3, 2, 1, 4\}$.

Permuting both rows (indexed by $q = 1, \dots, 5$) and columns (indexed by $l = 1, \dots, 5$) according to σ, we obtain $\bar{\boldsymbol{b}} = T_\sigma(\boldsymbol{b})$:

$$\begin{array}{cc} l & \\ & \sigma(l) \end{array} \quad \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 5 & 3 & 2 & 1 & 4 \end{array}$$

$$\bar{\boldsymbol{b}} = \begin{array}{cc} q & \sigma(q) \\ 1 & 5 \\ 2 & 3 \\ 3 & 2 \\ 4 & 1 \\ 5 & 4 \end{array} \quad \begin{pmatrix} 0 & b_{35} & b_{25} & b_{15} & b_{45} \\ 0 & 0 & b_{23} & a_{13} & b_{34} \\ 0 & 0 & 0 & b_{12} & b_{24} \\ 0 & 0 & 0 & 0 & b_{14} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

All the nonzero element in $\boldsymbol{b}$ are present also in $\bar{\boldsymbol{b}}$.

We perform a simulation study to empirically verify that a permutation in the order of the covariates does not affect the inference. Following the data generating mechanism described in Section 2.5.1, but considering only binary variables as effect

modifiers and no random effect, we generate a dataset with $n = 150$ observations. The dataset has the following dimension: $J = 10$ categories for the response, $P = 5$ continuous covariates and $Q = 5$ discrete covariates. In order to assess whether the permutation of the indices affected empirically our inference, we first fit our model to the data keeping the *original* ordering and then to the data with a random permutation in the indices of the covariates. We repeat this procedure for 100 different datasets and we report the results in terms of true positive rate, false positive rate and Matthews Correlation Coefficient in Table A.1.

Table A.1: Model selection performances: mean across 100 replicated datasets (standard errors are in parentheses). Evaluation is carried out on population parameters only.

	TPR	FPR	MCC
no permutation	0.7900 (0.0403)	0.0033 (0.0059)	0.8207 (0.1306)
permutation	0.8033 (0.0183)	0.0033 (0.0049)	0.8219 (0.1058)

As we can see the results are extremely similar, and the negligible difference should be attributed to the to the Monte Carlo error.

Permutation in covariate indices for the complete model

We will now address the effect on inference of the ordering of the covariates for the complete model. We will focus on continuous covariates and nonlinear effects since the discrete covariates are not associated with nonlinear effects.

After the decomposition is performed, we reparameterize $f_{pkj}(\boldsymbol{x}_k) = \boldsymbol{x}_k^\star \boldsymbol{\alpha}_{pkj}^\star + \boldsymbol{x}_k \alpha_{pkj}^0$, where $\boldsymbol{x}_k^\star$ is the orthogonal basis obtained from the spectral decomposition of the covariance of $\tilde{\boldsymbol{x}}_k \boldsymbol{\alpha}_{pkj}$ and $\boldsymbol{\alpha}_{pkj}^\star$ its the corresponding vector of coefficients, while α_{pkj}^0 is the coefficient of the linear term $\boldsymbol{x}_k$. This reparameterization is described in more detail in the manuscript (Section **??**). In order to restrict our focus on nonlinear terms let us consider the simpler case where binary covariates do not act as effect modifiers and no thresholding function is adopted. This results in the following expressions:

$$\beta_{pj}(\boldsymbol{x}_i) = \theta_{pj} + \sum_{k>p}^{P} \left(\boldsymbol{x}_{ik}^\star \boldsymbol{\alpha}_{pkj}^\star + x_{ik} \alpha_{pkj}^0 \right).$$

Considering continuous covariates also in their "predictor" role we have:

$$\sum_{p=1}^{P} \beta_{pj}(\boldsymbol{x}_i) x_{ip} = \sum_{p=1}^{P} \left[\theta_{pj} x_{ip} + \sum_{k>p}^{P} \left(\boldsymbol{\alpha}_{pkj}^\star (\boldsymbol{x}_{ik}^\star x_{ip}) + \alpha_{pkj}^0 (x_{ik} x_{ip}) \right) \right]. \tag{A.7}$$

It is apparent from equation (A.7) that, since $\boldsymbol{x}_k^\star$ is the orthogonal matrix corrisponding to the covariate $\boldsymbol{x}_k$, the considerations carried out previously on the invariance of the matrix of coefficients for linear interactions here hold only for the matrix $\boldsymbol{\alpha}_j^0$. The ordering of continuous covariates hence affects inference only through the nonlinear terms. Nonetheless, in the following simulation study we show that

a permutation of the ordering of the covariates does not critically undermine the inference we conducted with the proposed model.

In order to test the effect of permutation of covariates indices on inference we carried out a simulation study. Following the data generating mechanism described in Section 2.5.1, but considering only continuous variables as effect modifiers and no random effect, we generated a dataset with $n = 150$ observations. The dataset has the following dimension: $J = 10$ categories for the response, $P = 15$ continuous covariates and $Q = 5$ discrete covariates. In order to assess whether the permutation of the indices affected empirically our inference, we first fitted our model to the data keeping the "original" order and then to the data with a random permutation of the indices of the covariates. Note that this scenario is not comparable to those reported in Chapter 2. In fact this one has been designed purposefully to assess the effect of permutation in the labels for continuous covariates. We repeated this procedure for 100 different datasets, in the first study we don't consider the thresholding function, that is included in the second one. We report the results in terms of true positive rate, false positive rate, and Matthews Correlation Coefficient in Table A.2.

Table A.2: Model selection performances: mean across 100 replicated datasets (standard errors are in parentheses). Evaluation is carried out on population parameters only.

	permutation	TPR	FPR	MCC
no thresholding function	no	0.8201	0.0033	0.8169
		(0.0398)	(0.0029)	(0.0374)
	yes	0.8133	0.0081	0.8019
		(0.0425)	(0.0002)	(0.0360)
thresholding function	no	0.83933	0.0014	0.8337
		(0.0338)	(0.0384)	(0.0984)
	yes	0.8476	0.01737	0.8181
		(0.0375)	(0.0015)	(0.0274)

A.2 Choice of Hyperparameters and Spline Bases

Our subject-specific approach involves a large set of hyperparameters, whose specification needs to be discussed.

A.2.1 Spike-and-slab priors

As discussed and suggested in Wadsworth et al. (2017) we choose the prior for beta mixed binomial on main effects so that such specification could reflect the *a priori* expected number of included effects. Setting a_ω, b_ω such that $a_\omega + b_\omega = 2$ the prior expected mean value is $m = a_\omega/(a_\omega + b_\omega)$. A value of $m = \frac{a_\omega}{a_\omega + b_\omega}$ for $(a_\omega, b_\omega) = (0.05, 1.95)$ is equivalent to assuming that 2.5% of the associations are not active. It is common to choose a large value for the slab's variance (Chipman et al., 2001) to induce a noninformative prior (flat prior distribution) on the location of the coefficients.

A.2.2 Horseshoe priors

Hyperparameters for the $\{b_{pqj}\}$ and $\{b_{qlj}\}$ are set to default values. The horseshoe prior is a scale mixture of normals, with a half-Cauchy prior on the variance $\lambda_{pqj} \sim C^+(0,1)$ (Carvalho et al., 2009, 2010). The global hyperparameter ζ_{qj} was determined to be important towards the minimax optimality of the horseshoe posterior mean, and when a full Bayes version of the horseshoe prior is implemented, the truncated half-Cauchy prior is recommended by van der Pas et al. (2017), that is a half-Cauchy prior truncated to $[1/n, 1]$. The same choice is adopted for the $\{b_{qlj}\}$.

A.2.3 Number B of spline bases

Concerning the standard penalized spline framework, the approach proposed by Scheipl et al. (2012) results in a "post-processed set of orthonormal spline bases" that directly represent linear and nonlinear (penalized) effects. Specifically, in case of penalized splines with $k-$th order random walk prior, the space of unpenalized functions consists of all polynomials of order less than $k-1$. Separating these polynomials from the penalized part of the function allows the model to select the type of effect best supported by the data (i.e., the model can decide whether a continuous interaction should be included either as a nonlinear effect, as a linear effect, or be completely excluded from the model).

We model the smooth functions $\{f_{pkj}(\boldsymbol{x}_k)\}$ using cubic b-splines $f_{pkj}(\boldsymbol{x}_k) = \tilde{\boldsymbol{x}}_k \boldsymbol{\alpha}_{pkj}$, where $\tilde{\boldsymbol{x}}_k$ represents the design matrix of the spline bases for $\boldsymbol{x}_k$, and $\boldsymbol{\alpha}_{pkj}$ is the correspondent spline coefficient. In order to flexibly capture the nonlinearity of $\{f_{pkj}(\boldsymbol{x}_k)\}$ the number of bases B is set to a large value, namely 20. Since the spline coefficients are regularized by a roughness penalty, only the few bases that explain most of the variability of $\{f_{pkj}(\boldsymbol{x}_k)\}$ are retained. In fact, Scheipl et al. (2012)'s implementation makes use of an algorithm to compute only the largest r_k eigenvalues of $f_{pkj}(\boldsymbol{x}_k)$ and their associated eigenvectors. For example, only the first r_k eigenvectors and eigenvalues whose sum represent at least 99% of the sum of all eigenvalues are used to construct the reduced rank orthogonal basis $\boldsymbol{x}_k^\star$ with r_k columns. This leads to a large reduction in the dimension of $\boldsymbol{x}_k^\star$ and represents a robust criterion to select a pre-specified number of components. Note that 99% is commonly used as a default value since it results in a negligible loss of information but still in a significant dimensionality reduction. Usually, only 3 bases are retained in our case, hence a large B does not result in overfitting. Such a setting has been extensively tested also in Ni et al. (2019a).

A.2.4 peNMIG

Hyperparameters for peNMIG are set as follows: $(a_{\tilde{\omega}}, b_{\tilde{\omega}}) = (1,1)$, $v_0 = 0.00025$, $(a_{\tilde{\tau}}, b_{\tilde{\tau}}) = (5, 25)$ both for the linear part and the nonlinear one. This prior specification is a default setting proposed and discussed in Scheipl et al. (2012).

A.2.5 Threshold bounds

Prior elicitation for threshold parameters is thoroughly discussed in Supplemental Materials of Ni et al. (2019a) and we choose a standard uniform distribution for both t_x and t_z. The threshold parameter is interpreted as minimum effect size, so we

center the uniform prior at the smallest effect size we consider relevant with respect to the application. In order to avoid favoring the full model *a priori*, the upper bound is set to 1.

A.2.6 Intercepts

The prior for the variance parameter of the random intercepts $\{\iota_{s(i)}\}$ is $\sigma^2_\iota \sim$ Inverse-gamma$(1, 1)$, that is a weakly-informative prior (Gelman et al., 2006). The specification of the prior variance for the global intercept $\sigma^2_\mu = 10$ is meant to induce a weakly-informative prior also on the location of the coefficients.

A.3 Sensitivity Analysis

Since the analysis of CRC data required a careful tuning of some hyperparameters, we performed an investigation on the sensitivity of the results to these values. Results are reported in Table A.3. Specifically, we found that results were affected by the value of variance hyperparameters of the global intercept and by the value of the variance hyperparameters of the spike-and-slab prior. Hyperparameters on the Beta priors on $\{\theta_{pj}\}$'s and $\{\theta_{qj}\}$'s and the upper bound for the threshold parameter do not impact the results. In the case study, $a_\omega, b_\omega, \tau^2_j, \sigma^2_\mu$, and b_t have been set to values that ensured the best performance (in terms of MCC) in the sensitivity analysis conducted.

Table A.3: Model selection performances under different prior specification for parameters $a_\omega, \tau_j^2, \sigma_\mu^2$, and b_t. Mean across 100 replicated datasets (standard errors are in parentheses). For each parameter the best MCC is in bold. Evaluation is carried out on population parameters only.

$m = (a_\omega)/(a_\omega + b_\omega)$														
0.005			0.01			0.025			0.05			0.25		
TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC
0.66	0.04	0.43	0.67	0.04	0.42	0.68	0.03	**0.47**	0.67	0.05	0.40	0.71	0.04	0.44
(0.03)	(0.00)	(0.03)	(0.03)	(0.00)	(0.02)	(0.03)	(0.00)	(0.03)	(0.03)	(0.01)	(0.04)	(0.04)	(0.01)	(0.04)
τ_j^2														
1			5			10			50			100		
TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC
0.66	0.04	0.42	0.62	0.03	0.46	0.71	0.02	**0.56**	0.66	0.02	0.56	0.68	0.02	0.56
(0.04)	(0.00)	(0.03)	(0.04)	(0.00)	(0.04)	(0.04)	(0.00)	(0.05)	(0.01)	(0.00)	(0.02)	(0.03)	(0.00)	(0.04)
σ_μ^2														
0.01			0.1			1			10			100		
TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC
0.64	0.03	0.47	0.62	0.02	**0.53**	0.62	0.04	0.38	0.67	0.04	0.44	0.62	0.04	0.41
(0.03)	(0.00)	(0.02)	(0.03)	(0.00)	(0.04)	(0.05)	(0.01)	(0.04)	(0.03)	(0.01)	(0.02)	(0.03)	(0.01)	(0.04)
b_{t_x}, b_{t_z}														
.5			0.75			1.0			1.25			1.5		
TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC	TPR	FPR	MCC
0.62	0.04	0.38	0.67	0.04	0.43	0.67	0.04	**0.44**	0.65	0.04	0.42	0.62	0.04	0.41
(0.05)	(0.01)	(0.04)	(0.03)	(0.00)	(0.02)	(0.03)	(0.01)	(0.02)	(0.03)	(0.00)	(0.03)	(0.03)	(0.01)	(0.04)

A.4 Posterior Computation

We describe below the Markov Chain Monte Carlo (MCMC) we designed to obtain the posterior distribution of the parameters of interest. To construct an efficient algorithm and improve computational feasibility we adopted a data augmentation approach implemented in Wadsworth et al. (2017) and detailed in Koslovsky et al. (2020). The resulting MCMC is a Metropolis-Hastings within Gibbs. We will firstly describe the data augmentation scheme and then we will discuss parameters' sampling.

A.4.1 Data augmentation

Generating samples from the Dirichlet distribution using independent Gamma random variable is computationally efficient. Exploiting this property the data augmentation approach is based on a reparametrization of equation (2.1) and on the introduction of an auxiliary parameter. We first assume that the $J-$dimensional counts for the $i-$th sample $\boldsymbol{y}_i$ follows a Multinomial distribution:

$$\boldsymbol{y}_i \sim \text{ Multinomial}(y_i^+ \mid \boldsymbol{\phi}_i),$$

with $y_i^+ = \sum_{j=1}^J y_{ij}$ and $\boldsymbol{\phi}_i$ defined on the $J-$dimensional simplex $\mathcal{S}^{J-1}$. To account for extra-variation in the counts we specify a conjugate prior on the taxa probability:

$$\boldsymbol{\phi}_i \sim \text{ Dirichlet}(\boldsymbol{\gamma}_i)$$

with the $J-$dimensional vector $\boldsymbol{\gamma}_i = (\gamma_{i1}, \dots, \gamma_{iJ}), \gamma_{ij} > 0 \ \forall \ j$.

We introduce latent random variables $s_{ij} \overset{\text{iid}}{\sim} \text{Gamma}(\gamma_{ij}, 1)$ constructed such that $\phi_{ij} = s_{ij}/T_i$, where $T_i = \sum_{j=1}^J s_{ij}$, obtaining

$$\boldsymbol{y}_i \mid \boldsymbol{s}_i/T_i \sim \text{Multinomial}(y_i^+, \boldsymbol{s}_i/T_i),$$

where $\boldsymbol{s}_i$ is the $J-$dimensional vector $\boldsymbol{s}_i = (s_{i1}, \dots, s_{iJ})$. We can now write

$$p(\boldsymbol{y}_i, \boldsymbol{s}_i \mid \boldsymbol{\gamma}_i) \propto \frac{s_{i1}^{y_{i1}} \cdots s_{iJ}^{y_{iJ}}}{T_i^{y_i^+}} \prod_{j=1}^J \frac{1}{\Gamma(\gamma_{ij})} s_{ij}^{\gamma_{ij}-1} \exp(-s_{ij}). \tag{A.8}$$

The quantity $T_i^{y_i^+}$ is cumbersome to compute: this is why the augmentation step is implemented. We introduce n auxiliarity parameters u_i and let $u_i \mid T_i \sim \text{Gamma}(y_i^+, T_i)$ and by definition of the gamma density we have

$$\frac{1}{T_i^{y_i^+}} = \int_0^{+\infty} \frac{1}{\Gamma(y_i^+)} u_i^{y_i^+ - 1} \exp(-T_i u_i) \mathrm{d}u_i$$

we can rearrange equation (A.8) as

$$p(\boldsymbol{y}_i, \boldsymbol{s}_i \mid \boldsymbol{\gamma}_i) \propto \int_0^{+\infty} \frac{1}{\Gamma(y_i^+)} u_i^{y_i^+ - 1} \exp(-T_i u_i) s_{i1}^{y_{i1}} \cdots s_{iJ}^{y_{iJ}} \prod_{j=1}^J \frac{1}{\Gamma(\gamma_{ij})} s_{ij}^{\gamma_{ij}-1} \exp(-s_{ij}) \mathrm{d}u_i \tag{A.9}$$

hence

$$p(\boldsymbol{y}_i, \boldsymbol{s}_i, u_i \mid \boldsymbol{\gamma}_i) \propto \frac{1}{\Gamma(y_i^+)} u_i^{y_i^+ - 1} \exp(-T_i u_i) s_{i1}^{y_{i1}} \cdots s_{iJ}^{y_{iJ}} \prod_{j=1}^{J} \frac{1}{\Gamma(\gamma_{ij})} s_{ij}^{\gamma_{ij}-1} \exp(-s_{ij}).$$

The integral in equation (A.9) will be evaluated numerically with steps naturally embedded in the MCMC algorithm.

A.4.2 MCMC sampling

For posterior inference we designed a Metropolis within Gibbs sampler that employs component-wise adaptive Metropolis steps. The MCMC sampler goes as follows:

1. **Jointly update** $\{(\theta_{pj}, \xi_{pj})\}$. We employ the two-step scheme proposed in Savitsky et al. (2011), that comprises a between-model and a within-model step.

   - *between-model step.* Assuming uniform probabilities over the indices $j = 1, \ldots, J$ we randomly choose j. For each $p = 1, \ldots, P$ we jointly propose a new model such that
     - if $\xi_{pj} = 1$ we propose $\xi'_{pj} = 0$ and $\theta'_{pj} = 0$ (*delete step*). The proposal will be accepted with probability

$$\min\left\{\frac{p(\boldsymbol{s} \mid \boldsymbol{\theta}', \boldsymbol{\xi}', \cdot) p(\xi'_{pj})}{p(\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \cdot) p(\theta_{pj} \mid \xi_{pj}) p(\xi_{pj})}, 1\right\}$$

     - if $\xi_{pj} = 0$ we propose $\xi'_{pj} = 1$ and we sample θ'_{pj} from $N(\theta_{pj}, 0.5)$ (*add step*). The proposal will be accepted with probability

$$\min\left\{\frac{p(\boldsymbol{s} \mid \boldsymbol{\theta}', \boldsymbol{\xi}', \cdot) p(\theta'_{pj} \mid \xi'_{pj}) p(\xi'_{pj})}{p(\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \cdot) p(\xi_{pj})}, 1\right\}$$

   - *within-model step.* Assuming uniform probabilities over the indices $j = 1, \ldots, J$ we randomly choose j. For each $p = 1, \ldots, P$, if $\xi_{pj} = 1$, that is, if the covariate is currently included in the model, θ'_{pj} is proposed through an adaptive Metropolis-Hasting scheme. The new value θ'_{pj} is proposed using the proposal formulated in Roberts and Rosenthal (2009) but without the full covariance of the target, in a component-wise fashion:

$$\theta'_{pj} \sim 0.95 N(\theta_{pj}, 2.38^2 \times \hat{\sigma}^2_{pj}/J \times P) + 0.05 N(\theta_{pj}, 0.01/J \times P),$$

     where $\hat{\sigma}^2_{pj}$ is the estimate at the current iteration of the standard deviation of the target distributon. These values are updated via the recursive formula proposed in Haario et al. (2005). Each proposal is accepted with probability

$$\min\left\{\frac{p(\boldsymbol{s} \mid \boldsymbol{\theta}', \boldsymbol{\xi}', \cdot) p(\theta'_{pj} \mid \xi_{pj})}{p(\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \cdot) p(\theta_{pj} \mid \xi_{pj})}, 1\right\}$$

2. **Jointly update** $\{(\theta_{qj}, \xi_{qj})\}$. The same procedure described in Step 1 for the joint update of (θ_{pj}, ξ_{pj}) is carried out for (θ_{qj}, ξ_{qj}).

3. **Update** $\{b_{pqj}\}$. In order to update the parameters involved in the horseshoe prior we perform the following steps:

   - update $\{b_{pqj}\}$. Assuming uniform probabilities over the indices $j = 1, \ldots, J$ we randomly choose j. Given j for each $p = 1, \ldots, P$ and $q = 1, \ldots, Q$ we propose b'_{pqj} with the adaptive scheme in (1), and accept it with probability

   $$\min\left\{\frac{p(\boldsymbol{s} \mid b'_{pqj}, \cdot)p(b'_{pqj})}{p(\boldsymbol{s} \mid b_{pqj}, \cdot)p(b_{pqj})}, 1\right\}$$

   - update $\lambda_{pqj}, \zeta_{qj}$. The global-local scale parameters are updated through an adaptation of the slice sampling scheme given in the online supplement of Polson et al. (2014). We define $\varpi_{pqj} = 1/\lambda^2_{pqj}$ and $\varsigma_{pqj} = b_{pqj}/\zeta_{qj}$. This reparameterization allows us to employ slice sampler (Neal, 2003), as the conditional posterior distribution of ϖ_{pqj} is

   $$p(\varpi_{pqj} \mid \zeta_{qj}, \varsigma_{pqj}) \propto \exp\left\{-\frac{\varsigma^2_{pqj}}{2}\varpi_{pqj}\right\}\frac{1}{1+\varpi_{pqj}}.$$

   To sample λ_{pqj}:

   – draw a sample from Uniform distribution:

   $$u_{pqj} \mid \varpi_{pqj} \sim \mathrm{U}(0, 1/(1+\varpi_{pqj}));$$

   – draw a sample from truncated Exponential density, so that it has zero probability outside the interval $(0, (1-u_{pqj})/u_{pqj})$:

   $$\varpi_{pqj} \mid \varsigma_{pqj}, u_{pqj} \sim \mathrm{Exp}(2/\varsigma^2_{pqj}).$$

   Transforming back to the $\lambda-$scale it will ensure a sample from the conditional distribution of interest. The same applies for ζ_{qj}, replacing $\varpi = 1/\zeta^2_{qj}$ and $\varsigma^2_{qj} = \sum_{p=1}^{P} b^2_{pqj}/2$.

4. **Update** $\{b_{qlj}\}$. The same procedure described in Step 3 is carried out for $\{b_{qlj}\}$, with the constraint that the matrices $\boldsymbol{b}_j$ are strictly triangular.

5. **Update** $\{\alpha^0_{pkj}\}$ **and** $\{\boldsymbol{\alpha}^\star_{pkj}\}$. In this article, we decompose the nonlinear function $f_{pkj}(\boldsymbol{x}_k) = f^{pen}_{pkj}(\boldsymbol{x}_k) + f^0_{pkj}(\boldsymbol{x}_k)$ into a polynomial part (f^0_{pkj}) and a nonlinear part (f^{pen}_{pkj}) that are orthogonal to each other and we apply variable selection technique separately to each part. We will detail the step only for $\{\boldsymbol{\alpha}^\star_{pkj}\}$, as for $\{\alpha^0_{pkj}\}$ is defined analogously.

   Assuming uniform probabilities over the indices $j = 1, \ldots, J$ we randomly choose j. Note that r_k is the number of columns of the orthogonal matrix of the splines for $\boldsymbol{x}_k$, that is $\boldsymbol{x}^\star_k$. The relative index is $\kappa = 1, \ldots, r_k$.

   - For each $p = 1, \ldots, P,\ k = 1, \ldots, K$ we propose η'_{pkj} obtained from a random walk proposal and accept it with probability

   $$\min\left\{\frac{p(\boldsymbol{s} \mid \eta'_{pkj}, \cdot)p(\eta'_{pkj})}{p(\boldsymbol{s} \mid \eta_{pkj}, \cdot)p(\eta_{pkj})}, 1\right\};$$

- update $\{m_{pkj}\}$ by Gibbs:

$$p(m_{pkj} = 1 \mid \tilde{\boldsymbol{\psi}}_{pkj}) = \frac{1}{1 + \exp\left\{-2\tilde{\boldsymbol{\psi}}_{pkj}\right\}};$$

- update $\{\tilde{\boldsymbol{\psi}}_{pkj}\}$ in blocks. For $p = 1, \ldots, P$ we propose $\tilde{\boldsymbol{\psi}}'_{pkj}$, obtained from a random walk proposal and we accept it with probability

$$\min\left\{\frac{p(\boldsymbol{s} \mid \tilde{\boldsymbol{\psi}}'_{pkj}, \cdot)p(\tilde{\boldsymbol{\psi}}'_{pkj})}{p(\boldsymbol{s} \mid \tilde{\boldsymbol{\psi}}_{pkj}, \cdot)p(\tilde{\boldsymbol{\psi}}_{pkj})}, 1\right\};$$

- after updating the $\{\eta_{pkj}\}$s and $\left\{\tilde{\boldsymbol{\psi}}_{pkj}\right\}$s parameters, each vector $\tilde{\boldsymbol{\psi}}_{pkj}$ is rescaled so that $\mid \tilde{\boldsymbol{\psi}}_{pkj} \mid$ has mean 1 and the associated η_{pkj} is rescaled accordingly, so that $\boldsymbol{\alpha}^{\star}_{pkj} = \eta_{pkj}\tilde{\boldsymbol{\psi}}_{pkj}$ is unchanged:

$$\tilde{\boldsymbol{\psi}}_{pkj} \to \frac{r_k}{\sum_{\kappa=1}^{r_k} \mid \tilde{\psi}_{pk\kappa j} \mid}\tilde{\boldsymbol{\psi}}_{pkj}$$

and

$$\eta_{pkj} \to \frac{\sum_{\kappa=1}^{r_k} \mid \tilde{\psi}_{pk\kappa j} \mid}{r_k}\eta_{pkj}.$$

The rescaling is needed as $\{\tilde{\boldsymbol{\psi}}_{pkj}\}$ and $\{\eta_{pkj}\}$ are not identifiable. They could reach extreme regions of the space of the parameters without affecting the fit (e.g. the could compensate their values in the product $\boldsymbol{\alpha}^{\star}_{pkj} = \eta_{pkj}\tilde{\boldsymbol{\psi}}_{pkj}$). Rescaling avoids η_{pkj} becoming extremely large and allowing us to interpret it as a scaling factor representing the importance of the model it is associated with.

- update $\{\tilde{\tau}_{pkj}\}$ by Gibbs:

$$p(\tilde{\tau}_{pkj} \mid \eta_{pkj}, \tilde{\boldsymbol{\psi}}_{pkj}) = \text{Inverse-gamma}(a_{\tilde{\tau}} + \frac{1}{2}, b_{\tilde{\tau}} + \frac{\eta^2_{pkj}}{2\tilde{\xi}_{pkj}});$$

- update $\{\tilde{\xi}_{pkj}\}$ by Gibbs:

$$\frac{p(\tilde{\xi}_{pkj} = 1 \mid \eta_{pkj}, \tilde{\tau}_{pkj}, \tilde{\omega}_{pj})}{p(\tilde{\xi}_{pkj} = 0 \mid \eta_{pkj}, \tilde{\tau}_{pkj}, \tilde{\omega}_{pj})} = \frac{\sqrt{v_0}\tilde{\omega}_{pj}}{1 - \tilde{\omega}_{pj}} \exp\left\{\frac{(1 - v_0)\eta^2_{pkj}}{2v_0\tilde{\tau}_{pkj}}\right\};$$

- update $\{\tilde{\omega}_{pj}\}$ by Gibbs:

$$\tilde{\omega}_{pj} \mid \tilde{\xi}_{pkj} \sim \text{Beta}(a_{\tilde{\omega}} + \sum_{k=p+1}^{P} \delta_1(\tilde{\xi}_{pkj}), b_{\tilde{\omega}} + \sum_{k=p+1}^{P} \delta_{v_0}(\tilde{\xi}_{pkj}))$$

6. **Update** t_x, t_z. The threshold parameter is updated with a Metropolis-Hasting step. In order to have a proposal distribution which is both symmetric and has support only on a certain region we used a reflecting random walk. We assume $t_x \sim \text{Unif}(0, 1)$, and we propose $t'_x \sim \text{Unif}(t_x - \delta, t_x + \delta)$, where δ is some small positive value. If $t'_x < 0$ then we reassign it to be $\mid t'_x \mid$, if $t'_x > 1$ we reassign

it to be $2 - t_x^{'}$. This procedure draws samples from a symmetric proposal distribution and has support on $[0, 1]$ (Hoff, 2009). Then $t_x^{'}$ is accepted with probability

$$\min\left\{\frac{p(\boldsymbol{s} \mid t_x^{'}, \cdot)p(t_x^{'})}{p(\boldsymbol{s} \mid t_x, \cdot)p(t_x)}, 1\right\}.$$

The same applies for t_z.

7. **Update** $\{\mu_j\}$. The global intercept is updated with a Metropolis-Hastings step. For $j = 1, \ldots, J$ we propose μ_j, obtained from a random walk proposal and we accept it with probability

$$\min\left\{\frac{p(\boldsymbol{s} \mid \mu_j^{'}, \cdot)p(\mu_j^{'})}{p(\boldsymbol{s} \mid \mu_j, \cdot)p(\mu_j)}, 1\right\}.$$

8. **Update** $\{\iota_{s(i)}\}$. The random intercept is updated with a Metropolis-Hastings step. We propose $\iota_{s(i)}^{'}$, obtained from a random walk proposal and we accept it with probability

$$\min\left\{\frac{p(\boldsymbol{s} \mid \iota_{s(i)}^{'}, \cdot)p(\iota_{s(i)}^{'})}{p(\boldsymbol{s} \mid \iota_{s(i)}, \cdot)p(\iota_{s(i)})}, 1\right\}.$$

The variance of the intercept σ_ι is updated by Gibbs:

$$\iota_{s(i)} \mid \cdot \sim \text{Inverse-gamma}(a_\iota + \frac{1}{2}, b_\iota + \frac{\iota_{s(i)}^2}{2}).$$

9. **Update** $\{s_{ij}\}$. The latent variable s_{ij} for $i = 1, \ldots, n,\ j = 1, \ldots, J$ is updated by Gibbs:

$$s_{ij} \mid \cdot \sim \text{Gamma}(y_i^+ + \gamma_{ij}, (u_i + 1)^{-1}).$$

10. **Update** $\{u_i\}$. The auxiliary variable u_i for $i = 1, \ldots, n$ is updated by Gibbs:

$$u_i \mid \cdot \sim \text{Gamma}(y_i^+, T_i^{-1}).$$

A.4.3 DAG of the model

In figure A.1 is represented the DAG of the model. Note that the two different peNMIG hierarchical prior for the penalized and the unpenalized part are represented jointly, in order to reduce the complexity of the representation.

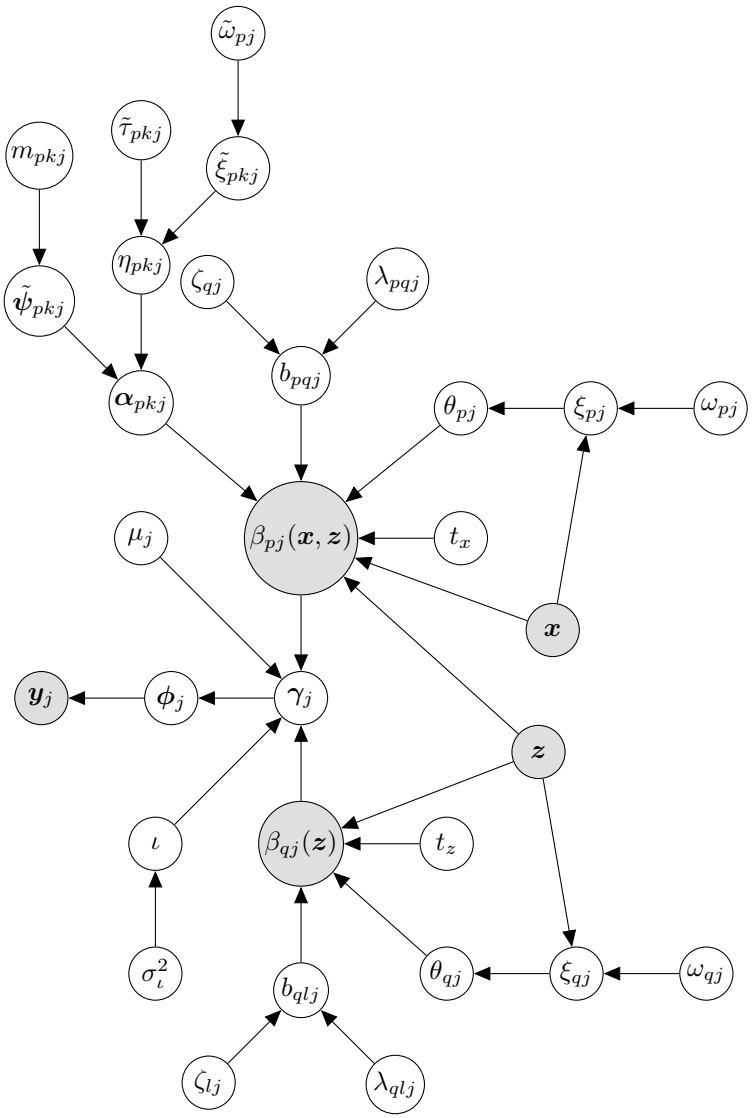


Figure A.1: Directed acyclic graph of the model. White circles are stochastic nodes. Grey circles are observed variables or deterministic nodes.

A.5 Additional Simulation Study Results

In the present Section we report additional results that are mentioned in Chapter 2. In section A.5.1 we report evaluation of our method under scenarios that account for the random intercept in the generating mechanism. In Section A.5.2 we assess the sensitivity of our method to misspecification, following the approach by Chen and Li (2013b), while in Section A.5.3 we test our method under scenarios characterized by larger values of the *taxa* considered. Finally, in Section A.5.4 we evaluate the ability of the method to capture flexible relationships between outcomes and covariates with respect to a kernel-based method proposed in Tsagris et al. (2021). In this latter study, the methods are compared regarding their predictive performance.

A.5.1 Random Intercept

In order to assess the associations' recovery in the hierarchical framework that is a peculiar feature of CRC data, SSDM's performance is evaluated on datasets whose generating mechanism includes random intercepts. These scenarios, reported in Tables A.4 and A.5, are constructed on the reference scenario discussed in Chapter 2, for varying sample sizes. The aim is to assess the performances of the proposed method in the presence of repeated measurements. As expected, for increasing sample size, MCC improves.

Table A.4: Selection of population parameters for model with random intercept. Mean across 100 replicated datasets (standard errors are in parentheses). We evaluate the performances of the proposed approach SSDM in terms of TRP, FPR and MCC.

	i) $n = 33$ j) $n = 100$ k) $n = 150$					
	TPR		FPR		MCC	
i)	0.4258	(0.0405)	0.0107	(0.0022)	0.4025	(0.0510)
j)	0.5409	(0.0377)	0.0055	(0.0023)	0.6236	(0.0353)
k)	0.7504	(0.0483)	0.0174	(0.0040)	0.6229	(0.0569)

Table A.5: Selection of subject-specific parameters for model with random intercept. Mean across 100 replicated datasets (standard errors are in parentheses). We evaluate the performances of the proposed approach SSDM in terms of TRP, FPR and MCC.

	i) $n = 33$ j) $n = 100$ k) $n = 150$					
	TPR		FPR		MCC	
i)	0.3342	(0.0335)	0.0363	(0.0084)	0.3517	(0.0397)
j)	0.3770	(0.0475)	0.0340	(0.0096)	0.4150	(0.0585)
k)	0.3871	(0.0293)	0.0284	(0.0047)	0.4249	(0.0332)

A.5.2 Model misspecification

To assess the sensitivity of our approach to model misspecification, we simulate samples using the linear growth model instead of the exponential growth model, as proposed to test misspecification in Chen and Li (2013b). In the linear growth model the proportion of taxon j for sample i is given by

$$\phi_{ij} = \frac{\gamma_j}{\sum_{j=1}^{J} \gamma_{ij}}.$$

To assess the robustness of our proposed model we generate samples not according to our model assumptions. Scenario l), reported in Tables A.6 and A.7 demonstrated that our model is quite robust to model misspecification and that the loss in MCC is moderate.

Table A.6: Selection of population parameters for misspecified model. Mean across 100 replicated datasets (standard errors are in parentheses). We evaluate the performances of the proposed approach SSDM in terms of TRP, FPR and MCC.

l) linear growth						
	TPR		FPR		MCC	
SSDM	0.3333	(0.0248)	0.0020	(0.0005)	0.5068	(0.0209)
DMBVS	0.2166	(0.0283)	0.0174	(0.0117)	0.2767	(0.0408)
regZ	0.1083	(0.0122)	0.0016	(0.0004)	0.1167	(0.0235)
penCL	0.3166	(0.0461)	0.0141	(0.0025)	0.3105	(0.0264)

Table A.7: Selection of subject-specific parameters for misspecified model. Mean across 100 replicated datasets (standard errors are in parentheses). We evaluate the performances of the proposed approach SSDM in terms of TRP, FPR and MCC.

	TPR		FPR		MCC	
l)	0.4049	(0.0335)	0.0687	(0.0096)	0.3179	(0.0311)

A.5.3 Scalability with respect to number of taxa

In order to assess the scalability of our method with respect to the number of *taxa* considered we carried out a simulation study. Following the data generating mechanism described in Section 2.5.1, but considering no random effect, we generated a dataset with $n = 100$ observations. The dataset has the following dimension: $P = 5$ continuous covariates and $Q = 5$ discrete covariates. In order to assess whether our method could adapt to more challenging scenarios, that are met when considering lower levels of the phylogenetic tree, we conducted our analysis for increasing J, in particular setting the number of categories for the response to $\{10, 50, 100, 150\}$. We replicate the estimation on each scenario 100 times. We report in Table A.8 results in terms of Matthews Correlation Coefficient, true positive rate, and false positive rate. We compared the proposed subject-specific DM approach (SSDM) with Dirichlet-multinomial Bayesian Variable Selection (DMBVS) (Wadsworth et al., 2017).

Table A.8: Model selection performances: mean across 100 replicated datasets (standard errors are in parentheses). Evaluation is carried out on population parameters only.

J	Method	TPR		FPR		MCC	
10	SSDM	0.7446	(0.0113)	0.0882	(0.0104)	0.4886	(0.2958)
	DMBVS	0.7363	(0.1169)	0.0835	(0.0891)	0.4476	(0.1580)
50	SSDM	0.7010	(0.0371)	0.0104	(0.0201)	0.5272	(0.0436)
	DMBVS	0.7536	(0.1342)	0.0327	(0.0410)	0.4972	(0.1330)
100	SSDM	0.6847	(0.0403)	0.0170	(0.0203)	0.5191	(0.0514)
	DMBVS	0.6595	(0.0938)	0.0144	(0.0311)	0.4930	(0.0555)
150	SSDM	0.5815	(0.0208)	0.0103	(0.0046)	0.3976	(0.0803)
	DMBVS	0.5613	(0.1160)	0.0267	(0.0210)	0.3844	(0.0531)

In order to assess the scalability of the method with respect to the computational time, we report detailed running times for our method letting J vary on the set $\{10, 50, 100, 150\}$. Using the R package rbenchmark (Kusnierczyk, 2012) we replicate 10 estimates for each value of J, considering a scenario with $n = 100$, $P = 5$, $Q = 5$. For each estimate we ran the MCMC for 10000 iterations, discarding the first half for burnin and then we thinned, keeping every $10-$th sampled value. The times are recorded on a PC running Ubuntu 20.04 OS with Intel Core i7-9750 2.60 GHz processor. Results are reported in Table A.9.

Table A.9: Running times performances: mean across 10 replicated runs. The column *elapsed* is the wall clock time taken to execute the function, *relative* gives the time ratio with the fastest test, *user* (CPU time) gives the CPU time spent by the current process (i.e., the current R session) and *system* (CPU time) gives the CPU time spent by the kernel (the operating system) on behalf of the current process.

J	elapsed (s)	relative	user.self (s)	sys.self (s)
10	44.81	1.00	44.80	0.01
50	126.02	2.81	126.01	0.02
100	221.37	4.94	221.12	0.19
150	328.53	7.33	327.85	0.50

A.5.4 Prediction performance

In order to evaluate the flexibility of our approach in terms of goodness-of-fit, we compared our model with a non-parametric model for compositional data. There are several example of microbiome data analysis through nonparametric and kernel methods, but they are mainly focused on a host trait prediction exploiting compositional covariates (Chen and Li, 2013a; Randolph et al., 2018). The approach proposed in Tsagris et al. (2021), through the use of the $\alpha-$transformation (Aitchison, 2003; Tsagris et al., 2011) extends the classical $k-NN$ regression to what is termed $\alpha-k-NN$ regression, yielding a highly flexible non-parametric regression model for compositional response. The $\alpha-k-NN$ regression is developed for the case in which the response data is compositional and, in contrast to other non-parametric regressions, the method allows for zero values in the data. A disadvantage of $\alpha-k-NN$

regression, and of $k-NN$ regression in general, is that it lacks the framework for classical statistical inference. This is counterbalanced by its higher predictive performance compared to parametric models.

In order to perform prediction with our proposed method, SSDM, as it is common in the Bayesian framework, we rely on posterior predictive distributions.

Following the data generating mechanism described in Section 2.5.1 we compared $\alpha-k-NN$, DMBVS and SSDM in terms of prediction performances. We designed 4 scenarios similar to those used in the simulation study to test performances on increasingly complex settings. The reference scenario in Chapter 2 is evaluated first, that is $n = 100$ samples, $J = 10$ categories for the response variable, and 10 covariates, half continuous and half binary. Strong heredity is assumed and counts are generated with a moderate level of overdispersion ($\theta_0 = 0.01$). The second scenario is obtained by increasing the number of categories ($J = 50$), while the third one is obtained considering 20 covariates. The last scenario is obtained from the reference scenario, relaxing the heredity assumption (hence assuming weak heredity) and considering large overdispersion ($\theta_0 = 0.1$) in the generating mechanism. Since $\alpha-k-NN$ is specifically designed for compositional data, the counts of each sample obtained from our generative process have been divided by the total counts in that sample, that is $u_{ij} = y_{ij}/T_i$, for $j = 1, \dots, J$ and $i = 1, \dots, n$, where $T_i = \sum_{j=1}^{J} y_{ij}$. Moreover, as $\alpha-k-NN$ can't explicitly model interaction terms, we expanded the covariate matrix adding a column for each interaction effect.

Monte Carlo simulation studies were implemented to assess the predictive performance of the three methods. For each repetition we used a 75%-25% training-validation set split.The $\alpha-k-NN$ regression requires to specify the tuning parameters (α, k); we followed the 10-fold cross-validation procedure proposed by Tsagris et al. (2021). In order to evaluate the goodness of the prediction the root mean square error (RMSE) and the average Aitchison distance (AAD) were computed. RMSE measures prediction accuracy, while AAD is an overall indicator of compositional analysis, which describes the distance between actual and predicted compositions. The equations for RMSE and AAD are defined as:

$$RMSE = \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{j=1}^{J}(y_{ij}-\hat{y}_{ij})^2},$$

$$AAD = \frac{1}{n_t}\sum_{i=1}^{n_t}\sqrt{\sum_{j=1}^{J}\left(\log\left(\frac{y_{ij}}{m_g(\boldsymbol{y}_i)}\right)-\log\left(\frac{\hat{y}_{ij}}{m_g(\hat{\boldsymbol{y}}_i)}\right)\right)^2},$$

where n_t denotes the number of samples in the training set, $\hat{y}_{ij}$ denotes the predicted value for y_{ij} and m_g denotes the geometric mean. For all examined case scenarios the results were averaged over 100 repetitions. The results are reported in Table A.10. Procedures regarding $\alpha-k-NN$ were carried out using the R package Compositional developed by Tsagris and Athineou (2021).

As expected the nonparametric method outperforms our proposed approach in all scenarios both in terms of RMSE (prediction accuracy) and AAD (compositional proximity). On the other hand, nonparametric models are hard to interpret and are not meant for selection, while our model was specifically design to detect association and account for complex pattern of interactions.

Table A.10: Prediction performances: mean across 100 replicated datasets (standard errors are in parentheses).

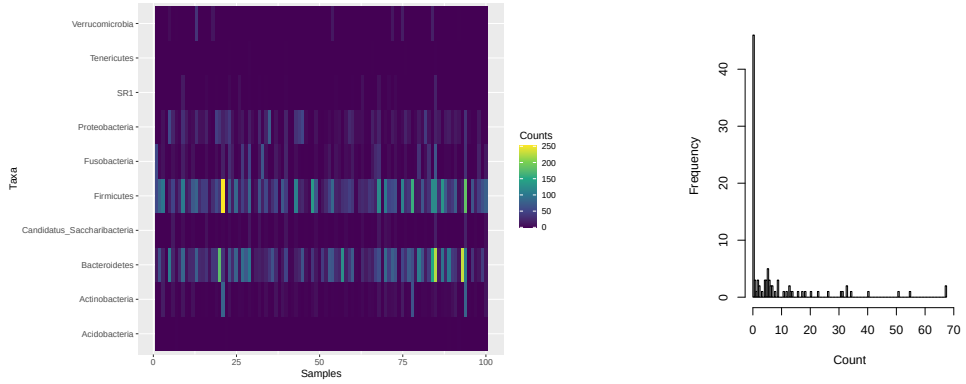
	Scenario 1			
	RMSE		AAD	
$\alpha - k - NN$	2.4558	(0.5153)	10.5998	(1.8958)
DMBVS	2.7387	(0.3693)	11.2813	(2.4317)
SSDM	3.2264	(0.5032)	11.6631	(2.0193)
	Scenario 2			
	RMSE		AAD	
$\alpha - k - NN$	0.6807	(0.1455)	36.7944	(1.5423)
DMBVS	0.8191	(0.1384)	39.7630	(1.7313)
SSDM	0.9388	(0.2249)	41.3711	(2.0698)
	Scenario 3			
	RMSE		AAD	
$\alpha - k - NN$	0.8931	(1.3531)	3.8920	(5.7756)
DMBVS	1.8932	(0.5163)	5.0323	(5.5583)
SSDM	1.1163	(1.6640)	4.2512	(6.2532)
	Scenario 4			
	RMSE		AAD	
$\alpha - k - NN$	3.2888	(0.3994)	18.9748	(1.2004)
DMBVS	4.7268	(0.4394)	23.6383	(3.7831)
SSDM	3.8275	(0.5140)	19.8972	(1.1278)

A.6 Additional results for application

In this Section we provide some more details on overdispersion and zero-inflation in the CRC dataset (Section A.6.1). Full results from the case study are reported in Section A.6.2, followed by an example of inference on personalized features. Finally we report MCMC convergence diagnostics and model checking for CRC data analysis in Section A.6.3 and A.6.4, respectively.

A.6.1 Overdispersion and zero-inflation

The dataset of our case study is characterized by large overdispersion and zero-inflation. Specifically, 47.5% of the counts were 0; Figure A.2a reports a heatmap of microbiome counts, while A.2b reports an illustrative example of zero-inflated counts from CRC data.



(a) Heatmap of CRC data.

(b) Distribution of bacterial counts for Fusobacteria taxon.

Figure A.2: Graphical Representation of Zero-Inflation in CRC data. Counts are in thousands.

In order to quantify the magnitude of overdispersion we fitted a DM model to the microbiome counts, adopting an alternative parameterization of the DM distribution (Chen and Li, 2013b). Denoting $\boldsymbol{y} = (y_1, \ldots, y_J)$ as the counts

$$f_{\boldsymbol{Y}|\boldsymbol{\phi},\theta_0}(y_1, \ldots, y_J \mid \boldsymbol{\phi}, \theta_0) = \binom{y^+}{\boldsymbol{y}} \frac{\prod_{j=1}^{J} \prod_{k=1}^{y_j} \{\phi_j(1-\theta_0) + (k-1)\theta_0\}}{\prod_{k=1}^{y^+} \{1 - \theta_0 + (k-1)\theta_0\}}, \tag{A.10}$$

where $y^+ = \sum_{j=1}^{J} y_j$, ϕ_j is the mean and θ_0 is the dispersion parameter. It is easy to see that for $\theta_0 = 0$, equation (A.10) is the multinomial distribution. Using the dirmult function in the dirmult R package (Tvedebrink, 2010), we estimated the parameter θ_0 to be equal to 0.1087 in the microbial counts in the case of study data. According to Chen and Li (2013b) and Wadsworth et al. (2017) $\theta_0 = 0.1087$ denotes a large level of overdispersion. To test our method under scenarios resembling our case of study we generate counts under settings that exhibit moderate overdispersion ($\theta_0 = 0.01$) and large overdispersion ($\theta_0 = 0.1$); we followed the same approach proposed by Chen and Li (2013b). Accordingly, these scenarios are characterized by zero inflation, which ranges from 17.2% to 45.95% in settings with low overdispersion, and ranges from 44.25% to 53.55% in settings with large overdispersion.

A.6.2 Subject-specific inference in the case study

In this Section we report all the results that are discussed in Section 2.6.1. Table A.11 lists posterior mean of main effects of covariates on microbiota relative abundances and their MPPI. Coefficients displayed in bold are included in the median probability model (Barbieri et al., 2004). Associations supported by the data and displayed in Table A.11 need to be analyzed jointly with the interaction effects in *Firmicutes* and *Bacteroidetes phyla* reported in Table A.12.

In our case study, interactions found to be significant were among binary factors. In this case inference on personalized features leads to the characterization of subgroups, determined by the combination of binary variables. If the data supported interactions among continuous covariates, it would have been possible to represent the varying coefficient as a smooth non-linear function of the covariates. Taking

as illustrative example an interaction pattern arising in the case study, we would like to make some considerations regarding subject-specific inference induced by the thresholding function.

Considering equation (2.5b), we will focus on the second term of the varying coefficient for binary covariates ($\beta_{qj}(\boldsymbol{z}_i)$), that is:

$$h\Big(\sum_{l>q} b_{qlj} z_{il}, t_Z\Big).$$

We can see that when the effects of interactions among binary factors $\{b_{qlj}\}_{l>q}$ are large, the varying coefficient tends to match the sum of the interactions. The only difference comes with the thresholding function, that pushes negligible values to an atom at zero. On the contrary the thresholding function has a more pronounced impact when interaction among binary factors $\{b_{qlj}\}_{l>q}$ are negligible. Since binary factors define subgroups in the populations, it is interesting to compare them.

We take as an example the interactions of variable *Gender* with *Stool* and *Adenocarcinoma* for the *Firmicutes taxon*, reported in Table A.12.

Table A.11: Results of CRC data analysis. Top part: posterior mean of main effects. Bottom part: MPPI of main effects. Associations included in the median model are reported in bold.

	BMI	Vegetables	Meat	Physical	Age	Mouthwash	Gender	Stool	Saliva	Adenoc.
Actinobacteria	0.10	0.06	-0.04	-0.01	0.01	-0.04	0.02	0.00	0.06	-0.09
Bacteria D	-0.00	0.02	-0.00	-0.04	0.00	0.01	-0.05	-0.19	-0.25	-0.16
Bacteroidetes	-0.00	**1.62**	**2.25**	0.61	-0.00	0.38	**0.68**	**0.77**	0.32	**1.26**
Candidatus Saccharibacteria	0.02	-0.01	-0.05	-0.03	0.06	-0.10	-0.06	-0.07	0.10	-0.36
Firmicutes	0.0	0.51	**1.41**	0.48	0.02	0.30	0.26	**0.80**	**0.94**	0.53
Fusobacteria	-0.01	-0.12	-0.12	-0.01	-0.01	-0.10	-0.05	-0.34	0.04	0.03
Proteobacteria	0.01	0.07	0.12	0.04	0.00	-0.01	0.07	-0.04	0.00	0.01
Spirochaetes	-0.01	-0.06	-0.06	-0.03	-0.13	-0.03	-0.01	-0.08	-0.06	-0.21
SR1	0.01	-0.10	-0.09	-0.04	0.08	-0.07	-0.03	-0.08	0.03	-0.12
Verrucomicrobia	0.01	-0.03	-0.08	-0.05	0.03	-0.06	-0.07	0.04	-0.16	-0.20
Residual	0.03	-0.02	-0.05	-0.03	0.02	-0.10	-0.06	0.04	-0.10	-0.19
Marginal Posterior Probability of Inclusion										
	BMI	Vegetables	Meat	Physical	Age	Mouthwash	Gender	Stool	Saliva	Adenoc.
Actinobacteria	0.24	0.30	0.31	0.24	0.10	0.20	0.23	0.21	0.24	0.25
Bacteria D	0.12	0.24	0.23	0.29	0.11	0.15	0.23	0.29	0.30	0.36
Bacteroidetes	0.06	**0.69**	**0.86**	0.41	0.18	0.27	**0.68**	**0.62**	0.33	**0.63**
Candidatus Saccharibacteria	0.18	0.27	0.25	0.25	0.18	0.27	0.24	0.27	0.29	0.43
Firmicutes	0.05	0.30	**0.66**	0.34	0.14	0.31	0.30	**0.53**	**0.64**	0.30
Fusobacteria	0.12	0.26	0.33	0.24	0.10	0.27	0.24	0.34	0.23	0.15
Proteobacteria	0.12	0.29	0.24	0.25	0.08	0.17	0.32	0.24	0.11	0.20
Spirochaetes	0.15	0.27	0.27	0.25	0.27	0.23	0.27	0.29	0.25	0.30
SR1	0.15	0.29	0.28	0.25	0.23	0.25	0.23	0.25	0.27	0.27
Verrucomicrobia	0.20	0.28	0.28	0.27	0.16	0.25	0.26	0.24	0.28	0.29
Residual	0.19	0.24	0.28	0.25	0.17	0.25	0.24	0.25	0.25	0.31

Table A.12: Posterior mean of interaction effects among discrete covariates in *Bacteroidetes* and *Firmicutes* (Marginal 90% credible set in parenthesis). The interaction terms included in the model are in bold.

	Bacteroidetes					Firmicutes				
	Mouth.	Gender	Stool	Saliva	Adenoc.	Mouth.	Gender	Stool	Saliva	Adenoc.
Mouth.	-	0.27	-0.30	-0.06	**1.40**	-	0.45	-0.06	**2.34**	-0.23
	-	(-0.39, 0.96)	(-0.95, 0.58)	(-0.46, 0.37)	**(1.04, 1.72)**	-	(-0.73, 0.77)	(-1.20, 0.65)	**(1.96, 2.56)**	(-0.79, 0.48)
Gender	-	-	0.35	-0.01	-0.43	-	-	**2.85**	0.76	0.03
	-	-	(-0.35, 0.86)	(-0.84, 0.61)	(-0.99, 0.27)	-	-	**(1.64, 3.88)**	(-0.33, 1.11)	(-0.51, 0.68)
Stool	-	-	-	-	**2.49**	-	-	-	-	0.06
	-	-	-	-	**(1.63, 3.04)**	-	-	-	-	(-0.86, 0.95)
Saliva	-	-	-	-	**0.87**	-	-	-	-	0.04
	-	-	-	-	**(0.17, 1.46)**	-	-	-	-	(-0.59, 0.70)
Adenoc.	-	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-	-

In order to assess the effect of the thresholding function, we can compare the posterior distribution of the quantities

$$(b_{23}Z_3 + b_{24}Z_4 + b_{25}Z_5)Z_2 \tag{A.11}$$

and

$$h(b_{23}Z_3 + b_{24}Z_4 + b_{25}Z_5, t_z)Z_2, \tag{A.12}$$

where the vector $\boldsymbol{Z} = (Z_1, Z_2, Z_3, Z_4, Z_5)^T$ is the vector of binary variables that gives rise to interactions and act as effect modifiers for the varying coefficient. This allows us to "isolate" the effect of the thresholding function on patients that share similar profiles. This comparison can be potentially performed for each subgroup defined by the values of the binary covariates.

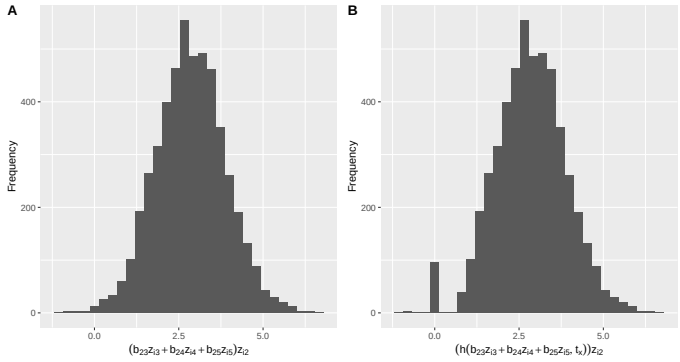
In the following, the subscript j is dropped because we are referring to the *Firmicutes taxon*. Since we are focusing on the interactions with the variable *Gender*, that is Z_2, we will only consider $\{Z_l\}_{l>2}$ and $\{b_{2l}\}_{l>2}$. The covariates *Stool*, *Saliva* and *Adenocarcinoma* are respectively labeled as Z_3, Z_4, Z_5.

In order to produce the distributions reported in Figure A.3, we draw 5000 samples from the posterior distributions of $\{b_{ql}\}$ and t_z, and obtained the distributions of the quantities in (A.11), (A.12) for different groups identified by the $n-$dimensional vectors $\{\boldsymbol{z}_l\}_{l>2}$.

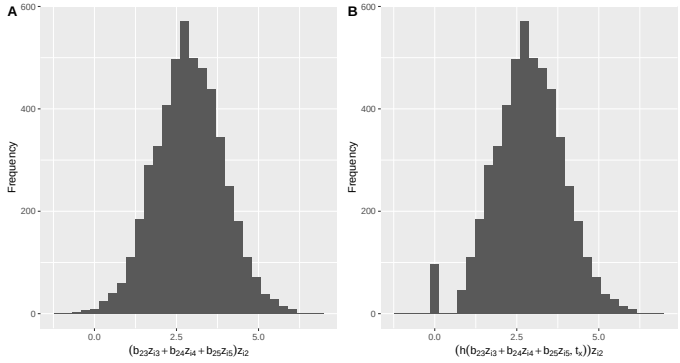
The magnitude of the interaction between *Stool* and *Gender* is large, hence it is not really affected by the thresholding mechanism. The proportion of zeros in pane B of Figures A.3a and A.3b is 1.75% and 1.92%, respectively. On the other hand the magnitude of the interaction between *Adenocarcinoma* and *Gender* is negligible and it does not affects the distribution of (A.11) and (A.12).

The magnitude of the interaction between *Saliva* and *Gender* is moderate, hence the thresholding mechanism strongly induces sparsity (the proportion of zero is 45.58% and 40.54% in pane B of Figures A.3c and A.3d, respectively). The magnitude of the interaction between *Adenocarcinoma* and *Gender* is negligible and it does not affect the mean distribution of (A.11) and (A.12), but it seems to add some variability.

Similar consideration can be conducted on the interactions with Mouthwash in both Bacteroidetes and Firmicutes taxa (see Table A.12).



(a) Distribution of interaction between gender and stool sample in patients not affected by adenocarcinoma, $\boldsymbol{z}^T = (-, -, 1, 0, 0)$. In pane A we do not consider the thresholding function, that is considered in pane B.



(b) Distribution of interaction between gender and stool sample in patients affected by adenocarcinoma, $\boldsymbol{z}^T = (-, -, 1, 0, 1)$. In pane A we do not consider the thresholding function, that is considered in pane B.

Figure A.3: Distributions of the quantities in (A.11) (panes A) and (A.12) (panes B) for different groups identified by $\boldsymbol{z}$.

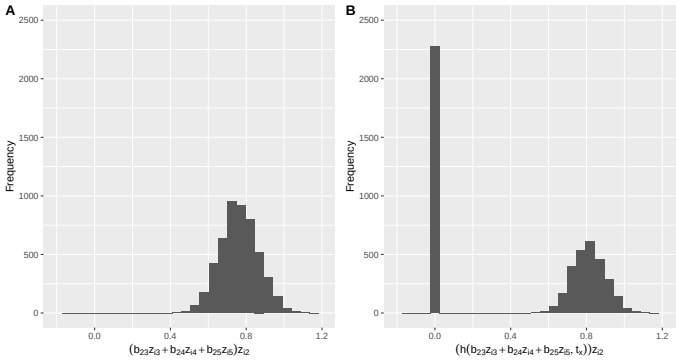
A.6.3 MCMC diagnostic

The acceptance rates for Metropolis steps are 0.2612 ($\{\theta_{pj}\}$), 0.2142 ($\{\theta_{qj}\}$), 0.4198 ($\{b_{qlj}\}$), 0.583 (t_z) and 0.634 (t_z).

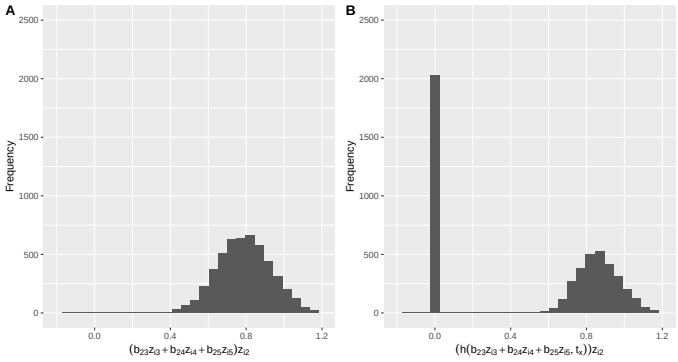
Convergence has been assessed through Gelman-Rubin potential scale reduction factor (PSRF) (Gelman et al., 1992) for continuous parameters and Pearson correlation coefficient of posterior probabilities for binary parameters. The 95% ranges of PSRF for $\{\theta_{pj}\}$, $\{\theta_{qj}\}$ and $\{b_{qlj}\}$ are $(1.0000, 1.1128)$, $(1.0000, 1.0443)$ and $(1.0000, 1.0373)$ respectively. The correlation for $\{\xi_{pj}\}$ and $\{\xi_{qj}\}$ are 0.8916 and 0.9859 respectively.

A.6.4 Model criticism

In order to assess the goodness of fit we draw samples of the posterior predictive OTUs table using the procedure described in Section A.4. The range between quantiles of level $0.025, 0.975$ for each cell of the posterior predictive OTUs tables contains the observed value for y_{ij} 84.17% of the times.



(c) Distribution of interaction between gender and saliva sample in patients not affected by adenocarcinoma, $\boldsymbol{z}^T = (-, -, 0, 1, 0)$. In pane A we do not consider the thresholding function, that is considered in pane B.



(d) Distribution of interaction between gender and saliva sample in patients affected by adenocarcinoma, $\boldsymbol{z}^T = (-, -, 0, 1, 1)$. In pane A we do not consider the thresholding function, that is considered in pane B.

Figure A.3: Distributions of the quantities in (A.11) (panes A) and (A.12) (panes B) for different groups identified by $\boldsymbol{z}$ (cont.)

Appendix B

Supplementary Material for Chapter 3

B.1 Hyperparameter Settings and Sensitivity Analysis

Our method involves several hyperparameters whose specification needs to be discussed. Moreover, to ensure a careful tuning we performed an investigation on the sensitivity of the results to these values. In particular, we constructed the sensitivity study on the reference scenario (Scenario 1) presented in Section 3.7.3, where 152 were assigned to 2 competing treatments and $K = 3$ levels of the ordinal response are assumed. Analysis is performed with a LOOCV strategy.

The parameters are set at the following default values: $\kappa = 1$, $\sigma = 0.25$, $\boldsymbol{\Lambda}_0 = diag(10, 10, 10)$, $\boldsymbol{S}_0 = diag(1.0, 1.0, 1.0)$, $v_0 = 1$. Keeping all other parameters fixed, the pairs (κ, σ) and $(\boldsymbol{\Lambda}_0, \boldsymbol{S}_0)$ and the scalar v_0 are evaluated over the following values:

- $\kappa = \{0.5, 1.0, 2.0\}$;
- $\sigma = \{0.01, 0.05, 0.25\}$;
- $\Sigma_{kk} = \{1, 10, 50\}$, for $k = 1, \ldots, K$;
- $S_{0_{kk}} = \{0.1, 1.0, 10.0\}$, for $k = 1, \ldots, K$;
- $v_0 = \{1, 2\}$.

The values for the parameters κ and σ are fixed such that, for $n = 75$ observations (that is approximately the number of patients assigned to each treatment), the prior expected number of clusters induced by the $NGGP(\kappa, \sigma)$ is reported in Table B.1.

We evaluated the model in terms of prediction, goodness-of-fit and clustering production. To assess treatment selection we used the summary measures discussed in Section 3.7.2. To evaluate the fit of the model we also report WAIC and *lpml*. WAIC is the Watanabe–Akaike information criterion, a generalized version of the Akaike information criterion. In particular, WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate (Gelman et al., 2014). *lpml* represents the log pseudo marginal likelihood, which is a goodness-of-fit metric (Christensen et al., 2011) that takes into account model

Table B.1: Prior expected number of components for $NGGP(\kappa, \sigma)$ for different specifications of σ (rows) and κ (columns).

		κ		
		0.5	1	2
σ	0.01	3.2321	5.0176	7.9814
	0.05	5.0176	5.5271	8.6342
	0.25	6.6982	9.2131	13.0813

complexity. Finally, to account for the cluster arrangement produced by the model we reported the a posteriori average number of clusters (# clu) and the variation of information (VI) (Wade and Ghahramani, 2018), presented in Section **??**. We ran the algorithm for $52,000$ iterations, with a burn-in period of $12,000$ iterations; chains were thinned and we kept every $10-$th sampled value. Analysis of each configuration was replicated and results averaged over 30 runs. Results are reported in Tables B.2, B.3, B.4, B.5, B.6 and B.7.

We found little or no sensitivity for parameters (κ, σ) and v_0. In particular, to induce a moderate number of clusters we set $(\kappa, \sigma) = (1, 0.01)$, while we set $v_0 = 1$ as default value. Results were affected to some extent by the values of $(\boldsymbol{\Lambda}_0, \boldsymbol{S}_0)$ in the multivariate normal model for the random intercept, hence we set $(\boldsymbol{\Lambda}_0, \boldsymbol{S}_0) = (10, 1.0)$, since this specification ensured (overall) the best performance.

Table B.2: Treatment selection and goodness-of-fit under different specifications of the parameters (κ, σ). Mean across 30 replicated datasets, standard deviation in parenthesis.

σ		0.01		0.05		0.25	
$\kappa = 0.5$	MOT	13.6667	(3.5266)	13.4333	(3.0590)	13.1000	(3.1552)
	$\%\Delta MTU_\ell$	0.8325	(0.0509)	0.8357	(0.0471)	0.8396	(0.0466)
	NPC	81.9333	(7.3622)	82.3667	(7.1847)	82.1667	(6.9782)
	WAIC	181.6152	(7.0690)	182.1244	(7.1719)	184.0507	(7.3738)
	$lpml$	-126.9757	(4.0635)	-127.0844	(4.0885)	-127.7218	(4.1292)
$\kappa = 1.0$	MOT	14.1000	(3.4276)	13.3667	(3.3372)	13.1000	(3.6232)
	$\%\Delta MTU_\ell$	0.8268	(0.0508)	0.8348	(0.0507)	0.8412	(0.0533)
	NPC	82.2667	(7.7010)	81.8000	(7.2130)	82.5000	(7.3614)
	WAIC	181.6095	(7.1117)	182.0948	(7.1387)	184.1644	(7.3763)
	$lpml$	-126.8929	(4.0680)	-127.0336	(4.1309)	-127.7568	(4.1364)
$\kappa = 2.0$	MOT	13.5000	(3.2137)	13.3000	(3.3026)	13.3000	(3.2393)
	$\%\Delta MTU_\ell$	0.8342	(0.0487)	0.8363	(0.0496)	0.8373	(0.0482)
	NPC	82.1333	(7.3001)	81.6000	(7.4861)	81.8333	(7.4467)
	WAIC	181.6902	(7.1881)	182.1560	(7.1690)	184.2528	(7.3638)
	$lpml$	-126.9060	(4.1041)	-127.0732	(4.0365)	-127.7572	(4.1559)

Table B.3: Cluster production under different specifications of the parameters (κ, σ). Since clustering is performed independently across treatments, results are reported separately for each treatment. Here trt 1 and trt 2 refer to Treatment 1 and Treatment 2, respectively. Mean across 30 replicated datasets, standard deviation in parenthesis.

σ		0.01		0.05		0.25	
		trt 1	trt 2	trt 1	trt 2	trt 1	trt 2
$\kappa = 0.5$	# clu	10.7184	10.0341	10.6510	9.9638	10.3284	9.6351
		(0.1848)	(0.2945)	(0.1848)	(0.2942)	(0.1832)	(0.2893)
	VI	3.6476	4.4147	3.6296	4.4325	3.6140	4.3936
		(0.5586)	(1.2150)	(0.5836)	(1.2109)	(0.5999)	(1.2416)
$\kappa = 1$	# clu	10.7198	10.0352	10.6498	9.9625	10.3222	9.6283
		(0.1870)	(0.2944)	(0.1844)	(0.2952)	(0.1813)	(0.2894)
	VI	3.6529	4.4627	3.6660	4.4482	3.6050	4.3645
		(0.5886)	(1.1913)	(0.5903)	(1.1905)	(0.6076)	(1.2391)
$\kappa = 2$	# clu	10.7141	10.0260	10.6421	9.9548	10.3108	9.6173
		(0.1863)	(0.2954)	(0.1850)	(0.2936)	(0.1826)	(0.2889)
	VI	3.6482	4.4656	3.6502	4.4300	3.6103	4.3794
		(0.5673)	(1.1897)	(0.5868)	(1.1689)	(0.6171)	(1.2243)

Table B.4: Treatment selection and goodness-of-fit under different specifications of the parameters $(\boldsymbol{\Lambda}_0, \boldsymbol{S}_0)$. $\{\Lambda_{0_{kk}}\}$ and $\{S_{0_{kk}}\}$ denote the set of K elements on the diagonal of $\boldsymbol{\Lambda}_0$ and $\boldsymbol{S}_0$, respectively. Mean across 30 replicated datasets, standard deviation in parenthesis.

$\{S_{0_{kk}}\}$		0.01		1.0		10.0	
$\{\Lambda_{0_{kk}}\} = 1$	MOT	15.0000	(3.4441)	15.0667	(3.2156)	15.2000	(3.1775)
	$\%\Delta MTU_\ell$	0.8033	(0.0577)	0.8036	(0.0530)	0.7992	(0.0533)
	NPC	80.3000	(6.8940)	80.2667	(6.8578)	80.2000	(6.7180)
	WAIC	224.3706	(6.6165)	224.2296	(6.6136)	224.0380	(6.6005)
	$lpml$	-142.3963	(3.7877)	-142.3744	(3.7829)	-142.3019	(3.7523)
$\{\Lambda_{0_{kk}}\} = 10$	MOT	13.3000	(3.2393)	13.4333	(3.1588)	14.3000	(3.1089)
	$\%\Delta MTU_\ell$	0.8373	(0.0482)	0.8360	(0.0474)	0.8203	(0.0462)
	NPC	81.8333	(7.4467)	82.1667	(7.3301)	81.9000	(7.3313)
	WAIC	184.2528	(7.3638)	183.6102	(7.3822)	181.6917	(7.2289)
	$lpml$	-127.7572	(4.1559)	-127.4659	(4.1125)	-126.7951	(4.0430)
$\{\Lambda_{0_{kk}}\} = 50$	MOT	32.2333	(21.0430)	32.4000	(20.6190)	27.7667	(14.6727)
	$\%\Delta MTU_\ell$	0.6336	(0.2455)	0.6336	(0.2395)	0.6918	(0.1651)
	NPC	73.5333	(6.4420)	74.5000	(5.6797)	74.5667	(6.5899)
	WAIC	256.8828	(10.9495)	252.8698	(11.1382)	241.6399	(10.0656)
	$lpml$	-152.9823	(4.2645)	-151.6020	(4.2887)	-147.3477	(4.1023)

Table B.5: Cluster production under different specifications of the parameters $(\boldsymbol{\Lambda}_0, \boldsymbol{S}_0)$. $\{\Lambda_{0_{kk}}\}$ and $\{S_{0_{kk}}\}$ denote the set of K elements on the diagonal of $\boldsymbol{\Lambda}_0$ and $\boldsymbol{S}_0$, respectively. Since clustering is performed independently across treatments, results are reported separately for each treatment. Here trt 1 and trt 2 refer to Treatment 1 and Treatment 2, respectively. Mean across 30 replicated datasets, standard deviation in parenthesis.

$\{S_{0_{kk}}\}$		0.01		1.0		10.0	
		trt 1	trt 2	trt 1	trt 2	trt 1	trt 2
$\{\Lambda_{0_{kk}}\} = 1$	# clu	17.4550	16.9271	17.4564	16.9277	17.4571	16.9271
		(0.2412)	(0.4232)	(0.2412)	(0.4231)	(0.2408)	(0.4237)
	VI	5.9976	9.0013	5.9818	8.9978	5.9482	8.9682
		(1.1646)	(0.8694)	(1.1616)	(0.9064)	(1.1617)	(0.9231)
$\{\Lambda_{0_{kk}}\} = 10$	# clu	10.3108	9.6173	10.3300	9.6348	10.3891	9.6963
		(0.1826)	(0.2889)	(0.1829)	(0.2897)	(0.1844)	(0.3010)
	VI	3.6103	4.3794	3.6114	4.3557	3.5818	4.3491
		(0.6171)	(1.2243)	(0.5986)	(1.1948)	(0.5661)	(1.1487)
$\{\Lambda_{0_{kk}}\} = 50$	# clu	5.3712	4.9241	5.4412	4.9834	5.6271	5.1582
		(0.0996)	(0.0922)	(0.1052)	(0.1022)	(0.1030)	(0.1168)
	VI	3.0050	3.0068	3.0252	3.0311	3.0693	3.0958
		(0.0088)	(0.0748)	(0.0234)	(0.1101)	(0.0541)	(0.1512)

Table B.6: Treatment selection and goodness-of-fit under different specifications of the parameter v_0. Mean across 30 replicated datasets, standard deviation in parenthesis.

v_0	1.0		2.0	
MOT	13.30	(3.24)	12.20	(3.46)
$\%\Delta MTU_\ell$	0.84	(0.05)	0.86	(0.05)
NPC	81.83	(7.45)	81.23	(7.57)
WAIC	184.25	(7.36)	217.00	(9.43)
$lpml$	-127.76	(4.16)	-138.74	(4.36)

Table B.7: Cluster production under different specifications of the parameter v_0. Since clustering is performed independently across treatments, results are reported separately for each treatment. Here trt 1 and trt 2 refer to Treatment 1 and Treatment 2, respectively. Mean across 30 replicated datasets, standard deviation in parenthesis.

		trt1		trt2	
$v_0 = 1.0$	# clu	10.3108	(0.1826)	9.6173	(0.2889)
	VI	3.6103	(0.6171)	4.3794	(1.2243)
$v_0 = 2.0$	# clu	6.9438	(0.1387)	6.5599	(0.1875)
	VI	3.0829	(0.1593)	2.7048	(0.5023)

B.2 Computational Details

To construct an efficient algorithm and improve computational feasibility we adopt a data augmentation approach to represent the Dirichlet distribution as independent

latent Gamma random variables (see Appendix B.2.1). This greatly facilitates the sampling procedure.

The core part of the algorithm is the update of cluster membership. The computation associated with (3.11) is based on Neal (2000)'s Algorithm 8 with Reuse (Favaro et al., 2013). The pivotal step in Algorithm 8 is to augment the state space of permanent parameters with M additional auxiliary parameters. Auxiliary parameters need to be defined such that the marginal distribution of the permanent variables after integrating out the auxiliary ones is the appropriate posterior distribution. These auxiliary parameters will be temporary and represent possible values for parameters of components that are not associated with any observation. After the update of the cluster labels, the augmentation variables can be discarded along with empty clusters, avoiding the need to perform analytical integrations that may not be available.

This algorithm is simple to implement and has showed excellent mixing speed, but each time we update the cluster label for an observation we need to sample and subsequentially discard the auxiliary variables. Noting that after updating the cluster assignment of each observation the parameters of any unused empty cluster are already independently and identically distributed, Favaro et al. (2013) propose to reuse them for the update of the next observation. We implement Algorithm 8 with Reuse, since it represents a computationally efficient strategy.

Conditional on the updated cluster labels, all the remaining parameters are easily updated with Gibbs sampler or Metropolis-Hastings steps (Section B.2.2).

B.2.1 Augmented data scheme

Generating samples from the Dirichlet distribution using independent Gamma random variable is computationally efficient. Exploiting this property our data augmentation approach is based on a reparameterization of equation (3.1) and on the introduction of an auxiliary parameter. Let $\pi_{ik}^a = d_{ik}^a / D_i^a$, where $D_i^a = \sum_{k=1}^K d_{ik}^a$ and assume that

$$d_{ik}^a \sim \text{Gamma}(\gamma_{ik}^a(\eta_{jk}^{a\star}, \boldsymbol{\beta}_k), 1).$$

Quantity $H^a(\boldsymbol{\eta}^{a\star}, \Pi^a, \boldsymbol{\beta}; \boldsymbol{y}^a, \boldsymbol{x}^a, \boldsymbol{\pi}^a)$ in (3.12) can be restated as

$$H^a(\boldsymbol{\eta}^{a\star}, \Pi^a, \boldsymbol{\beta}; \boldsymbol{y}^a, \boldsymbol{x}^a, \boldsymbol{d}^a) = \prod_{i=1}^{n_a} \frac{d_{iy_i^a}^a}{D_i^a} \prod_{j=1}^{C^a} \prod_{i \in S_j^a} \text{Gamma}(\boldsymbol{\gamma}_i^a(\boldsymbol{\eta}_j^{a\star}, \boldsymbol{\beta}_k), 1), \tag{B.1}$$

where $\boldsymbol{d}^a$ is a $n \times K$ matrix that contains d_{ik}^a elements, for $i = 1, \ldots, n^a$ and $k = 1, \ldots, K$. Moreover we introduce n auxiliary parameters and let, for $a = 1, \ldots, T$

$$u_i^a \mid D_i^a \sim \text{Gamma}(1, D_i^a).$$

From the Gamma density function we obtain that

$$\frac{1}{D_i^a} = \int_0^\infty \exp(-D_i^a u_i^a) \mathrm{d}u_i^a,$$

so from equation (B.1):

$$
\begin{aligned}
H^a(\boldsymbol{\eta}^{a\star}, \Pi^a, \boldsymbol{\beta}; \boldsymbol{y}^a, \boldsymbol{x}^a, \boldsymbol{\pi}^a) = & \prod_{i=1}^{n^a} d^a_{iy_i^a} e^{-D_i^a u_i^a} \times \prod_{j=1}^{C^a_{n^a}} \prod_{i \in S_j^a} \prod_{k=1}^{K} \frac{{d^a_{ik}}^{\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k)-1} e^{-d^a_{ik}}}{\Gamma(\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k))} \\
= & \prod_{i=1}^{n^a} d^a_{iy_i^a} e^{-u_i^a \sum_k d^a_{ik}} \times \prod_{j=1}^{C^a_{n^a}} \prod_{i \in S_j^a} \prod_{k=1}^{K} \frac{{d^a_{ik}}^{\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k)-1} e^{-d^a_{ik}}}{\Gamma(\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k))} \\
= & \prod_{i=1}^{n^a} d^a_{iy_i^a} \times \prod_{j=1}^{C^a_{n^a}} \prod_{i \in S_j^a} \prod_{k=1}^{K} \frac{{d^a_{ik}}^{\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k)-1} e^{-d^a_{ik}(u_i^a+1)}}{\Gamma(\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k))}
\end{aligned}
$$

B.2.2 MCMC sampling

For posterior inference we designed a Metropolis within Gibbs sampler. We use a generalization of Algorithm 8 by Neal (2000) proposed by Favaro et al. (2013) to update the $i-$th subject's cluster label. Weights for each component are obtained comparing the unnormalized posterior for cluster j when the subject is excluded and when it is included. Conditional on the updated cluster labels, all the remaining parameters are updated with Gibbs sampler or Metropolis-Hastings steps. The MCMC sampler goes as follows.

Π: *Algorithm 8 with Reuse.* For $a = 1, \dots, T$, $i = 1, \dots, n^a$ let $\boldsymbol{e}^a = (e^a_1, \dots, e^a_{n^a})$ be the cluster allocation vector of indexes, with $e^a_i = j$ iff $i \in S^a_j$. Let $S^{a,-i}_j$ and $C^{a,-i}_{n^a}$ denote the j-th cluster and the total number of clusters when subject i assigned to treatment a is not considered. In the same way, we use $\boldsymbol{x}_j^{a\star,-i}$ to denote the matrix of predictive determinants of the patients in cluster j, when the $i-$th patient is not included. Cluster membership for patient i, that is e^a_i is drawn using the following unnormalized probabilities:

$$
\begin{aligned}
& P(e^a_i = j | \cdot) \propto \\
& \begin{cases}
\displaystyle\prod_{k=1}^{K} \frac{{d^a_{ik}}^{\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k)-1} e^{-d^a_{ik}(u^a_i+1)}}{\Gamma(\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k))} \frac{\rho(S^{a,-i}_j \cup \{i\}) \tilde{g}(\boldsymbol{x}_j^{a\star,-i} \cup \{\boldsymbol{x}^a_i\})}{\rho(S^{a,-i}_j) \tilde{g}(\boldsymbol{x}_j^{a\star,-i})} & \text{for } j = 1, \dots, C^{a,-i}_{n^a} \\
\displaystyle\prod_{k=1}^{K} \frac{{d^a_{ik}}^{\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k)-1} e^{-d^a_{ik}(u^a_i+1)}}{\Gamma(\gamma^a_{ik}(\eta^{a\star}_{jk}, \boldsymbol{\beta}_k))} \frac{\rho(\{i\}) \tilde{g}(\{\boldsymbol{x}^a_i\})}{M} & \text{for } j = C^{a,-i} + 1, \dots, C^{a,-i}_{n^a} + M
\end{cases}
\end{aligned}
\tag{B.2}
$$

where $\{\eta^{a\star}_{jk}\}$ for $j = C^{a,-i}_{n^a} + 1, \dots, C^{a,-i}_{n^a} + M$ are M auxiliary variables (Neal, 2000), associated with M empty clusters, independently and identically distributed according to some prior distribution p^e. The first terms are the likelihoods associated with observation i given the cluster parameters, while the second terms can be interpreted as being proportional to the covariate-informed

prior probability of being assigned to the corresponding cluster (with the M empty clusters sharing the probability of creating a new cluster).

The Algorithm 8 with Reuse proposes an efficient handling of the M auxiliary parameters. For $a = 1, \dots, T$ and for $i = 1, \dots, n^a$ it updates the cluster assignment of observation i according to the following scheme:

Algorithm 1: Algorithm 8 with Reuse

1 Remove i from the cluster it belongs, so that $S_j^a \in \Pi_n^a$, $|S_j^a| = n_j^a$ becomes $S_j^{a,-1}, |S_j^a| = n_j^a - 1$
2 **if** $|S_j^a| = 0$ *(S_j^a is empty)* **then**
3 sample $m \in \{1, \dots, M\}$ uniformly at random
4 replace $\boldsymbol{\eta}^{a\star}_{C^{a,-i}_{n^a}+m}$ with $\boldsymbol{\eta}_j^{a\star}$
5 remove S_j^a from $\Pi^a_{n^a}$
6 Assign i to the clusters with probabilities $P(e_i^a|\cdot)$ as defined in equation (B.2)
7 **if** $e_i^a \in \{C^{a,-i}_{n^a} + 1, \dots, C^{a,-i}_{n^a} + M\}$ *(the observation i is assigned to an empty cluster)* **then**
8 assign it to a new cluster in Π_n^a with parameter $\boldsymbol{\eta}^{a\star}_{e_i^a}$
9 replace $\boldsymbol{\eta}^{a\star}_{e_i^a}$ with a new independent draw from p^e.

After the loop on all the observations is over and we have the updated cluster assignments, the auxiliary parameters associated with empty clusters are sampled again as independent and identically distributed according from p^e. In Algorithm 8 as proposed by Neal (2000) after line 6 of Algorithm 1 all the auxiliary parameters associated with empty clusters would have been discarded and then generated again to update the cluster label of the next observation. Note that the only difference adopting the Reuse Algorithm implies is the way the parameters of the empty clusters are managed and retained across cluster assignment updates of multiple observations (Favaro et al., 2013).

Finally particular attention should be paid when condition at line 2 of Algorithm 1 is true. In fact, to avoid gaps in the cluster labels one should proceed with a relabeling of all the clusters $\{S^a_{j'}\}_{j'>j}$ (Page and Quintana, 2015).

$\boldsymbol{\eta}^\star$: *Metropolis step.* For $a = 1, \dots, T$, $i = 1, \dots, n_a$ we sample $\boldsymbol{\eta}_1^{a\star}, \dots, \boldsymbol{\eta}_{C^a}^{a\star}$ from the following distribution:

$$P(\boldsymbol{\eta}_j^{a\star}|\cdot) \propto \prod_{i \in S_j^a} \prod_{k=1}^{K} \frac{{d_{ik}^a}^{\gamma_{ik}^a(\eta_{jk}^{a\star}, \boldsymbol{\beta}_k)}}{\Gamma(\gamma_{ik}^a(\eta_{jk}^{a\star}, \boldsymbol{\beta}_k))} p(\boldsymbol{\eta}_j^{a\star}).$$

Hyperparameters update:

$\boldsymbol{\theta}^\star$ *Gibbs step.* For $j = 1, \dots, C^a_{n^a}$ we sample $\boldsymbol{\theta}_j^{a\star}$ from its full conditional:

$$\boldsymbol{\theta}_j^{a\star}|\boldsymbol{\eta}_j^{a\star}, \boldsymbol{\sigma}_j^{a\star} \sim N_k(\boldsymbol{\mu}^{a\star}_{n_j}, \boldsymbol{\sigma}^{a\star}_{n_j}),$$

where

$$\boldsymbol{\Lambda}^{a\star}_{n_j^a} = (\boldsymbol{\Lambda}_0^{-1} + n_j^a \boldsymbol{\Sigma}_j^{-1})^{-1}$$

and

$$\boldsymbol{\mu}_{n_j^a}^{a\star} = \boldsymbol{\Lambda}_{n_j^a}^{a\star}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + n_j^a\boldsymbol{\Sigma}_j^{-1}\bar{\boldsymbol{\eta}}_j).$$

$\boldsymbol{\Sigma}^\star$ *Gibbs step.* For $j = 1, \dots, C_{n^a}^a$ we sample $\boldsymbol{\Sigma}_j^{a\star}$ from its full conditional:

$$\boldsymbol{\Sigma}_j^{a\star}|\boldsymbol{\eta}_j^{a\star}, \boldsymbol{\theta}_j^{a\star} \sim IW(\nu_0 + n_j^a, [\boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}_j^{a\star}}^a]^{-1}),$$

where

$$\boldsymbol{S}_{\boldsymbol{\theta}_j^{a\star}}^a = \sum_{i=1}^{n_j^a}(\boldsymbol{\eta}_j^{a\star} - \boldsymbol{\theta}_j^{a\star})(\boldsymbol{\eta}_j^{a\star} - \boldsymbol{\theta}_j^{a\star})^T.$$

$(\boldsymbol{\kappa}, \boldsymbol{\sigma})$: *Gibbs step.* Since the normalizing constant of equation (3.5) ca not be computed, we produce samples from a discretized approximation of posterior distribution, evaluating $p((\kappa^a, \sigma^a)|\Pi_n^a, \cdot)$ at a finite, discrete grid of possible (κ, σ) values. Moreover, we assume κ and σ to be uniformly distributed over the set of values. For $a = \dots, T$, we evaluate (3.5) at each grid point. We then obtain a discrete approximation of the log posterior distribution at each grid point and then normalize the values, obtaining weights that sum to 1 across the grid's points. Finally we sample a new value for (κ^a, σ^a) from the grid with respect to their corresponding normalized posterior probability.

$\boldsymbol{\beta}$: *Metropolis step.* We exploit the factorization of H with respect to k. For $k = 1, \dots, K$, we sample $\boldsymbol{\beta}_k$ from the following distribution:

$$P(\boldsymbol{\beta}_k|\cdot) \propto \prod_{i=1}^{n} \frac{{d_{ik}^a}^{\gamma_{ik}^a(\eta_{jk}^{a\star}, \boldsymbol{\beta}_k)}}{\Gamma(\gamma_{ik}^a(\eta_{jk}^{a\star}, \boldsymbol{\beta}_k))} p(\boldsymbol{\beta}_k).$$

Hyperparameters update:

$\boldsymbol{\lambda}$ The global scale parameters are updated through an adaptation of the slice sampling scheme given in the online supplement of Polson et al. (2014). We define $\varpi_{pk} = 1/\lambda_{pk}^2$ and $\varsigma_{pk} = \beta_{pk}/\tau_k$. This reparameterization allows us to employ slice sampler (Neal, 2003), as the conditional posterior distribution of ϖ_{pk} is

$$p(\varpi_{pk} \mid \tau_k, \varsigma_{pk}) \propto \exp\left\{ -\frac{\varsigma_{pk}^2}{2}\varpi_{pk} \right\} \frac{1}{1 + \varpi_{pk}}.$$

To sample λ_{pk}:

1. draw a sample from uniform distribution:

$$h_{pk} \mid \varpi_{pk} \sim U(0, 1/(1 + \varpi_{pk}));$$

2. draw a sample from truncated Exponential density, so that it has zero probability outside the interval $(0, (1 - u_{pk})/u_{pk})$:

$$\varpi_{pk} \mid \varsigma_{pk}, h_{pk} \sim \ Exp(2/\varsigma_{pk}^2).$$

Transforming back to the $\lambda-$scale it will ensure a sample from the conditional distribution of interest.

$\boldsymbol{\tau}$ The same applies for τ_k, replacing $\varpi = 1/\tau_k^2$ and $\varsigma_k^2 = \sum_{p=1}^{P} \beta_{pk}^2/2$.

$\boldsymbol{d}$: *Gibbs step.* For $a = 1, \ldots, T$, $i = 1, \ldots, n^a$ we sample $\boldsymbol{d}_i^a$ from:

$$\boldsymbol{d}_i^a \mid \cdot \sim \text{Gamma}(\boldsymbol{\gamma}_i^a(\boldsymbol{\eta}_j^{a\star}, \boldsymbol{\beta}) + \boldsymbol{\delta}_1(y_i^a), (u_i + 1)^{-1})$$

where $\boldsymbol{\delta}_1$ is a $K \times 1$ vector of Dirac delta in 1.

$\boldsymbol{u}$: *Gibbs step.* For $a = 1, \ldots, T$, $i = 1, \ldots, n^a$ we sample u_i^a from:

$$u_i^a \mid \cdot \sim \text{Gamma}(1, D_i^{-1}).$$

Bibliography

J. Aitchison. *The statistical analysis of compositional data*. Caldwell, N.J.: Blackburn Press, 2003. A.5.4

R. Argiento, A. Guglielmi, and A. Pievatolo. A comparison of nonparametric priors in hierarchical mixture modelling for aft regression. *Journal of Statistical Planning and Inference*, 139(12):3989–4005, 2009. 3.3.2

R. Argiento, A. Guglielmi, and A. Pievatolo. Bayesian density estimation and model selection using nonparametric hierarchical mixtures. *Computational Statistics & Data Analysis*, 54(4):816–832, 2010. 3.1, 3.3.2, 3.3.2

R. Argiento, I. Bianchini, and A. Guglielmi. Posterior sampling from \ε-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics*, 10(2):3516–3547, 2016. 3.5

M. M. Barbieri, J. O. Berger, et al. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004. 2.5.3, A.6.2

D. Barry and J. A. Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993. 1.2.3

P. L. Bedard, A. R. Hansen, M. J. Ratain, and L. L. Siu. Tumour heterogeneity in the clinic. *Nature*, 501(7467):355–364, 2013. 3.1

R. Bescos, A. Ashworth, C. Cutler, Z. L. Brookes, L. Belfield, A. Rodiles, P. Casas-Agustench, G. Farnham, L. Liddle, M. Burleigh, et al. Effects of chlorhexidine mouthwash on the oral microbiome. *Scientific Reports*, 10(1):1–8, 2020. 2.6.2

R. A. Betensky, D. N. Louis, and J. Gregory Cairncross. Influence of unrecognized molecular heterogeneity on randomized clinical trials. *Journal of Clinical Oncology*, 20(10):2495–2499, 2002. 1.1

D. M. Blei and P. I. Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(8), 2011. 1.2.3

M. Bonetti and R. D. Gelber. A graphical method to assess treatment–covariate interactions using the cox model on subsets of the data. *Statistics in Medicine*, 19 (19):2595–2609, 2000. 3.1

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232, 2011. 3.1

A. Brezger and S. Lang. Generalized structured additive regression based on bayesian p-splines. *Computational Statistics & Data Analysis*, 50(4):967–991, 2006. 1.2.2

A. Brix. Generalized gamma measures and shot-noise cox processes. *Advances in Applied Probability*, 31(4):929–953, 1999. 3.3.2

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR, 2009. 2.3.2, 3.4, A.2.2

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010. 2.3.2, 3.4, A.2.2

J. Chen and H. Li. Kernel methods for regression analysis of microbiome compositional data. In *Topics in Applied Statistics*, pages 191–201. Springer, 2013a. A.5.4

J. Chen and H. Li. Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7 (1), 2013b. 2.1, 2.2.1, 2.2.2, 2.2.2, 2.5, 2.5.1, 2.5.3, A.5, A.5.2, A.6.1, A.6.1

H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996. 2.3.2

H. Chipman, E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001. 2.5.2, A.2.1

R. Christensen, W. Johnson, A. Branscum, and T. E. Hanson. *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC press, 2011. B.1

E. B. Claus, K. M. Walsh, J. K. Wiencke, A. M. Molinaro, J. L. Wiemels, J. M. Schildkraut, M. L. Bondy, M. Berger, R. Jenkins, and M. Wrensch. Survival and low-grade glioma: the emergence of genetic information. *Neurosurgical Focus*, 38 (1):E6, 2015. 3.8.1

W. S. Cleveland, E. Grosse, and W. M. Shyu. Local regression models. In *Statistical models in S*, pages 309–376. Routledge, 1991. 1.2.1

A. L. Cohen and H. Colman. Glioma biology and molecular markers. *Current Understanding and Treatment of Gliomas*, pages 15–30, 2015. 3.8.1

D. B. Dahl, R. Day, and J. W. Tsai. Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112(518): 721–732, 2017. 1.2.3

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979. A.1.1

P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero. Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229, 2013. 3.1

J. S. De Bono and A. Ashworth. Translating cancer research into targeted therapeutics. *Nature*, 467(7315):543–549, 2010. 1.1, 2.1

P. H. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121, 1996. 1.2.2, 1.2.2

J. J. Faith, J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. L. Goodman, J. C. Clemente, R. Knight, A. C. Heath, R. L. Leibel, et al. The long-term stability of the human gut microbiota. *Science*, 341(6141), 2013. 2.7

J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179, 2008. 1.2.1

S. Favaro, Y. W. Teh, et al. Mcmc for normalized random measure mixture models. *Statistical Science*, 28(3):335–359, 2013. 3.1, 3.5, B.2, B.2.2, 9

A. E. Gelfand and S. K. Sahu. Identifiability, improper priors, and gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94 (445):247–253, 1999. A.1.1, A.1.1

A. Gelman, D. B. Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. A.6.3

A. Gelman, A. Jakulin, M. G. Pittau, Y.-S. Su, et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008. 2.3.2

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014. B.1

A. Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006. 2.3.4, A.2.6

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. 2.3.2

E. I. George and R. E. McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997. 2.3.1

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999. 3.7.1

M. L. Goodenberger and R. B. Jenkins. Genetics of adult glioma. *Cancer Genetics*, 205(12):613–621, 2012. 3.8.1

J. Griffin, P. Brown, et al. Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, 12(1):135–159, 2017. 2.3.2

P. Gustafson. Bayesian regression modeling with interactions and smooth effects. *Journal of the American Statistical Association*, 95(451):795–806, 2000. 2.3.2

H. Haario, E. Saksman, and J. Tamminen. Componentwise adaptation for high dimensional mcmc. *Computational Statistics*, 20(2):265–273, 2005. 1

C. Haro, O. A. Rangel-Zuniga, J. F. Alcalà-Diaz, F. Gòmez-Delgado, P. Pèrez-Martinez, J. Delgado-Lista, G. M. Quintana-Navarro, B. B. Landa, J. A. Navas-Cortès, M. Tena-Sempere, et al. Intestinal microbiota is influenced by gender and body mass index. *PloS one*, 11(5):e0154090, 2016. 2.6.2

J. G. Harrison, W. J. Calder, V. Shastry, and C. A. Buerkle. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular Ecology Resources*, 20(2):481–497, 2020. 2.2.2

J. A. Hartigan. Partition models. *Communications in Statistics-Theory and methods*, 19(8):2745–2756, 1990. 1.2.3, 3.3, 3.3.1, 3.3.1

T. Hastie and R. Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3): 297 – 310, 1986. doi: 10.1214/ss/1177013604. URL https://doi.org/10.1214/ss/1177013604. 1.2.2

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. 1.2.1, 2.2.2

D. J. Hentges, B. R. Maier, G. C. Burton, M. A. Flynn, and R. K. Tsutakawa. Effect of a high-beef diet on the fecal bacterial flora of humans. *Cancer Research*, 37(2): 568–571, 1977. 2.6.2

P. D. Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009. 6

M. Huang, A. Shen, J. Ding, and M. Geng. Molecularly targeted cancer therapy: some lessons from the past decade. *Trends in Pharmacological Sciences*, 35(1): 41–50, 2014. 1.1

T. Ius, Y. Ciani, M. E. Ruaro, M. Isola, M. Sorrentino, M. Bulfoni, V. Candotti, C. Correcig, E. Bourkoula, I. Manini, et al. An nf-κb signature predicts low-grade glioma prognosis: A precision medicine approach based on patient-derived stem cells. *Neuro-oncology*, 20(6):776–787, 2018. 3.8.1

S. Jahani-Sherafat, M. Alebouyeh, S. Moghim, H. A. Amoli, and H. Ghasemian-Safaei. Role of gut microbiota in the pathogenesis of colorectal cancer; a review article. *Gastroenterology and Hepatology from bed to bench*, 11(2):101, 2018. 2.6.2

T. P. Jenkins, F. Formenti, C. Castro, C. Piubelli, F. Perandin, D. Buonfrate, D. Otranto, J. L. Griffin, L. Krause, Z. Bisoffi, et al. A comprehensive analysis of the faecal microbiome and metabolome of strongyloides stercoralis infected volunteers from a non-endemic area. *Scientific Reports*, 8(1):1–13, 2018. 2.6.2

C. Kang, H. Janes, and Y. Huang. Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, 70(3):695–707, 2014. 3.1, 3.8.3

X. Ke and L. Shen. Molecular targeted therapy of cancer: The progress and future prospect. *Frontiers in Laboratory Medicine*, 1(2):69–75, 2017. 1.1

J. F. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(1):1–15, 1975. 3.3.2

M. D. Koslovsky, K. L. Hoffman, C. R. Daniel, and M. Vannucci. A bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *The Annals of Applied Statistics*, 14(3):1471–1492, 09 2020. doi: 10.1214/20-AOAS1354. URL https://doi.org/10.1214/20-AOAS1354. 2.1, 2.2.2, 2.4, A.4

M. R. Kosorok and E. B. Laber. Precision medicine. *Annual Review of Statistics and its Application*, 6:263–286, 2019. 1.1, 2.7, 3.1

N. Krämer, J. Schäfer, and A.-L. Boulesteix. Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10(1):1–24, 2009. 3.1

W. Kusnierczyk. *rbenchmark: Benchmarking routine for R*, 2012. URL https://CRAN.R-project.org/package=rbenchmark. R package version 1.0.0. A.5.3

P. S. La Rosa, J. P. Brooks, E. Deych, E. L. Boone, D. J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W. D. Shannon. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one*, 7(12):e52078, 2012. 2.1

S. W. Lagakos et al. The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine*, 354(16):1667, 2006. 3.1

S. Lang and A. Brezger. Bayesian p-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004. 1.2.2

H. Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015. 2.1

A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-gaussian priors. *Journal of the American Statistical Association*, 100(472): 1278–1291, 2005. 3.3.2

A. Lijoi, R. H. Mena, and I. Prünster. Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007. 3.1, 3.3.2, 3.3.2, 3.8.4, 3.9

W. Lin, P. Shi, R. Feng, and H. Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014. 2.1

W. Lu, H. H. Zhang, and D. Zeng. Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22(5):493–504, 2013. 3.1

J. Ma, B. P. Hobbs, and F. C. Stingo. Statistical methods for establishing personalized treatment rules in oncology. *BioMed Research International*, 2015, 2015. 3.1

J. Ma, F. C. Stingo, and B. P. Hobbs. Bayesian predictive modeling for genomic based personalized treatment selection. *Biometrics*, 72(2):575–583, 2016. 3.1, 3.6, 3.7.1, 3.7.1, 3.7.2, 3.8.2, 3.8.3

J. Ma, B. P. Hobbs, and F. C. Stingo. Integrating genomic signatures for treatment selection with bayesian predictive failure time models. *Statistical Methods in Medical Research*, 27(7):2093–2113, 2018. 3.1

J. Ma, F. C. Stingo, and B. P. Hobbs. Bayesian personalized treatment selection strategies that integrate predictive with prognostic determinants. *Biometrical Journal*, 61(4):902–917, 2019. 3.1, 3.7, 3.7.1, 3.7.1, 3.7.3, 3.8.2, 3.8.3, 3.8.4

J. Mandrioli, A. Amedei, G. Cammarota, E. Niccolai, E. Zucchi, R. D'Amico, F. Ricci, G. Quaranta, T. Spanu, and L. Masucci. Fetr-als study protocol: a randomized clinical trial of fecal microbiota transplantation in amyotrophic lateral sclerosis. *Frontiers in Neurology*, 10:1021, 2019. 2.1

J. Mao and L. Ma. Dirichlet-tree multinomial mixtures for clustering microbiome compositions. *arXiv preprint arXiv:2008.00400*, 2020. 2.1

I. Martin, H.-W. Uh, T. Supali, M. Mitreva, and J. J. Houwing-Duistermaat. The mixed model for the analysis of a repeated-measurement multivariate count data. *Statistics in Medicine*, 38(12):2248–2268, 2019. 2.1, 2.3.4

M. Marty, F. Cognetti, D. Maraninchi, R. Snyder, L. Mauriac, M. Tubiana-Hulin, S. Chan, D. Grimes, A. Antón, A. Lluch, et al. Randomized phase ii trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2–positive metastatic breast cancer administered as first-line treatment: the m77001 study group. *Journal of Clinical Oncology*, 23 (19):4265–4274, 2005. 3.1

K. Matsuoka and T. Kanai. The gut microbiota and inflammatory bowel disease. In *Seminars in Immunopathology*, volume 37, pages 47–55. Springer, 2015. 2.1

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. 2.3.1

S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1):91–118, 2003. 3.7

J. E. Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82, 1962. 2.2.1

P. Müller, F. Quintana, and G. L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, 2011. 1.2.3, 3.1, 3.3, 3.3.3, 3.3.3, 3.6.1

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. 3.5, B.2, B.2.2, B.2.2, 9

R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003. 2, 3.5, 3, 9

Y. Ni, F. C. Stingo, and V. Baladandayuthapani. Bayesian graphical regression. *Journal of the American Statistical Association*, 114(525):184–197, 2019a. 2.2.2, 2.2.2, 3, 2.7, A.2.3, A.2.5

Y. Ni, F. C. Stingo, M. J. Ha, R. Akbani, and V. Baladandayuthapani. Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. *Journal of the American Statistical Association*, 114(525):48–60, 2019b. 2.2.2

E. Niccolai, E. Russo, S. Baldi, F. Ricci, G. Nannini, M. Pedone, F. C. Stingo, A. Taddei, M. N. Ringressi, P. Bechi, et al. Significant and conflicting correlation of il-9 with prevotella and bacteroides in human colorectal cancer. *Frontiers in Immunology*, 11, 2020. 1.1, 2.1

A. Olar and E. P. Sulman. Molecular markers in low-grade glioma—toward tumor reclassification. In *Seminars in Radiation Oncology*, volume 25, pages 155–163. Elsevier, 2015. 3.8.1

S. Paganin, A. H. Herring, A. F. Olshan, and D. B. Dunson. Centered partition processes: Informative priors for clustering (with discussion). *Bayesian Analysis*, 16(1):301–370, 2021. 1.2.3

G. L. Page and F. A. Quintana. Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates. *Bayesian Analysis*, 10(2): 379–410, 2015. 3.6.1, 9

G. L. Page and F. A. Quintana. Calibrating covariate informed product partition models. *Statistics and Computing*, 28(5):1009–1031, 2018. 1.2.3, 3.3.3, 3.3.3, 2

J.-H. Park and D. B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, pages 1203–1226, 2010. 1.2.3

J. Pearl. Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3): 370–389, 2017. 1.1

J. Pitman. Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996. 3.3.1

J. Pitman. Poisson-kingman partitions. *Lecture Notes-Monograph Series*, pages 1–34, 2003. 3.3.2

S. J. Pocock, S. E. Assmann, L. E. Enos, and L. E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in Medicine*, 21(19):2917–2930, 2002. 3.1

N. G. Polson, J. G. Scott, and J. Windle. The bayesian bridge. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 713–733, 2014. 2, 3.5, 3, 9

F. A. Quintana and P. L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 557–574, 2003. 3.3.1

F. A. Quintana, P. Müller, and A. L. Papoila. Cluster-specific variable selection for product partition models. *Scandinavian Journal of Statistics*, 42(4):1065–1077, 2015. 3.3.3

T. W. Randolph, S. Zhao, W. Copeland, M. Hullar, and A. Shojaie. Kernel-penalized regression for analysis of microbiome data. *The Annals of Applied Statistics*, 12 (1):540, 2018. A.5.4

B. Ren, S. Bacallado, S. Favaro, T. Vatanen, C. Huttenhower, L. Trippa, et al. Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis. *Annals of Applied Statistics*, 14(1):494–517, 2020. 2.1

G. O. Roberts and J. S. Rosenthal. Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009. 1

D. Rossell and D. Telesca. Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265, 2017. 2.1

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003. 2.2.2

E. Russo, G. Bacci, C. Chiellini, C. Fagorzi, E. Niccolai, A. Taddei, F. Ricci, M. N. Ringressi, R. Borrelli, F. Melli, et al. Preliminary comparison of oral and intestinal human microbiota in patients with colorectal cancer: a pilot study. *Frontiers in Microbiology*, 8:2699, 2018. 2.1

Y. Sanz, M. Olivares, Á. Moya-Pérez, and C. Agostoni. Understanding the role of gut microbiome in metabolic disease risk. *Pediatric Research*, 77(1-2):236–244, 2015. 2.1

T. Savitsky, M. Vannucci, and N. Sha. Variable selection for nonparametric gaussian process priors: Models and computational strategies. *Statistical Science*, 26(1): 130, 2011. 1, 1

F. Scheipl, L. Fahrmeir, and T. Kneib. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532, 2012. 1.2.2, 2.2.2, 2.2.2, 2.3.2, 3, 2.5.2, A.2.3, A.2.4

J. G. Scott and J. O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619, 2010. 2.3.1

K. Shuler, M. Sison-Mangus, J. Lee, et al. Bayesian sparse multivariate regression with asymmetric nonlocal priors for microbiome data analysis. *Bayesian Analysis*, 2018. 2.1

R. Simon. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized Medicine*, 7(1):33–47, 2010. 3.1

D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, et al. Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2. *New England Journal of Medicine*, 344(11):783–792, 2001. 3.1

X. Song and M. S. Pepe. Evaluating markers for selecting a patient's treatment. *Biometrics*, 60(4):874–883, 2004. 3.1, 3.8.3

F. C. Stingo, Y. A. Chen, M. G. Tadesse, and M. Vannucci. Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5(3), 2011. 2.4

H. Taghizadeh, L. Müllauer, J. Furtner, J. Hainfellner, C. Marosi, M. Preusser, and G. Prager. Applied precision cancer medicine in neuro-oncology. *Scientific Reports*, 9(1):1–8, 2019. 3.8.1

N. Tai, F. S. Wong, and L. Wen. The role of gut microbiota in the development of type 1, type 2 diabetes mellitus and obesity. *Reviews in Endocrine and Metabolic Disorders*, 16(1):55–65, 2015. 2.1

X. Tan, M. P. Shiyko, R. Li, Y. Li, and L. Dierker. A time-varying effect model for intensive longitudinal data. *Psychological Methods*, 17(1):61, 2012. 1.2.1

Y. Tang, L. Ma, D. L. Nicolae, et al. A phylogenetic scan test on a dirichlet-tree multinomial model for microbiome data. *The Annals of Applied Statistics*, 12(1): 1–26, 2018. 2.1

M. Tsagris and G. Athineou. *Compositional: Compositional Data Analysis*, 2021. URL https://CRAN.R-project.org/package=Compositional. R package version 4.8. A.5.4

M. Tsagris, A. Alenazi, and C. Stewart. Non-parametric regression models for compositional data, 2021. 2.5.3, A.5, A.5.4

M. T. Tsagris, S. Preston, and A. T. Wood. A data-based power transformation for compositional data. *arXiv preprint arXiv:1106.1451*, 2011. A.5.4

P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007. 2.1

T. Tvedebrink. Overdispersion in allelic counts and theta-correction in forensic genetics. *Theoretical Population Biology*, 78(3):200–210, 2010. URL http://dx.doi.org/10.1016/j.tpb.2010.07.002. A.6.1

S. van der Pas, B. Szabó, A. van der Vaart, et al. Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274, 2017. 2.4, A.2.2

S. M. Vieira, O. E. Pagovich, and M. A. Kriegel. Diet, microbiota and autoimmune diseases. *Lupus*, 23(6):518–526, 2014. 2.1

S. Wade and Z. Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018. B.1

W. D. Wadsworth, R. Argiento, M. Guindani, J. Galloway-Pena, S. A. Shelburne, and M. Vannucci. An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics*, 18(1):1–12, 2017. 2.1, 2.2.1, 2.2.2, 2.4, 2.5, 2.5.2, A.2.1, A.4, A.5.3, A.6.1

M. Weiler and W. Wick. Molecular predictors of outcome in low-grade glioma. *Current Opinion in Neurology*, 25(6):767–773, 2012. 3.8.1

W. Werft, A. Benner, and A. Kopp-Schneider. On the identification of predictive biomarkers: Detecting treatment-by-gene interaction in high-dimensional data. *Computational Statistics & Data Analysis*, 56(5):1275–1286, 2012. 3.1

M. West and J. Harrison. Bayesian forecasting and dynamic models. 1989. 1.2.1

K. White, K. Connor, J. Clerkin, B. Murphy, M. Salvucci, A. O'Farrell, M. Rehm, D. O'Brien, J. Prehn, S. Niclou, et al. New hints towards a precision medicine strategy for idh wild-type glioblastoma. *Annals of Oncology*, 31(12):1679–1692, 2020. 3.8.1

Y. Xie and B. P. Carlin. Measures of bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10):3458–3477, 2006. A.1.1

Z. Xun, Q. Zhang, T. Xu, N. Chen, and F. Chen. Dysbiosis and ecotypes of the salivary microbiome associated with inflammatory bowel diseases and the assistance in diagnosis of diseases using oral bacterial profiles. *Frontiers in Microbiology*, 9: 1136, 2018. 2.6.2

G. Zanella and G. Roberts. Multilevel linear models, gibbs samplers and multigrid decompositions. *Bayesian Analysis*, 2020. A.1.1

B. Zhang, A. A. Tsiatis, E. B. Laber, and M. Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012. 3.1

H. Zhang and L. Sun. When human cells meet bacteria: precision medicine for cancers using the microbiota. *American Journal of Cancer Research*, 8(7):1157, 2018. 2.1

L. Zhang, Y. Shi, R. R. Jenq, K.-A. Do, and C. B. Peterson. Bayesian compositional regression with structured priors for microbiome feature selection. *Biometrics*, 2020. 2.1

Y. Zhang, H. Zhou, J. Zhou, and W. Sun. Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, 26(1):1–13, 2017. 2.1, 2.2.1, 2.5

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012. 3.1

Y. Zhu, X. Lin, F. Zhao, X. Shi, H. Li, Y. Li, W. Zhu, X. Xu, C. Li, and G. Zhou. Meat, dairy and plant proteins alter bacterial composition of rat gut bacteria. *Scientific Reports*, 5(1):1–14, 2015. 2.6.2

Y. Zhu, X. Lin, H. Li, Y. Li, X. Shi, F. Zhao, X. Xu, C. Li, and G. Zhou. Intake of meat proteins substantially increased the relative abundance of genus lactobacillus in rat feces. *PloS one*, 11(4):e0152678, 2016. 2.6.2