Bayesian Compositional Regression with Microbiome Features via Variational Inference

DARREN A. V. SCOTT (Darren.Scott@lshtm.ac.uk)
London School of Hygiene & Tropical Medicine

ERNEST DIEZ BENAVENTE
University Medical Center Utrecht, Utrecht University

JULIAN LIBISELLER-EGGER
London School of Hygiene & Tropical Medicine

DMITRY FEDEROV
Research Institute of Physical Chemical Medicine

JODY PHELAN
London School of Hygiene & Tropical Medicine

ELENA ILINA
Research Institute of Physical Chemical Medicine

POLINA TIKHONOVA
Huck Institutes of Life Sciences, Pennsylvania State University

ALEXANDER KUDRYAVSTEV
Northern State Medical University

JULIA GALEEVA
Research Institute of Physical Chemical Medicine

TAANE CLARK
London School of Hygiene & Tropical Medicine

ALEX LEWIN
London School of Hygiene & Tropical Medicine

Research Article

Keywords: Compositional, variational inference, microbiome, singular multivariate normal, Markov chain Monte Carlo

Posted Date: March 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1401242/v1

Bayesian compositional regression with microbiome features via variational inference

DARREN A. V. SCOTT*[1], ERNEST DIEZ BENAVENTE[2], JULIAN LIBISELLER-EGGER[1], DMITRY FEDEROV[3], JODY PHELAN[1], ELENA ILINA[3], POLINA TIKHONOVA[3,4], ALEXANDER KUDRYAVSTEV[5], JULIA GALEEVA[3], TAANE CLARK[1], and ALEX LEWIN[1]

[1]*Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, UK, WC1E 7HT, Darren.scott@lshtm.ac.uk*
[2]*Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands*
[3]*Federal Research and Clinical Center of Physical-Chemical Medicine, Medicine, Moscow, Russia*
[4]*Bioinformatics and Genomics Intercollege Graduate Program, Huck Institutes of Life Sciences, Pennsylvania State University, USA*
[5]*Northern State Medical University, Arkhangelsk, Russia*

1 Abstract

The microbiome plays a key role in the health of the human body. Interest often lies in finding features of the microbiome, alongside other covariates, which are associated with a phenotype of interest. One important property of microbiome data, which is often overlooked, is its compositionality as it can only provide information about the relative abundance of its constituting components. Typically, these proportions vary by several orders of magnitude in datasets of large dimensions. To address these challenges we develop a Bayesian hierarchical linear log-contrast model which is estimated by mean field Monte-Carlo co-ordinate ascent variational inference (CAVI-MC). We use novel priors which account for the large differences in scale and constrained parameter space associated with the compositional covariates. A reversible jump Monte Carlo Markov chain guided by the data through univariate approximations of the variational posterior probability of inclusion, with proposal parameters informed by approximating variational densities via auxiliary parameters, is used to estimate intractable marginal expectations. We demonstrate that our proposed method outperforms standard methods of variable selection applied to compositional data. We then apply the CAVI-MC to the analysis of real data exploring the relationship of the gut microbiome to body mass index.

Key words: Compositional, variational inference, microbiome, singular multivariate normal, Markov chain Monte Carlo.

2 Introduction

The human microbiome is the combined genome of the microorganisms that live in the human body. It has been estimated that these microbes make up to 10 trillion cells, equivalent to the number of human cells (Sender et al., 2016). Advances in genome sequencing technologies has enabled scientists to study these microbes and their function and to research microbiome–host interactions both in health and disease. The decreasing cost and increasing accessibility of nucleotide sequencing means it is the primary tool used to study the microbiome (Franzosa et al., 2015). Any microbiome dataset is compositional (Gloor et al., 2017) as the magnitude of a single operational taxonomic unit (OTU) depends on the sum of all the OTUs counts, and only provides information about the relative magnitudes of the compositional components. This means that the standard methods of analysis such as linear regression are not applicable to microbiome data (Li, 2015), unless a transformation is performed.

The large dimensions of these datasets often present a problem in variable selection where the number of covariates p exceeds the number of observations n ($p >> n$) and the space of possible combinations of significant variables is large, imposing a high computational burden. Sparse variable selection of the p covariates is expected, where just a few microbes are associated with the response. Bayesian variable selection approaches have the advantage of being able to include prior knowledge and simultaneously incorporate many sources of variation. Shrinkage priors encourage the majority of regression coefficients to be shrunk to very small values when an estimator is applied identifying associations (Park and Casella, 2008). Alternatively, introducing latent variables produces posterior distributions of model inclusion and parameter values which enable model choice and a probabilistic understanding of the strength and nature of the association (Guan and Stephens, 2011). The different approaches within explicit variable selection are characterised by the location of the latent variable and its relationship with the covariates (George and McCulloch, 1993, Kuo and Mallick, 1998, Dellaportas et al., 2002).

To model compositional data, a transformation must be performed to transfer the compositional vectors into Euclidean space. Various log ratio transformations have been proposed including additive log-ratio (alr), centred log-ratio (clr) (Aitchison, 1982) and more recently isometric log-ratio (ilr) (Egozcue et al., 2003). The ilr transformation defines balances proportional to the log difference between two groups which are scale invariant. Only the first coordinate can be interpreted as it represents all the relevant information about the compositional part.

The alr transformation, which constrains the associated parameter space to sum to 0, has proved to be useful in frequentist regression problems (Aitchison and Bacon-Shone, 1984), allowing a direct inference between selected covariates and the compositional data set. Lin et al., 2014 propose an adaptive l_1 regularisation regression for sparsity with the constraint imposed by the log contrasts. This has been extended to multiple linear constraints for sub-compositional coherence across predefined groups of predictors (Shi et al., 2016). A general approach to convex optimisation, where the model has been extended to the high-dimensional setting via regularization has recently been proposed by Combettes and Müller, 2021. In the Bayesian framework Zhang et al., 2020 introduce a generalised transformation matrix on the parameters rather than the covariates, as a function of a tuning parameter c, similar to the generalized lasso. This ensures parameter estimates remain in the p space and as c reaches infinity the sum to zero constraint is imposed. By incorporating the matrix into conjugate prior and avoiding any singular distributions by not strictly imposing the zero sum constraint, a Gibbs sampler for the marginal posterior of the selection parameter can be derived. Alternative Bayesian approaches treat the the microbiome predictors as random, parameterised by a multivariate count model. Koslovsky et al., 2020 combine this with the ilr transformation in a predictive model which identifies correlations across the microbiome. Li et al., 2019 cluster on a categorical covariate via a Gaussian mixture model in an ANOVA type model, but both approaches do not allow a direct inference between the compositional predictors and the response.

The abundances of features in microbiome data often differ by orders of magnitude. As far as we know this has not been explicitly accounted for in the current literature. In the Bayesian lasso (Park and Casella, 2008) separate scale parameters can have a hierarchical prior placed on them rather than this component being marginalised over which results in the Laplace prior. In the regularisation case, the choice of hyperprior defines how the parameters are shrunk to zero. This model is easily extended to the adaptive lasso (Leng et al., 2014) by positing independent exponential priors on each scale parameter, and then augmenting each tuning parameter with additional hyperpriors.

Typically, model selection is performed using Markov chain Monte Carlo (MCMC) methods. Various stochastic search based methods have been used to explore the model space in a computationally efficient manner (Lamnisos et al., 2013, Nott and Kohn, 2005, Dellaportas et al., 2002). Despite this body of work, MCMC can still be considered too slow in practice for sufficiently large scale problems. Variational inference is an alternative technique which uses optimisation to achieve computational savings by approximating the marginal posterior densities. Its success in machine learning problems has led to concerted efforts in the literature to encourage its use by statisticians (Blei et al., 2017, Ormerod and Wand, 2010). The speed of variational inference gives it an advantage, particular for exploratory regression, where a very large model is fitted to gain an understanding of the data and identify a subset of the microbiome which can be explored in more detail.

Approximate solutions arise in variational inference by restricting the family of densities which can be used as a proxy for the exact conditional density. Typically, the mean field variational family is used where independence is assumed across the factors. Thus by specifying conjugate priors, approximate marginal posteriors are members of the exponential family (Carbonetto and Stephens, 2012). However, many models of interest such as logistic regression and non conjugate topic models, do not enjoy the properties required to exploit this algorithm. Using variational inference in these settings require algorithms to be adjusted to for the specific model requirement. A variety of strategies have been explored including alternative bounds (Jaakkola and Jordan, 1997, Bishop and Svensen, 2003), numerical quadrature (Honkela and Valpola, 2005) and Monte Carlo approximation (Ye et al., 2020).

We propose a Bayesian hierarchical linear log-contrast model for compositional data which is estimated by mean field Monte Carlo co-ordinate ascent variational inference. We use the alr transformation proposed by Lin et al., 2014, because it is symmetric and removes the need to specify a reference category. Sparse variable selection is performed through novel priors within a hierarchical prior framework which account for the constrained parameter space associated with the compositional covariates and the different orders of magnitude in the taxon abundances. As our constrained priors are not conjugate, Monte Carlo expectations are used to approximate intractable integrals. These expectations are obtained via a reversible jump Monte Carlo Markov chain (RJMCMC) (Green, 1995), which is guided by the data through univariate approximations of the intractable variational posterior probability of inclusion. We exploit the nested nature of variational inference by proposing parameters from approximated variational densities via auxiliary parameters. Model averaging over all the explored models can be performed and shrunk estimates of the regression coefficient (by the model uncertainty) are available. The approach accommodates high dimensional microbial data and offers the potential to be scaled up for models with multiple responses.

We compare the performance of the proposed modelling approach with lasso, group lasso and Ordinary Least Squares (OLS) regressions on simulated data. The methods are then applied to a subset of the "Know Your Heart" cross-sectional study of cardiovascular disease (Cook et al., 2018) in order to examine the association of the gut microbiome with body mass index (BMI). The study was conducted in two Russian cities Novosibirsk and Arkhangelsk, enrolling 4542 men and women aged between 35-69 years recruited from the general population. A health check questionnaire was completed, providing information on smoking, weight and levels of alcohol consumption. We analyse the microbiome of 515 subjects from the Arkhangelsk region at the phylum and genus level, as the 16S rRNA sequencing of faecal samples was only performed for these participants, alongside age and health covariates.

3 Methods

3.1 Microbiome Model

The microbiome data begins as raw counts for each taxon. Any zeros are replaced by a small pseudo-count (typically 0.5), before each row is standardised to sum to 1. The sample space of a vector of components is a simplex for each data point, where the rows of each vector make up the design matrix $\boldsymbol{Q}_{n\times d}$. The set of compositional explanatory variables can be transformed onto the unconstrained sample space $\mathbb{R}^{d-1}$ using the alr transformation

$$alr(\boldsymbol{q}_i) = \Big[\log\Big(\frac{q_{i1}}{q_{id}}\Big), \log\Big(\frac{q_{i2}}{q_{id}}\Big), ..., \log\Big(\frac{q_{id-1}}{q_{id}}\Big)\Big], \tag{3.1}$$

where $\boldsymbol{q}_i$ is the ith row of $\boldsymbol{Q}$ and the ratios have been arbitrarily chosen to involve the division of each of the first $d-1$ components by the final component. The log linear model, with the alr transformed variables as proposed by Aitchison and Bacon-Shone, 1984, can be expressed as

$$y_i = \alpha \mathbf{1}_n + alr(\boldsymbol{q}_i)\tilde{\boldsymbol{\theta}} + \epsilon_i \tag{3.2}$$

where $\tilde{\boldsymbol{\theta}} = (\theta_1, ..., \theta_{d-1})^T$ is the corresponding $(d-1)$ vector of regression coefficients and ϵ_i is independent noise distributed as $N(0, \sigma^2)$. Although convenient, the interpretation of the model depends on the arbitrary choice of the reference category. If we expand the dot product $alr(\boldsymbol{q}_i) \cdot \tilde{\boldsymbol{\theta}}$ and set

$$\theta_d = -\sum_{j}^{d-1} \tilde{\theta}_j \tag{3.3}$$

the linear model can be conveniently expressed in matrix form (Lin et al., 2014) as

$$\mathbf{y} = \alpha \mathbf{1}_n + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad \text{subject to } \sum_{j=1}^{d} \theta_j = 0 \tag{3.4}$$

where $\boldsymbol{Z} = (\log \boldsymbol{q}_1, ..., \log \boldsymbol{q}_d)$ is the $n \times d$ compositional design matrix and $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^T$ is a d-vector of regression coefficients constrained to the affine hyperplane.

This likelihood is used by Zhang et al., 2020 who specify a d dimensional multivariate normal distribution on θ within a "spike-and-slab" prior,

$$\boldsymbol{\theta}|\sigma^2, \psi, \mathbf{V} \sim N_d(\mathbf{0}, \sigma^2\psi\mathbf{V}), \qquad \mathbf{V} = \mathbf{I}_d - \frac{c^2}{1 + c^2 d}\mathbb{J}_d \tag{3.5}$$

where $\mathbb{J}_d$ is a matrix of ones and $\mathbf{V}$ is the generalised transformation matrix which incorporates the tuning parameter c to constrain the $\boldsymbol{\theta}$ parameter space and takes the form in (3.5) for the alr transformation. This approach allows the probability distribution to remain in the d dimensional space as $\mathbf{V}$ is a matrix of full rank, facilitating conjugate updates, as the sum to zero constraint is not imposed exactly.

Interest often lies in assessing the association of unconstrained data, in the form of categorical or continuous covariates against the response, alongside the microbiome. Two additional design matrices are added to the likelihood, $\boldsymbol{X}$ which comprises the scaled continuous covariates and $\boldsymbol{W}$ which contains the dummy variables for the $g = 1, ..., G$ categorical variables coded to indicate the m_g levels with respect to the intercept. The likelihood for our model is thus expressed as

$$\mathbf{y} = \alpha\mathbf{1}_n + \boldsymbol{X\beta} + \boldsymbol{W\zeta} + \boldsymbol{Z\theta} + \boldsymbol{\epsilon} \quad \text{subject to} \sum_{j=1}^{d} \theta_j = 0. \tag{3.6}$$

3.2 Compositional Priors

The linear constraint on the unconstrained vector can be expressed in matrix form as

$$\mathbf{T} = (\mathbf{I}_d - (1/d)\mathbb{J}_d) \tag{3.7}$$

where $\mathbf{T}$ is an idempotent matrix of rank $d-1$. If we originally parametrise $\theta_j \sim N(\mu_j, \psi_j)$, where the large differences in the order of magnitude of each row of the $\boldsymbol{Z}$ design matrix are accounted for by allowing each parameter θ_j to have a separate variance parameter ψ_j, then the constrained random variables associated with the compositional explanatory variables are from a singular multivariate normal distribution

$$\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\psi} \sim SMVN_d(\mathbf{T}\boldsymbol{\mu}, \mathbf{T}\text{diag}(\boldsymbol{\psi})\mathbf{T}^T). \tag{3.8}$$

with $\boldsymbol{\psi}$ a vector of scale parameters. This prior respects the sum to zero constraint imposed by the reparametrisation of the likelihood in (3.6). The distribution is degenerate, the transformation matrix $\mathbf{T}$ means the covariance matrix is singular, and will assign 0 values to all sets in the d dimensional space. Zhang et al., 2020 treat the constraint as a tuning parameter, restricting the values that $\boldsymbol{\theta}$ can take whilst still remaining in the d dimensional space so that the marginal posterior can be obtained in closed form. Our approach imposes the constraint exactly. The singular multivariate normal prior for the compositional data can be considered to be at the unobtainable limit of c in the alr transformation approach (3.5), when the tuning parameter creates a singular matrix where the standard normal prior is no longer appropriate.

We augment the prior on $\boldsymbol{\theta}$ with dependent latent indicator variables from a product of Bernoulli distributions which have been truncated to account for the alr transformation which prevents the selection of a single taxon into the model

$$p(\boldsymbol{\xi}|\kappa) \propto \prod_{j=1} \kappa^{\xi_j}(1-\kappa)^{1-\xi_j}\mathrm{I}\Big[\sum_j \xi_j \neq 1\Big], \tag{3.9}$$

where I is the indicator function. This truncation is particularly important in the presence of sparsity. The full singular multivariate normal spike-and-slab prior for $p(\boldsymbol{\theta}|\boldsymbol{\xi}) =$

$p(\boldsymbol{\theta}_\xi|\boldsymbol{\xi})p(\boldsymbol{\theta}_{\bar{\xi}}|\boldsymbol{\xi})$, where $\boldsymbol{\theta}_\xi$ and $\boldsymbol{\theta}_{\bar{\xi}}$ are subvectors of $\boldsymbol{\theta}$ such that

$$p(\boldsymbol{\theta}_\xi|\Sigma,\boldsymbol{\xi}) = \frac{1}{(\det^*(2\pi\Sigma_\xi^+))^{(-1/2)}}\exp\left(-\frac{1}{2}\boldsymbol{\theta}_\xi\Sigma_\xi^+\boldsymbol{\theta}_\xi\right) \quad \text{and} \quad p(\boldsymbol{\theta}_{\bar{\xi}}=0|\boldsymbol{\xi})=1, \tag{3.10}$$

Σ_ξ^+ denotes the pseudo inverse of the matrix $\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi$, $\det^*$ the psuedo inverse and $\boldsymbol{\xi}$ is a vector of zeros and ones. The $\boldsymbol{\theta}_\xi$ parameters are dependent (the covariance for unit scale is equal to the fraction $-1/d_\xi$ and for the case of $d_\xi = 2$ the correlation is 1). This prior implies a univariate spike-and-slab on the diagonal of the covariance matrix in (3.10).

$$p(\boldsymbol{\psi}|\boldsymbol{\xi}) = \prod_{j=1}^{d}\left[\frac{b_\psi^{a_\psi}}{\Gamma(a_\psi)}(\psi_j)^{-a_\psi-1}\exp\{-b_\psi\psi_j^{-1}\}\right]^{\xi_j}\delta_0(\psi_j)^{1-\xi_j} \quad \psi_j > 0 \ \forall\ j. \tag{3.11}$$

A beta distribution is placed on the sparsity parameter κ and the hyperparameter b_ψ is given a gamma prior. This approach can be interpreted as replacing the continuous mixing density in the Bayesian lasso, which can have either hierarchical structure (Leng et al., 2014) or be marginalised over (Park and Casella, 2008), with a discrete mixture. This set of explicit variable selection priors on the compositional data ensures that the marginal posterior of variable ξ_j represents the inclusion of the jth taxon in the model.

3.3 Priors

The choice of the remaining prior distributions is partly down to convenience. The prior distributions and likelihood are semi-conjugate pairs which means the optimal form for the mean field variational density is in the same exponential family form.

We employ a variable selection spike-and-slab prior George and McCulloch, 1997 for β_s associated with the continuous variables in the design matrix $\boldsymbol{X}$, where each s parameter is independent. The spike is a point mass at 0 (Dirac distribution) with probability $1 - p(\gamma_s) = 1 - \omega$ and the slab is a zero centred Gaussian with variance w which requires the variables to be standardised. The binary latent indicator variable γ_s represents the inclusion of the sth covariate in the model.

In the case of the categorical data matrix, we are interested in selecting the group of variables associated with the response into the model, rather than a particular level. Each factor variable (or group) $g = 1, .., G$ has $j = 1, ..., m_g, m_g + 1$ levels which are coded as dummy variables in $\boldsymbol{W}$ with reference to the intercept. Motivated by the Bayesian group lasso (Xu and Ghosh, 2015) who introduce binary indicators to perform selection both between and within the groups levels, we employ a variable selection spike-and-slab prior on the vector $\boldsymbol{\zeta}_g$ with dimension m_g. The spike is a point mass at 0 (Dirac distribution) with probability $1 - p(\chi_g) = 1 - \varrho$ and the slab is a zero centred Gaussian with variance v. The binary latent indicator variable χ_g represents the inclusion of the gth categorical variable into the model. In the case where there factors have just 2 levels, the prior reduces to the same form as its unrestricted continuous counterpart, with a different scale parameter.

Hierarchical priors are also included to fully incorporate the uncertainty surrounding these parameters. The probability that a given covariate in the design matrices of $\boldsymbol{X}$ and $\boldsymbol{W}$ affects the response is modelled by the parameters ω and ϱ, with beta priors. Inverse gamma distributions with gamma (shape and scale) hyperpriors on their respective scales are placed on the prior variance parameters w and v.

3.4 Variational Inference

We employ coordinate ascent variational inference (CAVI) (Blei et al., 2017) as our estimation procedure, rather than relying entirely on MCMC which often requires substantial computing resources when the dimensionality of the problem is large. We use the *mean field variational family*, but allow dependencies within each member (block), where the latent variables are mutually independent and each governed by a distinct factor in the variational density. We define the blocks to ensure the dependency between the latent indicator variable(s) and their associated parameter(s) is captured. An example of a block is the joint q approximating density for the prior parameters $q(\beta_s, \gamma_s)$ directly associated with the design matrix $\boldsymbol{X}$. The full mean field approximation distribution $q(\boldsymbol{\vartheta})$ is defined in the Supplementary Materials.

3.5 Unconstrained Updates

The variational inference updates are available analytically for all unconstrained parameters and hyperparameters in the model. Derivations are given in the Supplementary Material. The updates involve a combination of univariate and multivariate calculations. The regression parameters directly associated with the $\boldsymbol{X}$ and $\boldsymbol{W}$ design matrices have joint updates in the same spike-and-slab form as their priors. The conjugate update for $q(\beta_s, \gamma_s)$ is

$$q(\beta_s|\gamma_s, \mathbf{y}) = \mathcal{N}(\mu_{\beta_s}, \sigma^2_{\beta_s})^{\gamma_s}\delta_0(\beta_s)^{1-\gamma_s} \qquad q(\gamma_s|\mathbf{y}) = Bern((\gamma_s)^{(1)}).$$

with free parameters

$$\begin{aligned}
\sigma^2_{\beta_s} &= \left(\|X_s\|^2(\sigma^{-2})^{(1)} + (w^{-1})^{(1)}\right)^{-1}, \\
\mu_{\beta_s} &= (\sigma^{-2})^{(1)}\sigma^2_{\beta_s}X_s^T\left(\mathbf{y} - (\alpha)^{(1)}\mathbf{1}_n - \sum_{k\neq s} X_k(\beta_k)^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_g)^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta}_\xi)^{(1)}\right), \\
(\gamma_s)^{(1)} &= \left[1 + \exp\left\{(\log(1-\omega))^{(1)} - (\log\omega)^{(1)} - \frac{1}{2}\left((\log w^{-1})^{(1)} - \mu^2_{\beta_s}\sigma^{-2}_{\beta_s} - \log\left(\sigma^2_{\beta_s}\right)\right)\right\}\right]^{-1},
\end{aligned}$$

where $(\cdot)^{(1)}$ denotes the q expectation. The conjugate update for $q(\boldsymbol{\zeta}_g, \chi_g)$ is

$$q(\boldsymbol{\zeta}_g|\chi_g, \mathbf{y}) = \mathcal{N}_{m_g}(\boldsymbol{\mu}_{\zeta_g}, \Sigma_{\zeta_g})^{\chi_g}\delta_0(\boldsymbol{\zeta}_g)^{1-\chi_g} \qquad q(\chi_g|y) = Bern((\chi_g)^{(1)}),$$

where the free parameters for $\boldsymbol{\zeta}_g$ are updated by the multivariate extension of the previous univariate update,

$$\begin{aligned}
\Sigma_{\zeta_g} &= \left[(\sigma^{-2})^{(1)}\boldsymbol{W}_g^T\boldsymbol{W}_g + (\upsilon^{-1})^{(1)}\right]^{-1}, \\
\boldsymbol{\mu}_{\zeta_g} &=(\sigma^{-2})^{(1)}\Sigma_{\zeta_g}\boldsymbol{W}_g^T\Big(\mathbf{y} - (\alpha)^{(1)}\mathbf{1}_n - \sum_s X_s(\beta_s)^{(1)} - \sum_{k\neq g}\boldsymbol{W}_k(\boldsymbol{\zeta}_k)^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta})^{(1)}\Big), \\
(\chi_g)^{(1)} &=\Bigg[1 + \exp\Bigg\{(\log(1-\varrho))^{(1)} - (\log\varrho)^{(1)} - \frac{m_g}{2}(\log\upsilon^{-1})^{(1)} - \frac{1}{2}\boldsymbol{\mu}_{\zeta_g}^T\Sigma_{\zeta_g}^{-1}\boldsymbol{\mu}_{\zeta_g}+ \\
&\quad - \frac{1}{2}\log\big(\det(\Sigma_{\zeta_g})\big)\Bigg\}\Bigg]^{-1}.
\end{aligned}$$

The marginal expectation of $\boldsymbol{\zeta}_g$ and β_s is the mean of the conditional density when the parameter is included in the model, shrunk by the probability of being included in the model. The nested q density update for each free parameter(s) is the expectation of the log joint distribution with respect to all the other factors. Thus, any update involving a marginal expectation from a parameter with a spike and slab prior involves a form of regularisation.

The selection of the spike-and-slab priors for β_s, $\boldsymbol{\zeta}_g$ and $\boldsymbol{\theta}$ with sparsity inducing hyperparameters for variable selection, shrinks the parameters estimates in the variational updates rather then performing explicit variable selection as in MCMC. These estimates are a useful proxy for the final model effects, but as opposed to a model with regularisation priors, the expectation of the model indicator parameters gives us the probability of a covariate being associated with the response. In the case of $\boldsymbol{\zeta}_g$, which is associated with the gth categorical covariate, the parameterisation has a convenient interpretation. Each element in the vector is free to vary but all elements are shrunk by the same value. Thus the expectation $(\chi_g)^{(1)}$ is the probability of the categorical covariate (rather than the individual levels) being included in the model.

3.6 CAVI-MC

The conditional vector update $q(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{\xi})$ is available analytically and takes the form

$$q(\boldsymbol{\theta}_\xi|\boldsymbol{\xi},\mathbf{y}) = SMVN_{d_\xi}(\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_\xi},\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi), \quad q(\boldsymbol{\theta}_{\bar{\xi}}|\boldsymbol{\xi},\mathbf{y}) = \delta_0(\boldsymbol{\theta}_{\bar{\xi}}), \tag{3.12}$$

where δ_0 is the Dirac distribution on the subvector $\boldsymbol{\theta}_{\bar{\xi}}$ with updates

$$\boldsymbol{\mu}_{\theta_\xi} = \Sigma_{\theta_\xi}(\sigma^{-2})^{(1)}\boldsymbol{Z}_\xi^T(\mathbf{y} - (\alpha)^{(1)}\mathbf{1}_n - \sum_s X_s(\beta_s)^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_g)^{(1)}) \tag{3.13}$$

$$\Sigma_{\theta_\xi} = \big((\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)^+ + (\sigma^{-2})^{(1)}\boldsymbol{Z}_\xi^T\boldsymbol{Z}_\xi\big)^{-1} \tag{3.14}$$

The truncated Bernoulli prior distributions for $\boldsymbol{\xi}$ and unique scale parameter ψ_j for each element in $\boldsymbol{\theta}$, prevents a conjugate posterior update for the joint block $q(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\xi})$. All other updates are available analytically.

The difficult to compute joint $q(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\xi})$ update is performed by inserting a Monte Carlo step within the mean field variational inference approach. We take advantage of the structure of the target density $p(\boldsymbol{\vartheta},\mathbf{y}) \equiv f(\boldsymbol{\vartheta})$ (the data $\mathbf{y}$ is omitted for notational purposes as its fixed) which has the form

$$f(\boldsymbol{\vartheta}) = h(\boldsymbol{\vartheta})\exp(\langle\boldsymbol{\eta}, T(\boldsymbol{\vartheta})\rangle - A(\boldsymbol{\eta})), \quad \boldsymbol{\vartheta} \in S_p \tag{3.15}$$

for r-dimensional constant vector $\boldsymbol{\eta}$, vector function $T(\boldsymbol{\vartheta})$ and relevant scalar functions $h > 0$. In our case this admits the factorisation

$$h(\boldsymbol{\vartheta}) = h_{q(\vartheta_j)}(\vartheta_j)h_{q(\boldsymbol{\vartheta}_{-j})}(\boldsymbol{\vartheta}_{-j}), \qquad T_l(\boldsymbol{\vartheta}) = T_{l,j}(\vartheta_j)T_{l,-j}(\boldsymbol{\vartheta}_{-j}), \quad 1 \leq l \leq r, \text{ for all } j \notin \mathcal{J},$$

where $\mathcal{J}$ is the set of all analytically available updates. This allows us to avoid generating and storing the samples from the approximating densities which would involve considerable computational cost, by using the q marginal expectations in the Monte Carlo estimate for $q(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi})$. Ye et al., 2020 show that, under regularity conditions, an MC-CAVI recursion will get arbitrarily close to a maximiser of the evidence lower bound with any given high probability.

The MCMC approach involves two move types, within-model moves where the samples are generated from a Metropolis-Hastings sampler and between-model moves which are sampled from a RJMCMC. The samplers involve using some form of the joint approximating posterior $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi}|\boldsymbol{y}) \propto q(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{y})q(\boldsymbol{\xi}, \boldsymbol{\psi}|\boldsymbol{y})$ which is simplified as $q(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{y})$ has the conjugate spike-and-slab form (3.12).

Randomly choose either a between-model move which consists of sequentially updating $\boldsymbol{\xi}, \boldsymbol{\psi}|\boldsymbol{\xi}$ and $\boldsymbol{\theta}|\boldsymbol{\psi}$, $\boldsymbol{\xi}$ or a within-model move where $\boldsymbol{\xi}$ is not updated. This naturally leads to questions regarding the proposals for $\boldsymbol{\psi}$ which has a constrained support and $\boldsymbol{\xi}$ which has the potential to be a very large binary space.

3.6.1 Between-model RJMCMC - Approximating $q(\boldsymbol{\xi}, \boldsymbol{\psi}|\boldsymbol{y})$ to $p(\boldsymbol{\xi}|\boldsymbol{\vartheta})$ for the proposal distribution $j_m(\boldsymbol{\xi}, \boldsymbol{\xi}')$

The choice of priors for the parameters associated with microbiome features, the indicator vector $\boldsymbol{\xi}$ and set of scale parameters $\boldsymbol{\psi}_\xi$, prevents a conjugate update for $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$. An MCMC step is introduced to sample from the intractable q approximating posterior. To search the binary space we use a RJMCMC where the proposal for ψ_j conditional on $\xi_j = 1$ is from the q approximating density of the auxiliary parameter Ω_j

$$\pi(\psi_j|\xi_j = 1) = IG_q(a^*_{\Delta_j}, b^*_{\Delta_j}), \tag{3.16}$$

where the calculation of the free parameters $a^*_{\Delta_j}$ and $b^*_{\Delta_j}$ is explained in the next section. $\boldsymbol{\theta}$ is generated directly from the singular multivariate normal target distribution (3.12).

There is considerable research in sampling high-dimensional binary vectors. Lamnisos et al., 2009 propose a general model for the proposal which combines local moves with global ones by changing blocks of variables. They find that the acceptance rates for Metropolis-Hastings samplers that include, exclude or swap a single variable improves. Lamnisos et al., 2013 extend their model with adaptive parameters which change during the mixing of the MCMC. Motivated by incorporating information from data into the proposal parameters, we use the variational inference posterior distribution $q(\boldsymbol{\xi}, \boldsymbol{\psi}|\mathbf{y})$ which is only available up to a constant of proportionality

$$\begin{aligned}
&q(\boldsymbol{\xi}, \boldsymbol{\psi}|\mathbf{y}) \propto \ \exp\Bigg(\frac{1}{2}(\boldsymbol{\mu}^T_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi(\mathbf{T}_\xi^T\Sigma_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi)^+\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_{(\xi,\psi)}}) + \frac{1}{2}\log\Big(\det{}^*(\mathbf{T}_\xi\Sigma_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi)\Big)+\\
&\sum_j \xi_j(\log\kappa)^{(1)} - \frac{1}{2}\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)) + \sum_j(1-\xi_j)(\log(1-\kappa))^{(1)}+\\
&-(a_\psi+1)\sum_j \xi_j\log(\psi_j) - b_\psi\sum_j \xi_j\psi_j^{-1} + (a_\psi\log(b_\psi) - \log(\Gamma(a_\psi))\sum_j \xi_j\Bigg), \qquad (3.17)
\end{aligned}$$

to obtain a univariate approximation relative to the jth element to guide the RJMCMC. These normalised probabilities are used to obtain our proposal probabilities in a birth-death and swap sampling scheme. Similar to adaptive parameters in MCMC, these selection probabilities are updated at each iteration of the CAVI.

The pseudo determinant in (3.17) is approximated by removing the constraints $\mathbf{T}_\xi$ and taking the MCMC expectation conditional on $\xi_j = 1$. So for the jth element the approximation is

$$\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)) \approx \{\log(\psi_j)\}^{\{1\}}_{\emptyset} \qquad (3.18)$$

where the curly brackets $\{\}$ denote an MCMC expectation and $\emptyset$ defines an expectation over all non-zero values. A similar approach can be used to approximate the determinant containing Σ_{θ_ξ}

$$\log\bigl(\det{}^*(\mathbf{T}_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi)\bigr) \approx \log\Bigl(\bar{\sigma}^2_{\theta_j}\Bigr).$$

where $\bar{\sigma}^2_{\theta,tj}$ is the non-zero variance average over the MCMC iterations, obtained by extracting the diagonal from $\Sigma_{\theta_{(\xi,\psi)}}$ at each iteration. If the jth term has not been included in the model the term is approximated by

$$\log\bigl(\det{}^*(\mathbf{T}_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi)\bigr) \approx \log\Bigl(\bigl[\|Z_j\|^2 (\sigma^{-2})^{(1)}\bigr]^{-1}\Bigr) \tag{3.19}$$

After approximating Σ_{θ_ξ} to a scalar for each jth element the matrix dot product reduces to

$$\boldsymbol{\mu}^T_{\theta_\xi} \mathbf{T}_\xi (\mathbf{T}^T_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi} \approx \bar{\sigma}^2_{\theta_j} \Bigl(\sum_j (1 - 1/d_\xi) \mu^2_{\theta_{\xi_j}} - 2 \sum_{j<j'} (\mu_{\theta_{\xi_{j'}}}, \mu_{\theta_{\xi_j}} / d_\xi) \Bigr). \tag{3.20}$$

To account for the cross product terms which contains the elements of $\boldsymbol{\xi}$ not equal to j and the associated $\boldsymbol{\mu}_\theta$ terms, a combination of conditional expectations and marginal expectations which shrink the values in proportion to its probability of being zero, is used. As ξ_j can not be separated from the sum in the numerator d_ξ, two approximations of the matrix dot product are used conditional on the expectation from the previous chain.

Defining the expectations with respect to the parameter currently being updated from the previous MCMC by a curly bracket as:

- $\{\mu_{\theta_j}\}^{\{1\}}_{\not 0}$: Conditional expectation $\xi_j = 1$. Weighted average of the nonzero terms from previous chain,
- $\{\mu_{\theta_j}\}^{\{1\}}$: Expectation wrt q from the previous chain,
- $\{d_\xi\}^{\{1\}}$: Expectation wrt q from the previous chain,

the approximation of the dot product $(\mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi})^T \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi}$ is thus

$$\begin{array}{ll} \bar{\sigma}_{\theta,j}^{-2}\left(\sum_j (1 - 1/\{d_\xi\}^{\{1\}})\xi_j(\{\mu_{\theta_j}\}_{\emptyset}^{\{1\}})^2 - \frac{2}{\{d_\xi\}^{\{1\}}} \sum_{j<j'} \xi_j \{\mu_{\theta_{\xi_j}}\}_{\emptyset}^{\{1\}} \{\mu_{\theta_{\xi_{j'}}}\}^{\{1\}} \right) & \{d_\xi\}^{\{1\}} > 2 \\ \bar{\sigma}_{\theta,j}^{-2} \sum_j \xi_j(\{\mu_{\theta_j}\}_{\emptyset}^{\{1\}})^2 & \{d_\xi\}^{\{1\}} < 2. \end{array}$$

Although $\{d_\xi \in \mathbb{N}_0 | d_\xi \leq d, d_\xi \neq 1\}$, the support of the MCMC expectation $\{d_\xi\}^{\{1\}}$ is the positive real line so we threshold on 2. When $\{d_\xi\}^{\{1\}} > 2$ the probabilities used in the proposal distribution for the RJMCMC, derived from approximating Equation (3.17) and normalising is

$$\begin{aligned} \tilde{p}(\xi_j = 1|\boldsymbol{\vartheta}) \equiv \Bigg[\exp \Bigg\{ & (\log(1-\kappa))^{(1)} - \frac{1}{2\bar{\sigma}_{\theta,j}^2} \Bigg((1 - 1/\{d_\xi\}^{\{1\}})(\{\mu_{\theta_j}\}_{\emptyset}^{\{1\}})^2 + \\ & - \frac{2}{\{d_\xi\}^{\{1\}}} \{\mu_{\theta_{\xi_j}}\}_{\emptyset}^{\{1\}} \sum_{j' \neq j} \{\mu_{\theta_{\xi_{j'}}}\}^{\{1\}} \Bigg) - \frac{1}{2} \log\left(\bar{\sigma}_{\theta,j}^2\right) + \frac{1}{2}(\log \psi_j)_{\emptyset}^{\{1\}} - (\log \kappa)^{(1)} + \\ & (\log \Gamma(a_\psi) - a_\psi \log b_\psi) + +(a_\psi + 1)(\log \psi_j)_{\emptyset}^{\{1\}} + b_\psi(\psi_j^{-1})_{\emptyset}^{\{1\}} \Bigg\} + 1 \Bigg]^{-1}, \end{aligned} \tag{3.21}$$

which contains the variational expectations and an MCMC conditional expectation from the previous iterations. This is then used to propose the various move types in the RJMCMC.

3.6.2 Pseudo Updates for MCMC proposals

A conjugate update for the parameters associated with the microbiome features $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$ is prevented by the choice of priors for the indicator vector $\boldsymbol{\xi}$ and set of scale parameters $\boldsymbol{\psi}_\xi$. Samples from the intractable q approximating posterior are simulated from an MCMC step instead. The move types in the RJMCMC for $\boldsymbol{\xi}$ use an element-wise approximation of the joint $q(\boldsymbol{\xi})$ density (3.21). For the proposal distribution of $\boldsymbol{\psi}$, we use the model likelihood and an unconstrained approximation to the constrained priors. In order to do this we define auxiliary parameters (upper case Greek letters) which are unconstrained versions of the constrained parameters. We derive pseudo variational updates from an unconstrained model with a simpler prior parametrisation, then use the q approximating distribution of the relevant auxiliary parameter as our proposal for $\boldsymbol{\psi}$. We can think of the auxiliary parameters as introducing an alternative directed acyclic graph (DAG) which is updated first, helping us to approximate the model in order to guide the MCMC step. These updates are refined by the full variational inference updates which account for the constraint at each iteration. The parameter κ and the hyperparameters a_Δ, b_Δ which are set to a_ψ, b_ψ provide a link back to the constrained model.

The series of pseudo variational updates are determined from a simple prior parametrisation where the parameters associated with the compositional covariates are not constrained to sum to 0. This unconstrained model has the following prior parametrisation

$$p(\Omega_j|\Delta_j, \Upsilon_j) = N(\Omega_j|0, \Delta_j)^{\Upsilon_j}\delta_0(\Omega_j)^{1-\Upsilon_j} \qquad p(\Delta_j|\Upsilon_j) = IG(\Delta_j|a_\Delta, b_\Delta)^{\Upsilon_j}\delta_0(\Delta_j)^{1-\Upsilon_j}$$

$$p(\Upsilon_j) = Bern(\Upsilon_j|\kappa).$$

Where $\boldsymbol{\Omega}$ are the unconstrained version of the $\boldsymbol{\theta}$ parameters, $\boldsymbol{\Delta}$ are the variance parameters for $\boldsymbol{\Omega}$ which are both dependent on the model selection parameters $\boldsymbol{\Upsilon}$. The prior for the model selection parameter Υ_j is a simple Bernoulli distribution. The remaining priors and likelihood take the form defined in the initial prior parametrisation. The introduction of independence across each univariate $(\Omega_j, \Delta_j, \Upsilon_j)$ block, (where the data is being treated as unconstrained) ensures the q expectations are all available in closed form (derived in the Supplementary Section).

Despite the similarities of the prior parametrisation to (3.5), the addition of a separate scale parameter Δ_j for Ω_j prevents a joint conjugate update on the $(\Omega_j, \Delta_j, \Upsilon_j)$ block. Instead we update $q(\Omega_j, \Upsilon_j)$ (for $j = 1, ..., d$) before updating $q(\Delta_j|\Upsilon_j)$. Both require expectations conditional on Υ_j as well as the typical marginal expectations. The $q(\Omega_j, \Upsilon_j)$

update is

$$
\begin{aligned}
&q(\Omega_j, \Upsilon_j) \propto N(\Omega_j | \mu_{\Omega_j}, \sigma^2_{\Omega_j})^{\Upsilon_j} \delta_0(\Omega_j)^{1-\Upsilon_j} && (3.22)\\
&\Big\{ \exp\Big(\frac{1}{2} \log \sigma^2_{\Omega_j} + (\log \kappa)^{(1)} - \frac{1}{2} \mathbb{E}_q(\log \Delta_j | \Upsilon_j) + \frac{1}{2} \mu^2_{\Omega,j} \sigma^{-2}_{\Omega,j} + a_\Delta \log(b_\Delta) + && (3.23)\\
&- \log(\Gamma(a_\Delta)) - (a_\Delta + 1)\mathbb{E}_q(\log \Delta_j | \Upsilon_j) - b_\Delta \mathbb{E}_q[\Delta_j^{-1} | \Upsilon_j] \Big) \Big\}^{\Upsilon_j} \Big\{ 1 - \kappa)^{(1)} + \delta_0(\Delta_j) \Big\}^{1-\Upsilon_j}
\end{aligned}
$$

The binary form of the pseudo update for Ω_j and Υ_j enables us to determine the values for the conditional expectations. In Equation (3.22) we have under q, where we condition on the value of Υ_j

$$
q(\Omega_j | \Upsilon_j = 1, \mathbf{y}) = \mathcal{N}(\mu_{\Omega,j}, \sigma^2_{\Omega,j}) \quad q(\Omega_j | \Upsilon_j = 0, \mathbf{y}) = \delta_0(\Omega_j), \tag{3.24}
$$

which allows us to set the expectations in the normal variance update as $\mathbb{E}_q[\Delta_j^{-1} | \Upsilon_j = 1]$

$$
\sigma^2_{\Omega,j} = \left(\| Z_j \|^2 (\sigma^{-2})^{(1)} + \mathbb{E}_q[\Delta_j^{-1} | \Upsilon_j = 1] \right)^{-1} \tag{3.25}
$$

$$
\mu_{\Omega,j} = \sigma^2_{\Omega,j} Z_j^T \left\{ (\sigma^{-2})^{(1)} \left(\mathbf{y} - \sum_{k \neq j} Z_k (\Omega_k)^{(1)} - \sum_s X_s (\beta_s)^{(1)} \right) \right\}. \tag{3.26}
$$

The conditional expectation prevents us averaging over Υ_j which shrinks the marginal expectation, creating an update which has the same form as (3.5). Using the form of (3.23) to determine the conditional expectation and normalising gives the probability of inclusion

$$
\begin{aligned}
(\Upsilon_j)^{(1)} = \Bigg[\exp \Bigg\{ &\frac{\log\left(\sigma^{-2}_{\Omega,s}\right)}{2} + (\log(1-\kappa))^{(1)} - (\log \kappa)^{(1)} + \frac{\mathbb{E}_q(\log \Delta_j | \Upsilon_j = 1)}{2} + \log \Gamma(a_\Delta) \\
&- \frac{1}{2} \mu^2_{\Omega,j} \sigma^{-2}_{\Omega,j} - a_\Delta \log(b_\Delta) + (a_\Delta + 1)\mathbb{E}_q(\log \Delta_j | \Upsilon_j = 1) + b_\Delta \mathbb{E}_q[\Delta_j^{-1} | \Upsilon_j = 1] \Bigg\} + 1 \Bigg]^{-1}.
\end{aligned}
$$

The univariate approximation of $q(\boldsymbol{\xi}, \boldsymbol{\psi} | \boldsymbol{y})$ (3.21) can be interpreted as a refinement of $(\Upsilon_j)^{(1)}$ using MCMC expectations and information on all elements of $\boldsymbol{\xi}$ to partially account for the constraint in the probability of inclusion.

The spike-and-slab form of the pseudo update for $q(\Delta_j|\Upsilon_j)$ allows us to again back out the conditioning in the conditional expectation of $\mathbb{E}_q[\Omega_j^2|\Upsilon_j]$ in $b^*_{\Delta_j}$.

$$q(\Delta_j|\Upsilon_j = 1, \mathbf{y}) = IG\left(\Delta_j\middle|\frac{1}{2} + a_\psi, \frac{(\sigma^2_{\Omega,j} + \mu^2_{\Omega,j})}{2} + b_\psi\right), \quad q(\Delta_j|\Upsilon_j = 0, \mathbf{y}) = \delta_0(\Delta_j)$$

As the update Δ_j is conditional on Υ_j, the free parameters in the proposal distributions are not a function of shrunken estimates. The $q(\Delta_j|\Upsilon_j, \mathbf{y})$ auxiliary approximating density is then used to propose scale parameters with the appropriate support, which are informed by the data, for $\boldsymbol{\psi}_\xi$ in the MCMC move.

3.6.3 Algorithm

CAVI is performed by iterating through the analytical variational updates, maximising the evidence lower bound (ELBO) with respect to each coordinate direction whilst fixing the other coordinate values. For the $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$ block an MCMC is implemented to obtain Monte Carlo estimates of the intractable marginal expectations of the approximating densities. The proposal probabilities for the sampling scheme are a function of the data and the free parameters, and are updated at each iteration of the CAVI.

For each run we compute the ELBO (derived in Section 1 of the Supplementary Material), with the updated free parameters, until this converges to the local optimum. The ELBO is no longer monotonically increasing because of the Monte Carlo variability, but we are able to declare convergence when the random fluctuations are small around a fixed point. The implementation of the overall approach is described in Algorithm 1, with the MCMC move detailed in 2.

It is computationally inefficient to start with a large number of iterations m, when the current variational distribution can be far from the maximiser. The software allows the user to specify a smaller number of iterations to begin with before increasing the number of iterations as the algorithm becomes more stable, improving the accuracy of the Monte Carlo estimates.

```
Algorithm 1: MC - CAVI for variable selection.
Input    : A model $p(\mathbf{y}, \boldsymbol{\vartheta})$, a data set $\mathbf{y}$. Number of Monte Carlo samples $m$.
Output   : Variational densities $q(\boldsymbol{\vartheta}_{-(\theta,\psi,\xi)}) = \prod_v q_v(\vartheta_v)$ and Monte Carlo
           expectations.
Intialize: First and second order raw moments of the variational factors, prior
           hyperparameters.
for $k$ = 1,..,$K$ do
    for $v$ = 1,...,$V$ do
        Set $q_v(\vartheta_v) \propto \exp\{\mathbb{E}_{-v}[\log p(\vartheta_v|\boldsymbol{\vartheta}_{-v}, \mathbf{y})]\}$
    end
    Calculate the arguments for proposal distribution for $\boldsymbol{\psi}$ from the psuedo
      variational updates.
        $$a^*_{\Delta_j} = \frac{1}{2} + a_\Delta \qquad b^*_{\Delta_j} = \frac{1}{2}(\mu^2_{\Omega_j} + \sigma^2_{\Omega_j}) + b_\Delta$$
        $$\psi_j \sim IG(a^*_{\Delta_j}, b^*_{\Delta_j})$$
      Calculate the probabilities $\tilde{p}(\boldsymbol{\xi}|\boldsymbol{\vartheta})$ for the $\boldsymbol{\xi}$ proposal (by approximating $q(\boldsymbol{\xi}|\boldsymbol{y})$
      and normalising) in the RJMCMC.
        $$\tilde{p}(\xi_j = 1|\boldsymbol{\vartheta}) \equiv \left[\exp\left\{(\log(1-\kappa))^{(1)} - \frac{1}{2}\log\left(\bar{\sigma}^2_{\theta,j}\right) + \frac{1}{2}(\log\psi_j)^{\{1\}}_{\not{0}} - (\log\kappa)^{(1)} + \right.\right.$$
        $$\left. + (\log\Gamma(a_\psi) - a_\psi \log b_\psi) + (a_\psi + 1)(\log\psi_j)^{\{1\}}_{\not{0}} + b_\psi(\psi_j^{-1})^{\{1\}}_{\not{0}}\right\} +$$
        $$\left. - \frac{1}{2\bar{\sigma}^2_{\theta,j}}\left((1 - 1/\{d_\xi\}^{\{1\}})(\{\mu_{\theta_j}\}^{\{1\}}_{\not{0}})^2 - \frac{2}{\{d_\xi\}^{\{1\}}}\{\mu_{\theta_{\xi_j}}\}^{\{1\}}_{\not{0}} \sum_{j' \neq j}\{\mu_{\theta_{\xi_{j'}}}\}^{\{1\}}\right) + 1\right]^{-1}$$
      Perform MCMC step Algorithm:
      return $\mathbb{E}_q(\boldsymbol{\xi}|\mathbf{y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\psi}|\mathbf{y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\theta}|\mathbf{y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\theta}_\xi^T \boldsymbol{Z}_\xi^T \boldsymbol{Z}_\xi \boldsymbol{\theta}_\xi|\mathbf{y})^{[k]}$ and cross product
      terms in the ELBO calculation
      Compute ELBO.
end
return $q(\boldsymbol{\vartheta}_{-(\theta,\psi,\xi)})$, $\mathbb{E}_q(\boldsymbol{\xi}|\mathbf{y})$, $\mathbb{E}_q(\boldsymbol{\psi}|\mathbf{y})$, $\mathbb{E}_q(\boldsymbol{\theta}|\mathbf{y})$.
```

Algorithm 2: MCMC step for CAVI-MC.

```
Input: $k$ current loop of CAVI-MC, $q$ expectations, pseudo VB updates,
       normalised approximate marginal probability $p(\boldsymbol{\xi}|\boldsymbol{\vartheta})$.
for $i$ = 1,...,$m$ do
    if Between-model move proposed then
        Given the current position of the variational samples $\boldsymbol{\xi}$, $\boldsymbol{\psi}_\xi$ and $\boldsymbol{\theta}_{(\psi,\xi)}$,
          propose either a birth-death move or swap move.
        Propose a new model with probability $j_m(\boldsymbol{\xi},\boldsymbol{\xi}') \propto \tilde{p}(\boldsymbol{\xi}|\cdot)$.
        Draw $\boldsymbol{\psi}'$ proposals for all the nonzero elements in $\boldsymbol{\xi}'$ with probability

          $$\pi(\boldsymbol{\psi}'|\boldsymbol{\xi}', a^*_{\Delta_j}, b^*_{\Delta_j}) = \prod_j \left[ IG\Big(\psi_j|\frac{1}{2}+a_\Delta, \frac{1}{2}(\mu^2_{\Omega_j}+\sigma^2_{\Omega_j})+b_\Delta\Big)\right]^{\xi'_j}$$

          Calculate the corresponding target mean and variance given $\boldsymbol{\xi}'$ and $\boldsymbol{\psi}'$,
          draw the $\boldsymbol{\theta}'$ proposal

          $$\boldsymbol{\mu}'_{\theta_{(\xi,\psi)}} = \Sigma_{\theta_{(\xi,\psi)}}(\sigma^{-2})^{(1)}\boldsymbol{Z}^T_\xi(\mathbf{u}_{\not{y}})^{(1)} \quad \Sigma^{-1'}_{\theta_{(\xi,\psi)}} = \big((\mathbf{T}_\xi \text{diag}(\boldsymbol{\psi}'_\xi)\mathbf{T}_\xi)^+ + (\sigma^{-2})^{(1)}\boldsymbol{Z}^T_\xi\boldsymbol{Z}_\xi\big)$$
          $$\boldsymbol{\theta}'_{(\psi,\xi)} \sim \text{SMVN}_{d'_\xi}((\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_\xi})', (\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi)'|\cdot)$$

          Part of the target $q(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\xi})$ cancels with the proposal for $\theta$ to give the
          acceptance probability of

          $$\alpha_b = \min\left\{\frac{q(\boldsymbol{\psi}',\boldsymbol{\xi}'|\mathbf{y})j_m(\boldsymbol{\xi}',\boldsymbol{\xi})\pi(\boldsymbol{\psi}|\boldsymbol{\xi},, a^*_\Delta, b^*_\Delta)}{q(\boldsymbol{\psi},\boldsymbol{\xi}|\mathbf{y})j_m(\boldsymbol{\xi},\boldsymbol{\xi}')\pi(\boldsymbol{\psi}'|\boldsymbol{\xi}', a^*_\Delta, b^*_\Delta)}, 1\right\}$$

          with the simplified target density $q(\boldsymbol{\xi},\boldsymbol{\psi}|\mathbf{y})$:
        for $l$=1,...,$L$ do
            Perform within-model moves: Given the current position of the
              variational samples $\boldsymbol{\xi}$, $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ draw proposals $\boldsymbol{\psi}'|\boldsymbol{\xi}$ and $\boldsymbol{\theta}'|\boldsymbol{\psi}',\boldsymbol{\xi}$ using
              the same distributions as the between-model move.
              Proposed moved accepted with probability

              $$\alpha_w = \min\left\{\frac{q(\boldsymbol{\psi}',\boldsymbol{\xi}|\mathbf{y})\pi(\boldsymbol{\psi}|\boldsymbol{\xi}, a^*_\Delta, b^*_\Delta)}{q(\boldsymbol{\psi},\boldsymbol{\xi}|\mathbf{y})\pi(\boldsymbol{\psi}'|\boldsymbol{\xi}, a^*_\Delta, b^*_\Delta)}, 1\right\}.$$
        end
    else
        for $l$=1,...,$L$ do
            Perform within-model moves with probability $\alpha_w$.
        end
    end
end
```

4 Simulation Study

We validate the performance of our variational inference model against two frequentist variable selection approaches, ordinary least squares (OLS) (when $n >> p$) and group lasso regression which have software freely available on CRAN (R, 2017). Importantly, both of these approaches ignore the sum to zero constraint on the associated vector of parameters $\boldsymbol{\theta}$ after the columns of the compositional design matrix $\boldsymbol{Q}$ have been logged.

We generate the covariate data using an approach which is similar to Lin et al., 2014. An $n \times d$ data matrix $\mathbf{O} = (o_{ij})$ is drawn from a multivariate normal distribution $N_p(\boldsymbol{\mu}_o, \Sigma_o)$, and then the compositional covariate matrix $\mathbf{Q} = (q_{ij})$ is obtained via the transformation $q_{ij} = \exp(\tau o_{ij}) / \sum_{k=1}^{d} \exp(\tau o_{ik})$. The covariates thus follow a logistic normal distribution (Aitchison and Shen, 1980). To account for the differences in the order of magnitudes of the components, we fix $\tau = 2$ and let $\mu_{oj} = \log(d \times 0.5)$ for j = 1,...,5 and $\mu_{oj} = 0$ otherwise. As the correlations between the abundances of features in the microbiome can vary quite considerably according to the taxonomy class, we choose three settings for Σ_o: $\Sigma_o = \mathbf{I}$, $(\rho^{|i-j|})$ with $\rho = 0.2$ or 0.5. We vary the number of compositional features from 45 $(n = 100, d = 45)$ to 100 $(n = 100, d = 100)$ and $(n = 200, d = 100)$, but keep the total number of continuous covariates $p = 20$ and categorical covariates $G = 4$ with associated levels (3, 5, 5 and 5) fixed. Two scenarios are simulated from model (3.6), non-zero $\boldsymbol{\theta}$ elements only with $\boldsymbol{\theta} = (1, -1.3, 0.7, 0, 0, -1, 1.3, -0.7, 0, 0, ..., 0)$ ("simple" scenario) and additional non-zero elements of $\boldsymbol{\beta} = (1, -0.8, 0.6, -1.5, 0, 0, ..., 0)$ and the second categorical covariate with reference to the intercept $\boldsymbol{\zeta} = (1, -0.8, 0.6, 1.5)$ as the first level is included in the intercept ("mixed" scenario).

Fast OLS backward selection via Akaike information criterion is performed using "fastbw" (Harrell, 2021), where factors rather than columns are removed from the design matrix. A complete model is fitted and the approximate Wald statistics are computed via restricted maximum likelihood, assuming multivariate normality of estimates. The regularisation paths of the group lasso penalised learning for a sequence of regularization parameters are fitted by “gglaso” (Yang et al., 2020). Group lasso is used so that selection, as in the OLS approach, is performed on the categorical group rather than the individual levels within the factor. The penalty parameter selection is performed using cross validation over a grid of values and the mean squared error loss function. For the CAVI-MC model, vague priors are placed on the hyperparameters and initial q expectations are randomly sampled from the prior distributions. 30 variational inference iterations are performed (although the algorithm typically converges after approximately 8 iterations) for each run. The initial number of between-model MCMC iterations is set to 5000, before 10000 iterations are performed after the 5th set of variational inference updates.

We define the signal to noise ratio (SNR) as SNR = mean $|\boldsymbol{\beta}_\gamma + \boldsymbol{\zeta}_\chi + \boldsymbol{\theta}_\xi|/\sigma$. To generate the data with SNR of 0.5, 1 and 5 the SNR expression is solved for σ and 100 simulations for each setting are preformed. To assess the performance of the approaches we use metrics which evaluate the ability to select the correct variables and estimate the appropriate effects. We compute the l_2 loss $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} + \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} + \hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}||_2$ to assess the accuracy of the coefficient estimates, where the approximate posterior mean is used for the parameter estimate of the Bayesian model. To asses the accuracy of the variable selection, the true positive rate (TPR or sensitivity) and false positive rate (FPR or 1 - specificity) is reported, where positives and negatives in the context of the frequentist approaches refer to non-zero and zero coefficients respectively. Variable selection is performed by thresholding the marginal approximate posterior distributions $\mathbb{E}[q(\gamma_j|y)]$, $\mathbb{E}[q(\chi_j|y)]$ and $\mathbb{E}[q(\xi_j|y)]$ at 0.5. When there is a mixture of different parameters in the true model, the TPR and FPR are also decomposed in to the TPR($\boldsymbol{\theta}$) and FPR($\boldsymbol{\theta}$) for the compositional covariates and TPR($\boldsymbol{\beta}, \boldsymbol{\zeta}$) and FPR($\boldsymbol{\beta}, \boldsymbol{\zeta}$) for the unconstrained covariates.

The proposed CAVI-MC method performs much better than the existing methods in terms of estimation with low to moderate dimensionality. When the signal is moderate or strong the CAVI-MC approach provides a more accurate estimation of the model, both in terms of a lower false positive rate (FPR) and L2 loss. The approach works well even in the presence of high correlation with sufficient signal. This can be seen in Table 1 for the “mixed” scenario with a SNR of 1, and in the full table of results in the Supplementary material.

The lasso approach fails to capture the sparsity of the true model in each of the scenarios. This characteristic is particularly obvious when $n >> p$. In Table 2, where the SNR is 1, $n = 100$ and $\rho = 0$, the FPR of the compositional covariates for the group lasso is 35%. For $\rho = 0.2$, the FPR is approximately 70%. The presence of correlation between the compositional covariates appears to make this problem worse.

When the true model contains both types of covariates, the two alternative approaches which fail to account for the compositional nature are easily outperformed by the CAVI-MC. The lasso methods suffer from high FPR even when the SNR is high and the correlation is low. The OLS approach struggles to identify the correct unconstrained covariates. This maybe due to the much larger variability in the true β compared with θ, despite similar means.

Each of the methods perform poorly when the SNR is low and the correlation is high. Where as the lasso approaches are inclined to include unnecessary variables in the model (leading to a very high FPR), the OLS and the CAVI-MC tend to exclude relevant variables resulting in low TPR, whilst maintaining low FPR. This increases the l_2 loss as the non zero parameter estimates shrink to zero. High correlation tends to magnify the problems with low SNR. The between-model moves in the CAVI-MC rely on a RJMCMC which is guided by independent pseudo updates. These are analogous to the OLS regression model, which tends to drop true positive variables from the model when the signal reduces and the correlation increases. When this happens the low signal is coupled with a poor guide for searching the large binary space for ξ parameter. This may explain why in Table 1 for $n = 100, d = 100$, the CAVI-MC has a TPR for θ below that of the group lasso approach.

A snapshot of the failings of all three approaches is provided by the plot of the ROC curves for a SNR of 0.5 in the "simple" scenario (Figure 1) where the red and green dots and blue cross represent the TPR and FPR of the CAVI-MC, lasso and OLS approach respectively. When the correlation increases from 0.2 to 0.5, the green dot shifts to the right as the FPR increases, where as the blue cross and red drop down as the TPR decreases. The CAVI-MC outperforms the two alternative approaches easily in the first two scenarios by combination of a very high TP and very low FP. When $\rho = 0.5$ the TPR of 0.72 for the CAVI-MC is not as large as the lasso but the FPR of 0.01 is two orders of magnitude lower than the lasso. Despite the lower TPR for $\rho = 0.5$ the parameter estimation of the CAVI-MC remains far more accurate, with a considerably lower L2 loss than the lasso.

5 Data

We apply our proposed method to a subset of the main study in Arkhangelsk, containing 515 men and women aged between 35-69 years recruited from the general population, from the "Know your Heart" cross-sectional study of cardiovascular disease (Cook et al., 2018). As part of the study, participants were asked to volunteer faecal samples for analysis of the gut microbiome. The relative abundances of the microbes were then determined by 16S rRNA sequencing (using the variable regions V3-V4) followed by taxonomic classification using a Naive Bayes Classifier (Bokulich et al., 2018). A baseline questionnaire captured unconstrained covariate information on age, sex and smoking status. Information on alcohol consumption from the questionnaire and biomarker data was used to derive a categorical factor with four levels on alcohol use.

The gut microbiome plays an important role in energy extraction and obesity (Tseng and Wu, 2019), which we illustrate by regressing body mass index (BMI) against the microbiome at the phylum and genus level alongside the unconstrained covariates. The counts are transformed into relative abundances after adding a small constant of 0.5 to replace the zero counts (Aitchison, 2003) and then log transformed. BMI is also log transformed and the continuous age covariate is standardised. The same CAVI-MC variational inference set up described in the simulation study is applied to each regression model and the ELBO is monitored to confirm convergence. Four separate CAVI-MC runs are performed at different initial starting points for the q expectations.

Thresholding the marginal expectation of the approximate posterior distributions at 0.5, we find an increase in Firmicutes (which has a -0.8 correlation with Bacteroidetes) and a decrease in Synergistetes is associated with an increase of BMI at the phylum level. At the genus level, BMI is increased by an increase in *Roseburia* and a reduction in *Oscillospira*. The corresponding marginal expectation of the approximating posterior $\mathbb{E}[q(\boldsymbol{\xi}|y)]$ is plotted in Figure 2. We also find BMI to be positively associated with age. The corresponding ELBO for each model clearly indicates an optimum has been reached (Figure 3), with each run finding the same local optimum.

Our findings appear to be consistent with previous studies. The ratio of Firmicutes to Bacteroidetes at the phylum level is considered to be a biomarker for obesity (Armougom et al., 2009, Davis, 2016). Increases in physical training of rats has led to an increase in their levels of Synergistetes (de Oliveira Neves et al., 2020). At the genus level Yuan et al., 2021 identifies *Roseburia* to be positively correlated with obesity in children, and Chen et al., 2020 determines *Oscillospira* to be negatively associated with BMI.

6 Discussion

Our Bayesian hierarchical linear log-contrast model estimated by mean field Monte Carlo co-ordinate variational inference improves regression modelling for compositional data. Sparse variable selection is performed through priors which fully account for the constrained parameter space associated with the compositional covariates. We introduce Monte Carlo expectations to approximate integrals which are not available in closed form. These expectations are obtained via RJMCMC with proposal parameters informed by approximating variational densities via auxiliary parameters with pseudo updates. As long as there is sufficient signal to guide the RJMCMC, the approach leads to an increase in the TPR and a reduction in the FPR.

The CAVI-MC suffers when the SNR is low and the correlation is high. Addressing the correlation by adapting the prior parametrisation may help to improve the model in these settings. One approach to address this issue is to use a Markov Random Field prior (Chen and Welling, 2012) which imposes a structure on the selection of $\boldsymbol{\xi}$. Zhang et al., 2020 use this prior to incorporate the phylogenetic relationship among the bacterial taxa alongside a model which partially accounts for the constraint on the parameters. Alternatively, to avoid having to pre-define the structure of the taxa, a Dirichlet Process could be used to account for the correlation of the microbiome by clustering the covariates (Curtis and Ghosh, 2011) prior to the regression.

At the genus level, despite the CAVI-MCMC identifying associations between the BMI and *Roseburia* and *Oscillospira*, some of the other microbiome features which have been found to be associated with BMI were not detected. *Bifidobacterium* has been found to be negatively associated with BMI in children (Ignacio et al., 2016). This taxon was also found to be associated with BMI in adults, alongside a negative association between BMI and *Methanobrevibacter* (Schwiertz et al., 2010). However, associations between BMI and the gut microbiome at the genus level are subject to a high degree of variation across studies (Verdam et al., 2013). This maybe partly explained by the tools used to construct the microbiome datasets, which can identify quite different results from the same sample (Nearing et al., 2021).

As genetic sequencing becomes more widely available, interest grows in modelling the relationship between the microbiome and a complex set of phenotypes such as blood concentrations of lipids or other metabolites. Bayesian hierarchical models have been introduced for multiple outcomes (Ruffieux et al., 2017, Lewin et al., 2016), which leverage shared information improving predictor selection. These approaches often use the simplifying assumption of conditionally independent residuals to allow different covariates to be associated with different responses. In future work, we would like to explore this multiple response extension to our model, using a hierarchical approach to allow information on the shared parameters to be pooled whilst incorporating correlation between the responses to aid variable selection.

7 Supplementary Material

Supplementary material which contains the derivations of all of the analytical updates for the CAVI-MC is available online.

Author Information

Affiliations

Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom

Darren Scott, Alex Lewin, Julian Libiseller-Egger, Jody Phelan and Taane Clark.

Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Ernest Diez Benavente.

Federal Research and Clinical Center of Physical-Chemical Medicine, Medicine, Moscow, Russia

Dmitry Federov, Elena Ilina and Polina Tikhonova.

Bioinformatics and Genomics Intercollege Graduate Program, Huck Institutes of Life Sciences, Pennsylvania State University, USA

Polina Tikhonova.

Northern State Medical University, Arkhangelsk, Russia

Alexander Kudryavstev.

Contributions

DS and AL developed the statistical methods outlined in the paper. All authors except AL and DS helped create the Know Your Heart dataset. DS drafted the manuscript with support from AL. All authors read and approved the manuscript.

Corresponding author

Correspondence to Darren Scott.

Declarations

Ethics approval and consent to participate

The Know Your Heart study, which generate the applied dataset used in the article, complies with the Declaration of Helsinki. The study was approved by the ethical committees of ethics committees of the London School of Hygiene & Tropical Medicine (approval number 8808 received 24.02.2015), Novosibirsk State Medical University (approval number 75 approval received 21/05/2015), the Institute of Preventative Medicine, Novosibirsk (no approval number; approval received 26/12/2014), and the Northern State Medical University, Arkhangelsk (approval number 01/01–15 received 27/01/2015).

Signed informed consent was obtained both at baseline interview and at the health check. At baseline interview the consent was obtained for passing on name, address, and telephone number to the polyclinic medical team for those deciding to have a health check. Agreement for interview per se was obtained verbally. At the health check written informed consent was obtained for participation in the study. Participants were given the option also to consent to be re-contacted by the study team in the future.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from the IPCDR Steering Group but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of IPCDR Steering Group. Software in the form of Python code, together with a sample input data set and complete documentation is available on request from the corresponding author.

Competing interests

None declared.

Funding

This work was supported by the UK Medical Research Council grant MR/N013638/1 and, MR/M013138/1 “Methods and tools for structural models integrating multiple high-throughput omics data sets in genetic epidemiology”. The approach is applied to data from the the Know Your Heart study, a component of International Project on Cardiovascular Disease in Russia (IPCDR) and funded by Wellcome Trust Strategic Award [100217], UiT The Arctic University of Norway (UiT), Norwegian Institute of Public Health, and Norwegian Ministry of Health and Social Affairs. The funding bodies had no role in the design of the study, data collection, analysis, interpretation of data, or in writing the manuscript.

Acknowledgments

Not applicable.

References

Aitchison, J. (1982). “The statistical analysis of compositional data”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2, pp. 139–177.

— (2003). *The Statistical Analysis of Compositional Data.* Blackburn Press: Caldwell, NJ, USA.

Aitchison, J. and J. Bacon-Shone (1984). “Log contrast models for experiments with mixtures”. In: *Biometrika* 71.2, pp. 323–330.

Aitchison, J and S M Shen (1980). “Logistic-normal distributions: some properties and uses”. In: *Biometrika* 67.2, pp. 261–272.

Armougom, Fabrice et al. (2009). “Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and methanogens in anorexic patients”. In: *PLoS ONE* 4.9, pp. 1–8.

Bishop, Christopher M and Markus Svensen (2003). *Bayesian hierarchical mixtures of experts.* UAI, pp. 57–64.

Blei, David M. et al. (2017). “Variational inference: a review for statisticians”. In: *Journal of the American Statistical Association* 112.518, pp. 859–877.

Bokulich, Nicholas A. et al. (2018). “Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin”. In: *Microbiome* 6.90, pp. 1–17.

Carbonetto, Peter and Matthew Stephens (2012). “Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies”. In: *Bayesian Analysis* 7.1, pp. 73–108.

Chen, Yi et al. (2020). “High Oscillospira abundance indicates constipation and low BMI in the Guangdong gut microbiome project”. In: *Scientific Reports* 10.1, pp. 1–8.

Chen, Yutian and Max Welling (2012). "Bayesian structure learning for markov random fields with a spike and slab prior". In: *Uncertainty in Artificial Intelligence - Proceedings of the 28th Conference, UAI 2012*, pp. 174–184.

Combettes, Patrick L. and Christian L. Müller (2021). "Regression models for compositional data: general log-contrast formulations, proximal optimization, and microbiome data applications". In: *Statistics in Biosciences* 13.2, pp. 217–242.

Cook, Sarah et al. (2018). "Know your heart: rationale, design and conduct of a cross-sectional study of cardiovascular structure, function and risk factors in 4500 men and women aged 35-69 years from two Russian cities." In: *Wellcome Open Research* 3, pp. 1–29.

Curtis, S. Mc Kay and Sujit K. Ghosh (2011). "A Bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression". In: *Journal of Statistical Theory and Practice* 5.4, pp. 715–735.

Davis, Cindy D. (2016). "The gut microbiome and its role in obesity". In: *Nutrition Today* 51.4, pp. 167–174.

de Oliveira Neves, Viviano Gomes et al. (2020). "High-sugar diet intake, physical activity, and gut microbiota crosstalk: implications for obesity in rats". In: *Food Science and Nutrition* 8.10, pp. 5683–5695.

Dellaportas, Petros et al. (2002). "On Bayesian model and variable selection using MCMC". In: *Statistics and Computing* 12.1, pp. 27–36.

Egozcue, J. J. et al. (2003). "Isometric logratio transformations for compositional data analysis". In: *Mathematical Geology* 35.3, pp. 279–300.

Franzosa, Eric A. et al. (2015). "Sequencing and beyond: integrating molecular 'omics' for microbial community profiling". In: *Nature Reviews Microbiology* 13, pp. 360–372.

George, Edward I. and Robert E. McCulloch (1993). "Variable selection via Gibbs sampling". In: *Journal of the American Statistical Association* 88.423, pp. 881–889.

George, Edward I. and Robert E. McCulloch (1997). “Approaches for Bayesian variable selection”. In: *Statistica Sinica* 1.7, pp. 339–373.

Gloor, Gregory B. et al. (2017). “Microbiome datasets are compositional: and this is not optional”. In: *Frontiers in Microbiology* 8, pp. 1–6.

Green, P J (1995). “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”. In: *Biometrika* 82.4, pp. 711–732.

Guan, Yongtao and Matthew Stephens (2011). “Bayesian variable selection regression for genome-wide association studies and other large-scale problems”. In: *Annals of Applied Statistics* 5.3, pp. 1780–1815.

Harrell, Frank E (2021). *rms: regression modeling strategies.*

Honkela, Antti and Harri Valpola (2005). “Unsupervised variational Bayesian learning of nonlinear models”. In: *Advances in Neural Information Processing Systems.*

Ignacio, A. et al. (2016). “Correlation between body mass index and faecal microbiota from children”. In: *Clinical Microbiology and Infection* 22.3, 258.e1–258.e8.

Jaakkola, Tommi S and Michael I Jordan (1997). “A variational approach to Bayesian logistic regression models and their extensions”. In: *Sixth International Workshop on Artificial Intelligence and Statistics.*

Koslovsky, Matthew D. et al. (2020). “A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes”. In: *Annals of Applied Statistics* 14.3, pp. 1471–1492.

Kuo, L. and B. Mallick (1998). “Variable selection for regression models”. In: *The Indian Journal of Statistics* 60.1, pp. 65–81.

Lamnisos, Demetris et al. (2009). “Transdimensional sampling algorithms for Bayesian variable selection in classification problemswith many more variables than observations”. In: *Journal of Computational and Graphical Statistics* 18.3, pp. 592–612.

— (2013). “Adaptive Monte Carlo for Bayesian variable selection in regression models”. In: *Journal of Computational and Graphical Statistics* 22.3, pp. 729–748.

Leng, Chenlei et al. (2014). “Bayesian adaptive lasso”. In: *Annals of the Institute of Statistical Mathematics* 66.1, pp. 221–244.

Lewin, Alex et al. (2016). “MT-HESS: An efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues”. In: *Bioinformatics* 32.4, pp. 523–532.

Li, Hongzhe (2015). “Microbiome, metagenomics, and high-dimensional compositional data analysis”. In: *Annual Review of Statistics and Its Application* 2, pp. 73–94.

Li, Qiwei et al. (2019). “Bayesian modeling of microbiome data for differential abundance analysis”. In: *arXiv:1902.08741*.

Lin, Wei et al. (2014). “Variable selection in regression with compositional covariates”. In: *Biometrika* 101.4.

Nearing, Jacob T et al. (2021). “Microbiome differential abundance methods produce disturbingly different results across 38 datasets”. In: *bioRxiv* 13.1, p. 342.

Nott, David J. and Robert Kohn (2005). “Adaptive sampling for Bayesian variable selection”. In: *Biometrika* 92.4, pp. 747–763.

Ormerod, J. T. and M. P. Wand (2010). “Explaining variational approximations”. In: *American Statistician* 64.2, p. 154.

Park, Trevor and George Casella (2008). “The Bayesian lasso”. In: *Journal of the American Statistical Association* 103.482, pp. 681–686.

R, Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Ruffieux, Helene et al. (2017). “Efficient inference for genetic association studies with multiple outcomes”. In: *Biostatistics* 18.4, pp. 618–636.

Schwiertz, Andreas et al. (2010). “Microbiota and SCFA in lean and overweight healthy subjects”. In: *Obesity* 18.1, pp. 190–195.

Sender, Ron et al. (2016). “Revised Estimates for the Number of Human and Bacteria Cells in the Body”. In: *PLoS Biology* 14.8, pp. 1–14.

Shi, Pixu et al. (2016). “Regression analysis for microbiome compositional data”. In: *Annals of Applied Statistics* 10.2, pp. 1019–1040.

Tseng, Ching Hung and Chun Ying Wu (2019). “The gut microbiome in obesity”. In: *Journal of the Formosan Medical Association* 118, S3–S9.

Verdam, Froukje J. et al. (2013). “Human intestinal microbiota composition is associated with local and systemic inflammation in obesity”. In: *Obesity* 21.12, pp. 607–615.

Xu, Xiaofan and Malay Ghosh (2015). “Bayesian variable selection and estimation for group lasso”. In: *Bayesian Analysis* 10.4, pp. 909–936.

Yang, Yi et al. (2020). *gglasso: group lasso penalized learning using a unified BMD algorithm, R package.*

Ye, Lifeng et al. (2020). “Monte Carlo co-ordinate ascent variational inference”. In: *Statistics and Computing* 30, pp. 887–905.

Yuan, Xin et al. (2021). “The role of the gut microbiota on the metabolic status of obese children”. In: *Microbial Cell Factories* 20.1, pp. 1–13.

Zhang, Liangliang et al. (2020). “Bayesian compositional regression with structured priors for microbiome feature selection”. In: *Biometrics* 77.3, pp. 824–838.

Tables

Table 1: Subset of the results from the "mixed" scenario with SNR 1 for $d = 100$ compositional covariates, $G = 24$ categorical covariates, for the variational Bayes (VB) and group lasso approach. The true positive and false positive rates for the unconstrained and constrained covariates are reported alongside the L2 loss of the estimated parameters (2 decimal places).

n	ρ	Method	TPR	FPR	TPR(θ)	FPR(θ)	TPR(β, ζ)	FPR(β, ζ)	L2
100	0	VB	0.99	0.00	1.00	0.00	0.98	0.01	0.94
		GLasso	0.77	0.20	1.00	0.20	0.60	0.19	5.71
100	0.2	VB	0.99	0.00	1.00	0.00	0.98	0.01	0.99
		GLasso	0.74	0.65	0.96	0.71	0.57	0.58	2.79
100	0.5	VB	0.36	0.00	0.26	0.00	0.48	0.00	9.69
		GLasso	0.68	0.27	0.89	0.21	0.53	0.21	4.28
200	0	VB	1.00	0.00	1.00	0.00	1	0.00	0.37
		OLS	0.68	0.00	1.00	0.00	0.43	0.00	4.57
		GLasso	1.00	0.30	1.00	0.32	1.00	0.23	4.06
200	0.2	VB	1.00	0.00	1.00	0.00	1.00	0.01	0.40
		OLS	0.67	0.00	1.00	0.00	0.42	0.00	4.65
		GLasso	0.99	0.35	1.00	0.37	0.98	0.29	2.53
200	0.5	VB	1.00	0.00	1.00	0.00	1.00	0.00	0.02
		OLS	0.68	0.00	1.00	0.00	0.44	0.00	5.16
		GLasso	1.00	0.33	1.00	0.33	1.00	0.30	2.74

Table 2: Table of true positive rate, false positive rate and the L2 loss of the estimated parameters for the true model with elements $\boldsymbol{\theta}$ as the only significant parameter for the VB approach, OLS and group lasso for a SNR of 1. The total number of compositional, continuous and categorical covariates are represented by d, p and G respectively.

$(n, d, p+G)$	ρ	Method	TPR	FPR	L2 loss
(100, 45, 24)	0	VB	1.00	0.00	0.08
		OLS	0.94	0.08	2.32
		GLasso	0.98	0.35	3.86
(100, 45, 24)	0.2	VB	1.00	0.01	0.04
		OLS	0.97	0.16	2.13
		GLasso	0.99	0.68	3.63
(100, 45, 24)	0.5	VB	0.94	0.00	0.39
		OLS	1.00	0.16	2.41
		GLasso	1.00	0.62	3.84
(200, 100, 24)	0	VB	1.00	0.00	0.03
		OLS	0.99	0.00	0.23
		GLasso	1.00	0.22	0.16
(200, 100, 24)	0.2	VB	1.00	0.00	0.03
		OLS	1.00	0.00	0.13
		GLasso	1.00	0.15	0.13
(200, 100, 24)	0.5	VB	1.00	0.00	0.02
		OLS	1.00	0.00	0.88
		GLasso	1.00	0.23	0.25

Figures

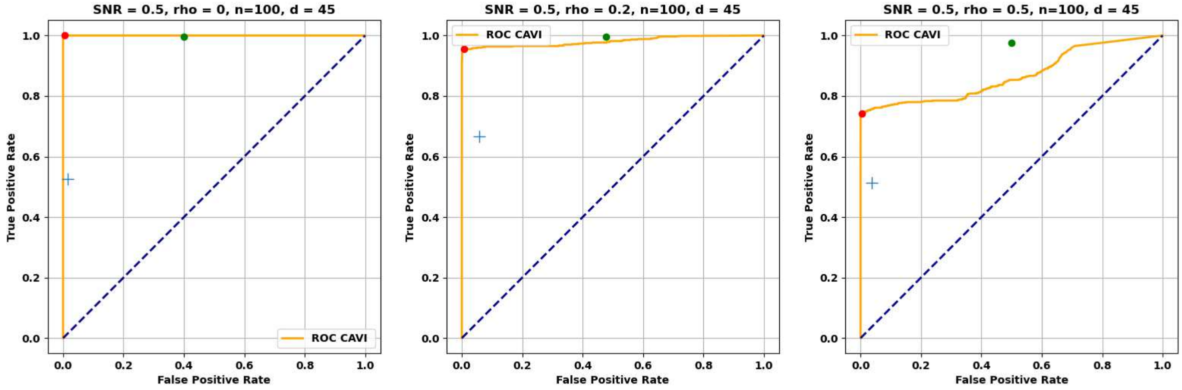

Figure 1: Plot of the ROC curves for the CAVI-MC from the "simple" scenario for a SNR of 0.5. The red and green dots and blue cross represent the TPR and FPR of the CAVI-MC, lasso and OLS respectively.

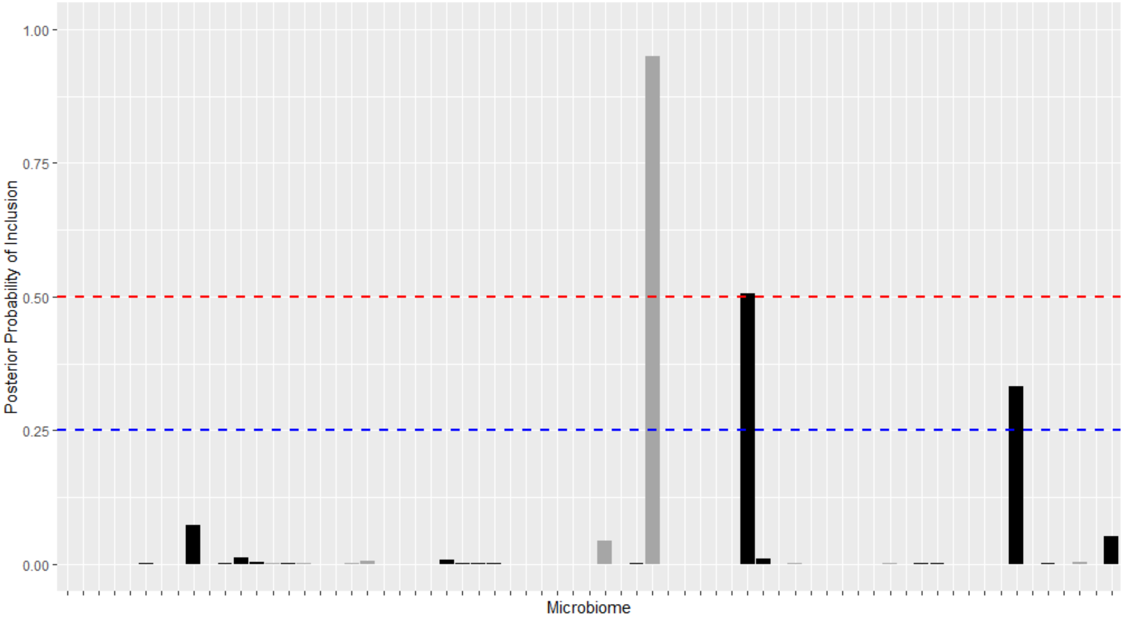

Figure 2: Plot of the marginal expectation of the approximating posterior $\mathbb{E}_q[p(\boldsymbol{\xi}|\mathbf{y})]$ at the genus level. The grey denotes a positive θ_j, black a negative θ_j. The bars above 0.25 probability of inclusion (blue dashed line) are *Roseburia*, *Oscillospira* and *Oxalobacter* respectively. The red dashed line at 0.5 probability of inclusion indicate the thresholding value used to determine a significant association.

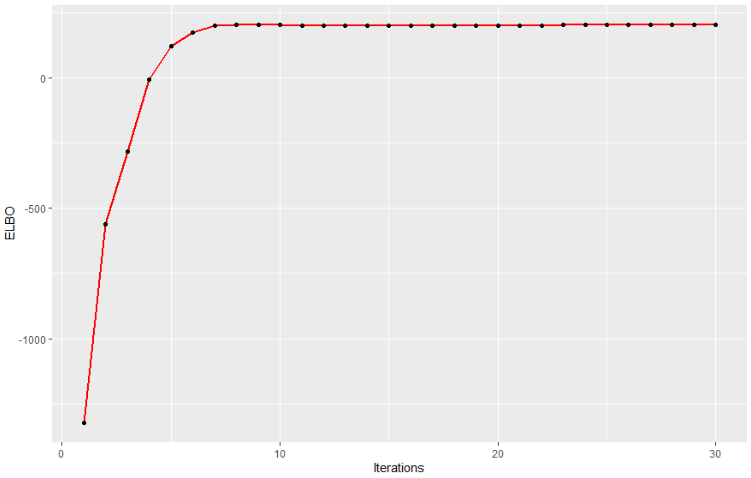

Figure 3: Plot of the ELBO against iterations for the CAVI-MC applied to the "Know Your Heart" data set with the microbiome grouped at the genus level. 30 iterations are performed, with 30,000 between state space moves by the RJMCMC after 4 iterations. The approximate straight line after only 7 iterations implies that the model has reached convergence. Despite the MCMC component removing the monotonic properties of the ELBO, the fluctuations are relatively small.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BMCBioinformaticsBayesiancompositionalregressionwithmicrobiomefeaturesviavariationalinferenceSupplementaryMaterial.pdf