A Bayesian Multi-Task Approach for Detecting Global Microbiome Associations

Farhad Hatami[1], Emma Beamish[3], Rachael Rigby[2] and Frank Dondelinger[1]
[1]Centre for Health Informatics, Computation and Statistics, Lancaster Medical School, Lancaster University
[2]Biomedicine and Life Sciences, Lancaster University
[3]Department of Musculoskeletal Biology, Institute of Ageing and Chronic Disease, University of Liverpool

Abstract

The human gut microbiome has been shown to be associated with a variety of human diseases, including cancer, metabolic conditions and inflammatory bowel disease. Current statistical techniques for microbiome association studies often rely on a measure of ecological distance, or on detecting associations with individual bacterial species. In this work, we develop a novel, Bayesian multi-task approach for detecting global microbiome associations. Our method is not dependent on a choice of distance measure, and is able to incorporate phylogenetic information about microbial species. We apply our method to simulated data and show that it allows for consistent estimation of global microbiome effects. Additionally, we investigate the performance of the model on a real-world microbiome study in inflammatory bowel disease (Beamish, 2017) and show that we can use it to detect microbiome-metabolome associations.

1 Introduction

Recent years have seen an explosion in the amount of genomic sequencing data collected on a variety of organisms. One area of particular interest is the nascent field of metagenomics. Metagenomics is concerned with the study of genetic material that can be found in samples from diverse environments ranging from soil and water to body cavities (Wooley et al., 2010). The collection of genetic material identified in a sample is called the microbiome. For human health, the human gut microbiome has been shown to be associated with a variety of conditions, including colorectal cancer (Ahn et al., 2013), metabolic diseases (Le Chatelier et al., 2013) and inflammatory bowel disease (Halfvarson et al., 2017).

The recent glut of microbiome studies has led to increased interest in developing robust and efficient statistical methods for analysing this type of data. A typical microbiome study will first sequence the genetic material contained in a sample, and will then use the genetic information to identify and quantify the organisms that the material originated from. In the case of the microbiome, the latter is most commonly done by using the 16S ribosomal RNA genes, which allow for the identification of bacterial species that colonize the human gut. Each fragment of 16S genetic material in the sample can then be classified by comparing it to known reference genomes, and species can be differentiated from each other by looking at genetic similarity. The result will be a count matrix counting the number of times genetic material from each species was detected.

Previous approaches to analysing microbiome data can be classified in two main categories: distance-based methods and regression-based methods. In distance-based methods we define a distance metric between observed microbiome samples. Common measures include Jaccard (dis-)similarity, Bray-Curtis dissimilarity and the UniFrac distance (Lozupone and Knight, 2005). The latter has the advantage that it takes into account the phylogeny, or genetic relatedness, of the species in the samples when calculating the distance; loosely speaking, samples with related species will be closer than samples with unrelated species. One can then use a statistical test such as the PERMANOVA to test whether the centroids and dispersion of samples within two or more groups are identical (see e.g. Chen et al., 2012).

Regression-based methods tend to treat the count data as following either a negative binomial (Zhang et al., 2017) or Dirichlet-multinomial distribution (Chen and Li, 2013). In the former, the effects of observed covariates on species are usually treated as independent, although variance parameters may be shared across species. In the latter, the covariance across species is

defined by the Dirichlet-multinomial distribution, but the regression will still define individual parameters representing the association between each species and the observed covariate. Unlike distance-based methods, regression methods do not try to infer a global effect of a covariate on the microbiome. However, they are better able to deal with multiple continuous covariates, and are not sensitive to a choice of distance measure.

In our work, we combine the advantages of the regression and distance-based approaches. Our idea is to develop a Bayesian multi-task log-ratio regression model, where the multi-task nature of the model allows us to model multiple species as multiple outcomes. Unlike many other regression approaches for microbiome data, we do not model the species counts directly, but rather treat the data as proportional and apply the log-ratio transform (Aitchison, 1982) to obtain multivariate-Gaussian distributed pseudo-observations. This allows us to bring in the phylogenetic information via the covariance matrix. Global associations between the microbiome and the covariates are estimated using a shared hyper-prior on the species-specific effect estimates. We demonstrate the robustness and efficiency of our method using an extensive simulation study, and apply it to a case study in inflammatory bowel disease.

2 Methods

2.1 Log Ratio Model

Let us assume that we are collecting microbiome data from N people, and in each sample we are counting the occurrence of S species or taxons. We treat the microbiome data as compositional, i.e., even though we are observing count data, we only consider the proportion of each species. If we observe an S-dimensional count vector $\mathbf{z}_i$ for sample i, then the proportion of species s can be estimated as $\rho_{i,s} = \frac{z_{i,s}}{\sum_{s'}^{S} z_{i,s'}}$. We can now transform the species proportions for species $s \in \{1, ..., S-1\}$ by using the log-ratio transform (Aitchison, 1982):

$$y_{i,s} = \log\left(\frac{\rho_{i,s}}{\rho_{i,S}}\right) \tag{1}$$

We assume that the $\mathbf{y}_i$ are approximately multivariate Gaussian distributed, and model them as:

$$\mathbf{y}_i \sim \mathcal{MVN}(\boldsymbol{\mu}_i, c\boldsymbol{\Sigma}) \tag{2}$$

In our application, $\boldsymbol{\Sigma}$ is informed by the phylogenetic tree inferred from the sequencing data. Briefly, under the assumption of a Brownian motion evolutionary model, the covariance of two species is equivalent to the branch length from the root to the most recent common ancestor in the phyogenetic tree. If no phylogenetic tree is available, then we set $\boldsymbol{\Sigma} = \mathbf{I}$, the $(S-1) \times (S-1)$ identity matrix. The hyperparameter c, $c > 0$ is a scaling parameter controlling the desired fit to the data. For the data in this study we found that $c = 1$ led to satisfactory models that showed no evidence of over- or under-fitting.

2.2 Multi-task regression

Having established the log ratio model, we now want to formulate $\boldsymbol{\mu}_i$ as a linear expression with information sharing across the species ("tasks", or response variables). Let:

$$\boldsymbol{\mu}_i = (\boldsymbol{\Gamma} \odot \mathbf{B})\mathbf{x}_i^T \tag{3}$$

where $\mathbf{x}_i$ is a vector of p observed covariates for subject i, $\mathbf{B}$ is an $(S-1) \times p$ matrix of positive linear coefficients and $\boldsymbol{\Gamma}$ is an $(S-1) \times p$ matrix with entries in $\{-1, 1\}$. Estimation of global effects is achieved via a prior on the elements of $\boldsymbol{\beta}_j$, the $S-1$-dimensional vectors corresponding to the columns in $\mathbf{B}$:

$$\beta_{s,j} \sim \mathcal{N}^+(a_j, \sigma^2) \; \forall s \in \{1, ..., S-1\} \tag{4}$$

where a_j denotes the global effect of covariate j on all species. Here $\mathcal{N}^+$ denotes the positive part of the Gaussian distribution truncated at zero. Note that the global effect only defines the prior for the strictly positive magnitude $\beta_{s,j}$ of the species-specific effects. Placing a prior on the signed effect size would not have the desired effect; since the size of each biological sample is fixed, an increase in some species must lead to a decrease in other species, and so the global effect size under such an alternative model would not be reflective of the change across species.

We also need to define a distribution for the elements $\gamma_{s,j}$ of $\boldsymbol{\Gamma}$. The natural approach would be to treat them as binary variables following a Bernoulli distribution. However, this requires us to perform inference over discrete variables, which can be onerous, and does not lend itself well to gradient-based inference techniques. Instead, we opt for a Gumbel-Softmax approximation to the Bernoulli distribution (Jang et al., 2016; Maddison et al., 2016). The Gumbel-Softmax is a continuous relaxation of the categorical distribution. In general, for a categorical random variable h with class probabilities π_1, π_2, ..., π_K, we can represent h as:

$$\mathbf{h} = \text{one-hot}(\arg\max_k \left[g_k + \log \pi_k\right]) \tag{5}$$

where one-hot denotes the one-hot encoding of a categorical variable as a K-dimensional vector (hence the change from h to $\mathbf{h}$). The g_i are sampled from a Gumbel(0,1) distribution. Note that this representation is

exactly equivalent to a categorical distribution; to obtain the continuous relaxation we need to replace the arg max by a softmax function, which gives:

$$\delta_k = \frac{\exp\left((\log(\pi_k) + g_k)/\tau\right)}{\sum_{k'=1}^{K} \exp\left((\log(\pi_{k'}) + g_{k'})/\tau\right)} \tag{6}$$

for $k \in \{1, ..., K\}$, where $\boldsymbol{\delta} \in \Delta^{K-1}$, the $(K-1)$-dimensional simplex. Here τ is a tuning parameter that controls the smoothness, with $\tau = 0$ reverting to the discrete case. In our case, $K = 2$, which means that eq. (6) simplifies to:

$$\delta_{s,j} = \frac{\exp\left((\log(\pi_1) + g_1)/\tau\right)}{\exp\left((\log(\pi_1) + g_1)/\tau\right) + \exp\left((\log(\pi_{-1}) + g_{-1})/\tau\right)} \tag{7}$$

where $\pi_{-1} = 1 - \pi_1$. We drop the subscript k as one value of δ fully defines the two-dimensional vector on the simplex, but introduce the subscript (s, j) to denote the value of δ for species s and covariate j. We use the subscripts 1 and -1 to denote Gumbel draws for $\gamma_{s,j} = 1$ and $\gamma_{s,j} = -1$ respectively. Since $\delta_{s,j}$ is either 0 or 1, we set:

$$\gamma_{s,j} = 2\delta_{s,j} - 1 \tag{8}$$

This has the desired property that if $\delta_{s,j} \approx 1$, $\gamma_{s,j} \approx 1$ and if $\delta_{s,j} \approx 0$, $\gamma_{s,j} \approx -1$.

We specify truncated Gaussian $\mathcal{N}^+(0, 1)$ priors for the global effects a_j, which also serves to regularize the model when the sample size is smaller than the number of parameters, which is the case for almost all the models studied here. We further place a Beta prior on the parameter π_1, where the hyperparameters can be used to encode any prior beliefs about the prevalence of positive or negative associations; for our study we set both equal to 1.

2.3 Inference

We are interested in performing fully Bayesian posterior inference over the parameters of this model. A convenient framework is Hamiltonian Monte Carlo (HMC), which allows us to take advantage of the differentiability of the Gumbel-Softmax to estimate gradients and speed up the Bayesian inference. We employ the Stan sampler to perform HMC sampling; for more details on the the Stan sampler and modelling language see Carpenter et al. (2017).

In microbiome research, it is common to encounter datasets consisting of large numbers of operational taxonomic units (OTUs). Despite the efficiency of HMC sampling, sampling hundreds or thousands of parameters $\beta_{s,j}$ and $\gamma_{s,j}$ is not currently feasible on a personal computer. We therefore propose an approximate estimation, as an alternative to the full log-ratio model, where we only estimate $\beta_{s,j}$ as part of the sampling, and the $\gamma_{s,j}$ parameters are estimated in a separate inference step. The approximate inference procedure is as follows:

1. Fit independent linear models $y_{i,s} = \mathbf{x}_i^T \boldsymbol{\beta}_s^* + \epsilon_{i,s}$, $\epsilon_{i,s} \sim \mathcal{N}(0, \sigma_s^2)$ to the log-transformed microbiome counts for all species s. Note that $\boldsymbol{\beta}_s^*$ is distinct from the rows of $\mathbf{B}$ in the full model.
2. Set $\gamma_{s,j} = \text{sign}(\beta_{s,j}^*)$ where the sign() operation returns -1 if $\beta_{s,j}^*$ is negative and 1 otherwise.

We refer to the model with $\boldsymbol{\Gamma}$ inferred from the data as the full log-ratio model, and to the model with $\boldsymbol{\Gamma}$ fixed according to the two-step procedure above as the conditional model, since inference is conditional on the estimate of $\boldsymbol{\Gamma}$. We can think of the conditional model as an approximation to the full joint model $P(\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}) = P(\mathbf{B}, \boldsymbol{\theta}|\boldsymbol{\Gamma}, \mathbf{Y}, \mathbf{X})P(\boldsymbol{\Gamma}|\mathbf{Y}, \mathbf{X})$, where $\boldsymbol{\theta}$ collects all parameters not in $\mathbf{B}$ or $\boldsymbol{\Gamma}$, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent the log-ratio response and the design matrix with the observed covariates, respectively. Instead of estimating $P(\boldsymbol{\Gamma}|\mathbf{Y}, \mathbf{X})$, which would require marginalising over all other parameters, we use $\hat{\boldsymbol{\Gamma}} = \text{sign}(\hat{\boldsymbol{\beta}}^*)$ where $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_S^*)$ is the maximum likelihood estimate of the effect sizes in independent linear models of $P(\mathbf{y}_s|\mathbf{x}_s, \boldsymbol{\beta}_s^*)$ for each s. Inference is performed on the conditional model $P(\mathbf{B}, \boldsymbol{\theta}|\boldsymbol{\Gamma}, \mathbf{Y}, \mathbf{X})$, which is less computationally expensive.

3 Simulation Study

3.1 Setup

We perform a simulation study under different scenarios to test the performance of the full and conditional log-ratio models. In particular, we study how well the full model can predict the global effect parameters a_j under scenarios with different sample size n, variance parameter σ^2, and number of species (or taxonomic units) S and compare the performance to that of the conditional model. In each experiment the variables that are not changing are set to $n = 200$, $\sigma^2 = 0.1$ and $S = 20$. The global effect sizes in the simulation model are set fixed with $a_j \in \{0, 1, 2\}$.

For each unique setting of the simulation parameters, we generate 20 random datasets according to the following procedure. First, a random phylogenetic tree is generated using the `ape` R package (Paradis and Schliep, 2018). Based on this tree, the covariance matrix $\boldsymbol{\Sigma}$ under the Brownian motion evolutionary model is calculated. Next, we generate an $N \times p$ matrix of covariates $\mathbf{X}$ with $x_{i,j} \sim \mathcal{N}(0, 1)$. For each covariate j and species s, $\gamma_{s,j}$ are generated from a

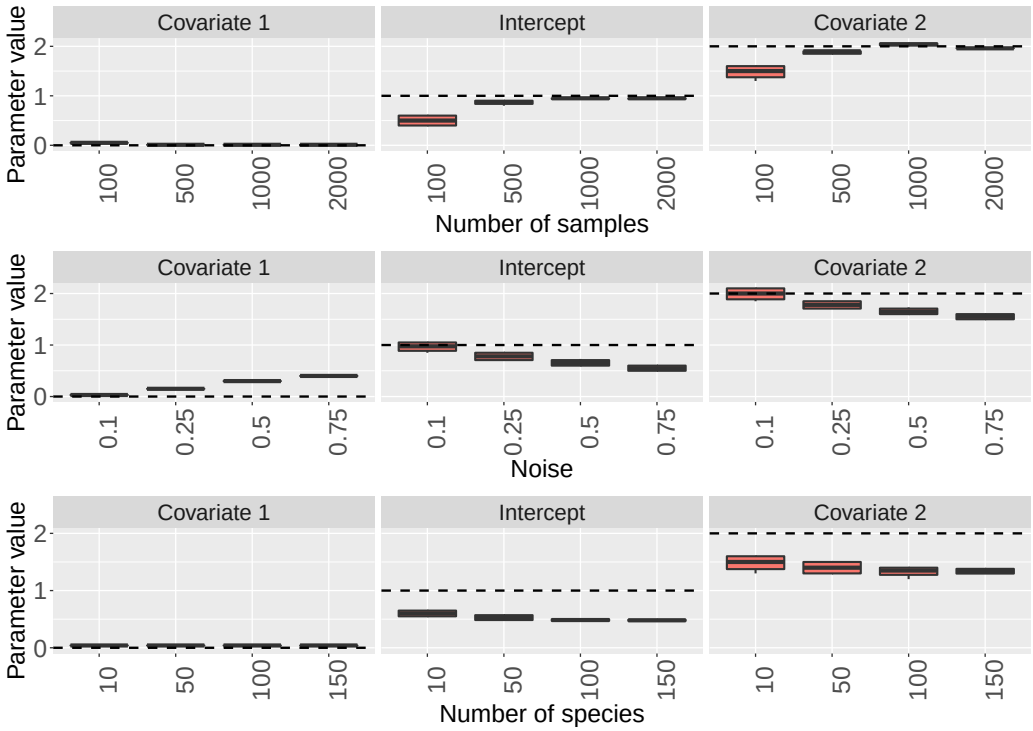


Figure 1: Performance of the full log-ratio model on simulated data when varying the sample size (top row), noise variance of the species-specific effects (middle row) and number of species or taxonomic units (bottom row). The linear simulation model consists of the intercept term and two covariates with global effect sizes 1, 2 and 0 respectively. The dashed horizontal lines show the size of the true effect, and the boxplots represent the posterior mean estimates obtained from 20 independent simulated datasets.

binomial distribution with $p = 0.5$. Species-specific coefficients $\beta_{s,j}$ are generated according to eq. (4). Then we simply calculate $\boldsymbol{\mu}_i$ according to eq. (3), and sample $\boldsymbol{y}_i$ according to eq. (2). The $\rho_{i,s}$ must sum to 1, and, for $s \in \{1, ..., S-1\}$, must be equal to $\rho_{i,S} \exp(y_{i,s})$ by eq. (1). To achieve this we set $\rho_{i,S} = 1/\sum_s^{S-1} \exp(y_{i,s})$. We can then calculate the $\rho_{i,s}$ and simulate count data $\mathbf{z}_i$ from a multinomial distribution with parameters $\boldsymbol{\rho}_i$, and number of trials equal to $200 * S$.

The bench-marking platform used for this study ran R-3.6.1 to generate datasets and invoke `rstan`. Models were fitted using the Stan software (Carpenter et al., 2017) version 2.19.1. In order to validate our models, a comparison with the generalized uniFrac method using the `GUniFrac` R package (Chen and Chen, 2018) has been done. The sampling was executed in parallel using 4 different chains each with 8000 iteration with a stopping criteria at convergence on 4 Intel Ivy Bridge cores each running at 1.15GHz speed and 12GB of RAM memory in total.

3.2 Full Log-Ratio Model

Figure 1 shows the performance of the full log-ratio model for three different experiments: changing the sample size, noise variance on the $\beta_{s,j}$ parameters, and number of species. Dashed lines represent the true values of the global effect (a_j) in the simulation model. It can be seen that increasing the number of samples n improves the accuracy of prediction and reduces the variance. On the other hand, increasing the variance parameter for the effect values and increasing the number of species (or taxonomic units) both make estimation of the global effect parameters harder. The only exception to this is the estimation of the association with covariate 1 (with global effect size 0) which is seen to improve as the number of species increases. We hypothesise that a larger number of species allows for better identification of zero effect sizes because we do not rely on identifying the $\gamma_{s,j}$ parameters; so while adding a species results in 2 additional parameters per non-zero covariates, it only results in one additional parameter for a zero covariate, while providing additional data to identify the global parameter a_j.

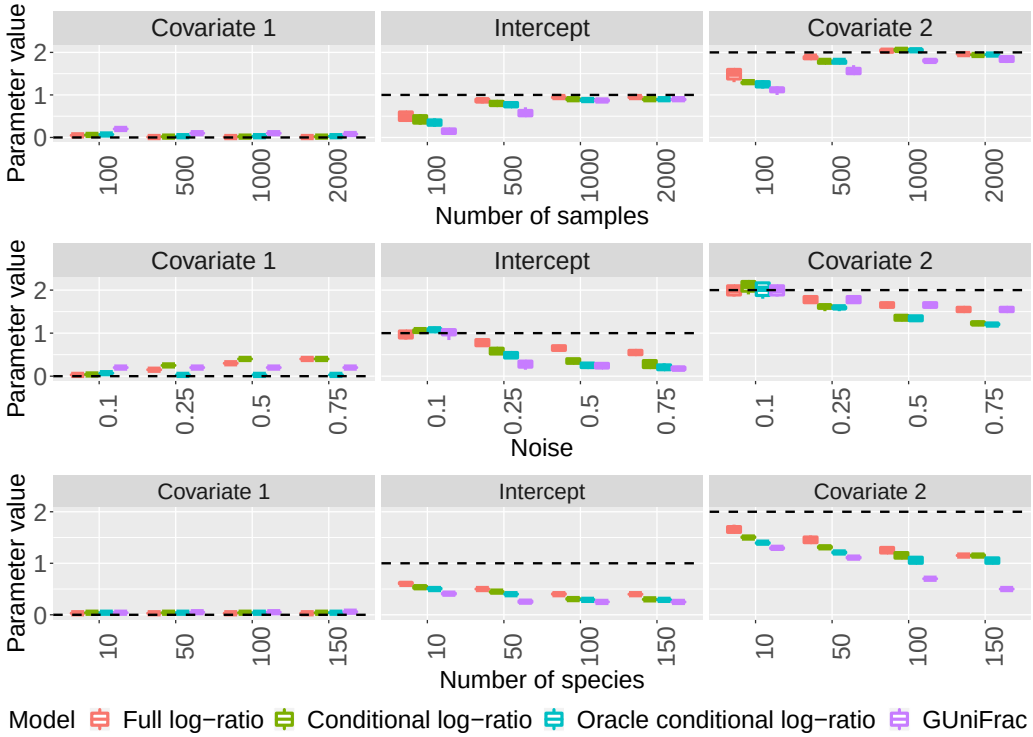


Figure 2: Comparison of the performance of the full, conditional and oracle conditional log-ratio models, and the PERMANOVA procedure based on the generalized UniFrac distance (Chen et al., 2012) on simulated data when varying the sample size (top row), noise variance of the species-specific effects (middle row) and number of species or taxonomic units (bottom row). The linear simulation model consists of the intercept term and two covariates with global effect sizes 1, 2 and 0 respectively. The dashed horizontal lines show the size of the true effect, and the boxplots represent the posterior mean estimates obtained from 20 independent simulated datasets.

3.3 Conditional Model

In this section we investigate the performance of the conditional log-ratio model, which uses a linear marginal model for each species to determine the species-specific effect signs $\gamma_{s,j}$ and then infers the size of the species-specific effects $\beta_{s,j}$ and the global effect a_j conditional on the $\gamma_{s,j}$. We compare this model to the full log-ratio model, and to a model that we call the "oracle conditional model", where the $\gamma_{s,j}$ parameters are set to the true values used for simulating the data. Additionally, we compare to a PERMANOVA test using the generalized UniFrac distance (Chen et al., 2012).

Figure 2 shows a comparison between the performance of these models. We see that there is not much difference between the model performance for the three log-ratio models. However, all three models outperform the generalized UniFrac distance, in particular when increasing the number of species. This is not terribly surprising, as the generalized UniFrac approach estimates a single effect parameter using a distance measure based on all species; by comparison, our method uses a hierarchical model to estimate both global effects and species-specific effect, making it more flexible.

We also investigate the computational efficacy of the full and conditional log-ratio models. Figure 3 shows the run time to convergence under each of the models. We see that the sampler for the conditional model takes longer for large S.

Taken together, Figures 2 and 3 indicate that for small sample sizes ($n < 100$) one could use either the full or conditional log-ratio models. For large sample sizes the conditional model could be deployed to improve computational efficiency, but there is a trade-off in terms of estimation accuracy. For large number of species S ($S > 50$), the full log-ratio model seems more efficient than the conditional model, perhaps because it can take advantage of Hamiltonian paths in the posterior landscape that are closed off when conditioning

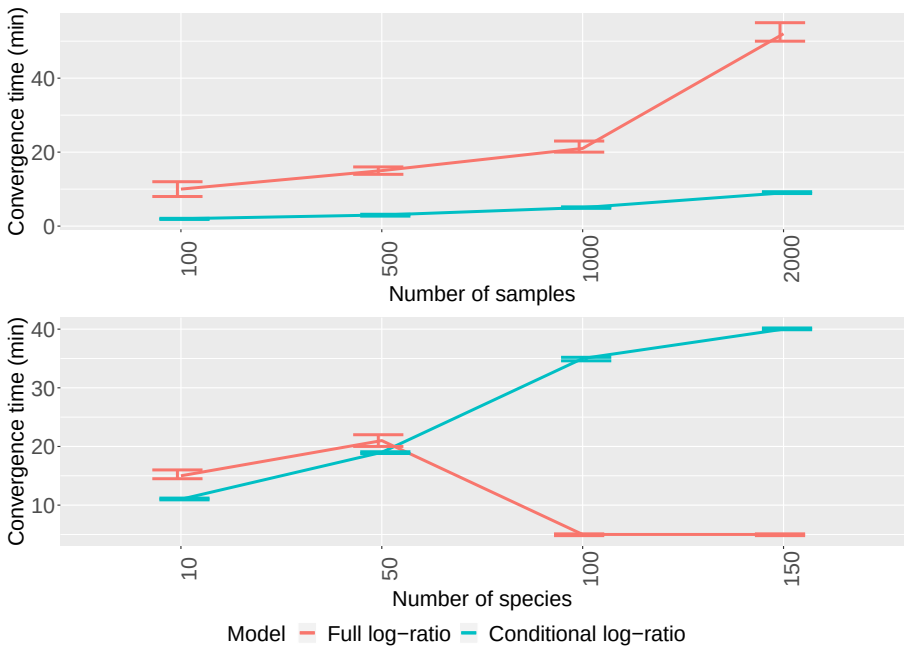


Figure 3: Comparison of the the computational time to convergence of the full and conditional log-ratio models, on the simulated data when increasing sample size n and number of species (or taxonomic units) S. The error bars show the 95% confidence interval calculated over 20 independent simulated dataset.

on $\mathbf{\Gamma}$.

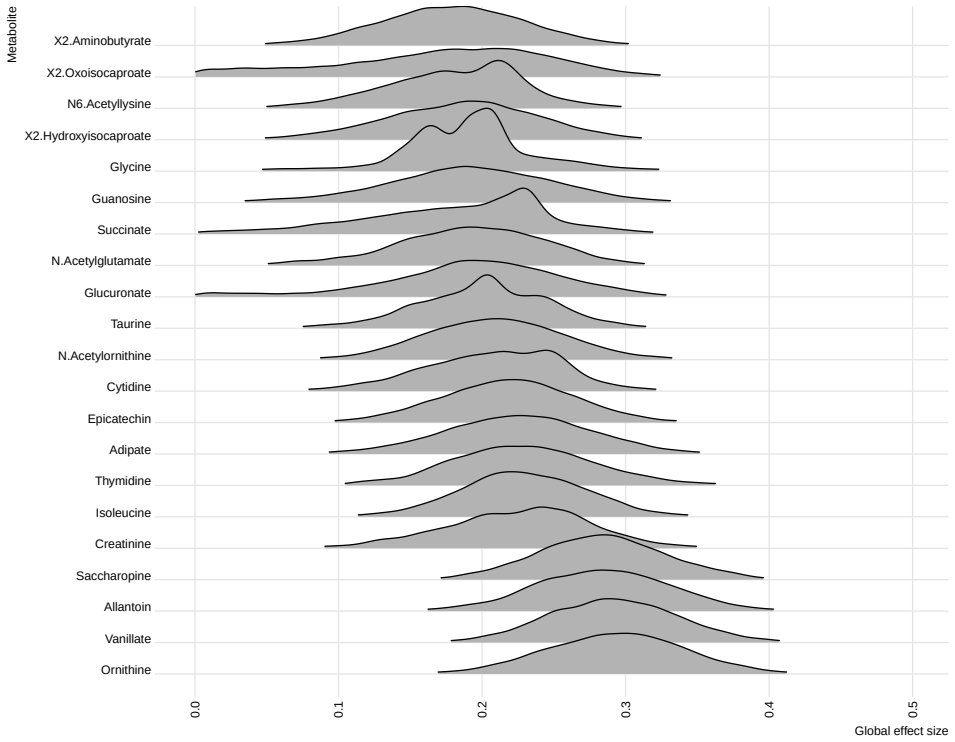
4 Real-World Application: Detecting Global Microbiome-Metabolome Associations

Using metabolomics as a measure of health, or to predict response to pharmacologics is a very attractive prospect, as measuring metabolites in urine, for example, may provide a non-invasive and potentially inexpensive screening method that has potential for application to a plethora of pathologies.

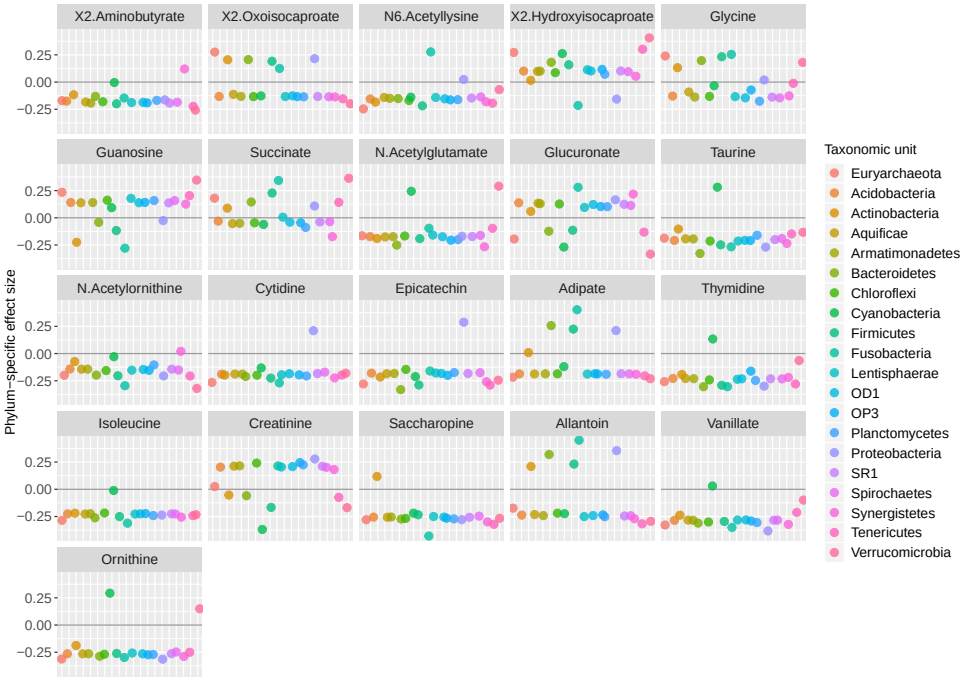
We apply the full log-ratio model to a dataset collected as part of a pilot study into inflammatory bowel disease (IBD) (Beamish, 2017). IBD is an umbrella term used to describe two conditions (Crohn's disease and ulcerative colitis) that are characterized by chronic inflammation of the gastrointestinal tract. 42 gut bacteria samples were obtained from patients undergoing colonoscopy at Lancashire Teaching Hospitals Trust (LTHTR) and University Hospitals of Morecambe Bay Trust (UHMBT), and metagenomic profiling was performed to obtain the bacterial species counts. The patients were additionally asked to provide a urine sample, and metabolite concentrations found in those samples were recorded. One objective of the study was to investigate associations between the gut microbiome and the urinary metabolite profiles.

After quality control, our cohort consists of $n = 40$ people (IBD cases and controls), for which we have measured a set of 3 clinical covariates: age, sex, and the hospital where the sample was collected. Note that we do not investigate any associations with IBD status, as number of cases was too small. Instead, we focus on detecting associations between the microbiome and individual metabolites. We consider the set of the $S = 21$ most abundant operational taxonomic units (OTUs or species) at the phylum level. As a result, our design matrix $\mathbf{X}$ is a 40×5 matrix consisting of the intercept, the metabolite level, age, sex and sample collection location. As the data is limited, we do not try to infer the variance parameter σ^2, but keep it constant and equal to 0.1. Finally, we have a total of 261 metabolites.

Figure 4a shows the posterior distribution of estimated global effect size a_j for the metabolites with the largest posterior mean effect estimate. We see that out of these 21 metabolites, saccharopine, allantoin, vanillate and ornithine have the strongest association with the microbiome. Some metabolites show interesting bimodal patterns in their posterior effect distribution; this is particularly obvious for glycine. Such patterns could potentially be indicative of more than one mechanism that links the microbiome to metabolite levels.



(a)



(b)

Figure 4: Application of the log-ratio model to a microbiome-metabolme association study (Beamish, 2017). (a) Density plots showing the posterior distribution of the global effect size associating each metabolite to changes in the microbiome. (b) Posterior mean estimates of the species-specific effect size for each metabolite across the different bacterial phyla.

Figure 4b shows the posterior mean estimate of the species-specific effect sizes $\beta_{s,j}$, specifically the association of each phylum with each metabolite. We note that for the metabolites with the largest posterior effects, the majority of species-specific effects are strongly negative, indicating that these metabolites are associated with a decrease in most phylum counts, compensated by an increase in just a couple of phyla.

We note that our analysis identified several phyla-associated metabolites that have previously been shown to be linked to markers of intestinal health or nutrition, demonstrating potential impacts on health that are reflective of microbiota changes. First, our study suggests that adipate is strongly associated with changes in the microbiome. Adipate, the salts and esters of adipic acid, are increased in the urine of breast-fed vs. formula-fed neonates (Dessì et al., 2016), a now well-known essential factor in the establishment of a diverse, healthy microbiome. We observe effect sizes that are indicative of an association between adipate and an increase in Firmicutes, Proteobacteria and Bacteriodes, the major components of a healthy adult intestinal flora. Therefore, clinical measurement of adipate may have potential as biomarkers for intestinal health.

Epicatechin is a polyphenolic flavonoid abundantly present in plants, cocoa and that offers protection against diabetes by its ability to mimic the effects of insulin and has strong antioxidant properties (Kanehisa and Goto, 2000; Samarghandian et al., 2017). The association we observed between epicatechin and proteobacteria abundance may prove useful in the prediction of pathologies associated with proteobacteria expansion, such as IBD or nonalcoholic fatty liver disease (Rizzatti et al., 2017). Of course, any potential mechanisms that underpin these associations will require further investigation but collectively this demonstrates the broad and significant potential for clinical application of our model. Several metabolites, including isoleucine, thymidine, ornithine and vanillate, were associated with an increase in cyanobacteria (Figure 4b). Cyanobacteria are unusual intestinal microflora (Ley et al., 2005), in that they perform oxygenetic photosynthesis and cyanobacteria-based supplements, such as Spirulina, regarded as functional foods, have been shown to support immunological function (Finamore et al., 2017). These associations may allow us to get a clearer picture of the functionality of the gut microbiome, by assessing the functions of these metabolites.

5 Conclusions

We have developed a Bayesian multi-task regression model for detecting global microbiome associations, and have shown that the model produces consistent estimates in a simulation study. We then applied the model to investigate microbiome-metabolome associations in a pilot study of inflammatory bowel disease patients. The model is superior to existing approaches in that it allows for integrating phylogenetic information, it estimates global as well as species-specific effects, and it does not rely on a specific choice of ecological distance measure.

There are some limitations to our study. The IBD dataset used in this paper only consists of a small number of samples, which reduced our ability to extend the inference to a larger number of species. For a larger sample size, the full log-ratio model could also be applied to taxonomic units at the 'order' or 'genus' level, as our simulation study showed that inference for up to 150 species is feasible without major modifications. However, we have seen that for large numbers of samples ($n \geq 500$), the conditional log-ratio model can be deployed as it showed comparable prediction accuracy to the full model at potentially greater computational efficiency.

An interesting avenue for future research is whether our model can scale to larger numbers of species S and sample sizes n. We note that time to convergence seems to increase polynomially with n; we would therefore not expect to be able to scale to tens of thousands of samples, and a different inference method would need to be employed. One option is variational inference, which has been successfully used as an alternative to full Bayesian inference in other settings (Hensman et al., 2013).

More crucially, we would like the method to scale to larger numbers of species. Currently, our results indicate that HMC inference in the full model converges faster as S increases. Our hypothesis is that for small S, the gradient estimate for a_j is unreliable as it depends on the estimates of $\beta_{s,j}$. For large S, we have more information to estimate the gradients, and thus experience fewer rejections during HMC sampling. However, estimation of the parameters will be poor unless n also increases. In future work, we will investigate regularization approaches to estimate sparse $\beta_{s,j}$ parameters in situations with small sample sizes.

Acknowledgements

We thank University Hospitals of Morecambe Bay Trust SIFT and the Academy of Medical Sciences Springboard scheme for funding and Dr Albert Davies

for collection of patient samples.

References

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J. J., Hayes, R. B., and Yang, L. (2013). Human gut microbiome and risk for colorectal cancer. *Journal of the National Cancer Institute*, 105(24):1907–1911.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Beamish, E. (2017). *Investigating Dysbiosis as a Cause and Predictor of Intestinal Pathology*. PhD thesis, Lancaster University.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113.

Chen, J. and Chen, M. J. (2018). Package 'gunifrac'.

Chen, J. and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1).

Dessì, A., Murgia, A., Agostino, R., Pattumelli, M. G., Schirru, A., Scano, P., Fanos, V., and Caboni, P. (2016). Exploring the role of different neonatal nutrition regimens during the first week of life by urinary GC-MS metabolomics. *Int. J. Mol. Sci.*, 17(2):265.

Finamore, A., Palmery, M., Bensehaila, S., and Peluso, I. (2017). Antioxidant, immunomodulating, and Microbial-Modulating activities of the sustainable and ecofriendly spirulina. *Oxid. Med. Cell. Longev.*, 2017:3247528.

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2(5):17004.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. *Uncertainty in Artificial Intelligence*, pages 282–290.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U. S. A.*, 102(31):11070–11075.

Lozupone, C. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–8235.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The CONCRETE distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Paradis, E. and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C., and Gasbarrini, A. (2017). Proteobacteria: A common factor in human diseases. *Biomed Res. Int.*, 2017:9351507.

Samarghandian, S., Azimi-Nezhad, M., and Farkhondeh, T. (2017). Catechin treatment ameliorates diabetes and its complications in Streptozotocin-Induced diabetic rats. *Dose Response*, 15(1):1559325817691158.

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, 6(2):e1000667.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1):4.