Koslovsky and Vannucci *BMC Bioinformatics* (2020) 21:301
https://doi.org/10.1186/s12859-020-03640-0

BMC Bioinformatics

SOFTWARE Open Access

MicroBVS: Dirichlet-tree multinomial regression models with Bayesian variable selection - an R package

Matthew D. Koslovsky* and Marina Vannucci

*Correspondence:
mkoslovsky12@gmail.com
Department of Statistics, Rice University, Houston, TX, USA

Abstract

Background: Understanding the relation between the human microbiome and modulating factors, such as diet, may help researchers design intervention strategies that promote and maintain healthy microbial communities. Numerous analytical tools are available to help identify these relations, oftentimes via automated variable selection methods. However, available tools frequently ignore evolutionary relations among microbial taxa, potential relations between modulating factors, as well as model selection uncertainty.

Results: We present MicroBVS, an R package for Dirichlet-tree multinomial models with Bayesian variable selection, for the identification of covariates associated with microbial taxa abundance data. The underlying Bayesian model accommodates phylogenetic structure in the abundance data and various parameterizations of covariates' prior probabilities of inclusion.

Conclusion: While developed to study the human microbiome, our software can be employed in various research applications, where the aim is to generate insights into the relations between a set of covariates and compositional data with or without a known tree-like structure.

Keywords: Bayesian analysis, Compositional data, Dirichlet-tree multinomial regression, Microbiome, Variable selection

Background

The human microbiome is a collection of prokaryotes, archaea, fungi, and viruses which may vary in composition depending on an individual's health, diet, and environment [1, 2]. High-throughput sequencing technologies enable researchers to characterize the composition of the microbiome by quantifying richness, diversity, and abundances (see [2] for a detailed review). Characterization of the microbiome is especially critical to the study of chronic diseases such as cancer and diabetes that may be associated with key changes in the microbiome [2].

Models developed to investigate microbial taxa abundance data collected on the human microbiome must be able to handle numerous analytical challenges observed in practice, including overdispersion, complex correlation structures, sparsity, high-dimensionality, and known biological information [2]. Recently, the Dirichlet-multinomial (DM) distribution has been used to model microbial count data, since it can accommodate overdispersion induced by sample heterogeneity and varying proportions among samples [3–6]. However, the DM model only assumes that counts are negatively correlated. Alternatively, the Dirichlet-tree multinomial model (DTM) inherits the DM's ability to handle overdispersed data and can model general correlation structures between counts as well as naturally incorporate structural information [7, 8]. Microbial abundance data, in particular, have been shown to depend on the evolutionary relations among taxa represented by a phylogenetic tree [9–11].

An important question in human microbiome research is to identify associations between microbial abundance data and clinical covariates, such as KEGG orthology pathways or dietary intake [5, 6, 9, 12–16]. For this, researchers often use penalized likelihood methods to simultaneously estimate regression coefficients and select covariates [6, 9]. These models are typically quite efficient and have shown good predictive accuracy [6, 9]. However, the ability of these models to incorporate information about known relations between covariates is limited due to the requirement of complex optimization routines [9]. Additionally, they do not accommodate model selection uncertainty while performing selection.

Alternatively, Bayesian variable selection methods are able to accommodate complex, high-dimensional data structures and fully account for model uncertainty over covariate selection [17, 18]. A common approach for Bayesian variable selection is to employ a spike-and-slab prior for regression coefficients that depends on a latent inclusion indicator for each covariate [18]. In this model formulation, unassociated covariates are pushed out of the model and associated covariates' regression coefficients are freely estimated. Recently, Wadsworth et al. [5] developed an approach for identifying KEGG orthology pathways that were associated with multivariate count data using a DM regression model with spike-and-slab priors. Through simulations, they demonstrate improved performance of their method on selecting covariates when compared to alternative methods, including the penalized likelihood approach of [6].

We present MicroBVS, an `R` package for Dirichlet-tree multinomial models with Bayesian variable selection, for the identification of covariates associated with microbial taxa abundance data. The underlying Bayesian model extends the work of Wadsworth et al. [5] by accommodating tree-like structure between the compositional data and also includes various parameterizations of covariates' prior probabilities of inclusion. While developed to study the human microbiome, our software can be employed in various research applications, where the aim is to generate insights into the relations between a set of covariates and compositional data with or without a known tree-like structure.

Implementation

Software implementation

Our contributed `R` package provides a general approach for identifying covariates associated with compositional data. At the core is a Markov chain Monte Carlo (MCMC) algorithm that generates posterior samples of model parameters for inference. The

MCMC algorithm is written in C++ to increase performance time and accessed through R wrapper functions using Rcpp and RcppArmadillo [19, 20]. The package extends the work of Wadsworth et al. [5] by accommodating tree-like structure between the compositional data via a DTM regression model. As a result, our approach incorporates the contributions of [5] as a special case and additionally is flexible to various prior probability of inclusion parameterizations. The package has built-in functionality to simulate data in user-specified research scenarios to assess selection performance and conduct sensitivity analyses. Additionally, various auxiliary R functions are incorporated to help researchers assess convergence, draw inference from the MCMC samples, and plot results. The package includes a vignette with worked examples using simulated data.

Data input and output

While designed to study microbial abundance data, our package can handle any research setting aimed at identifying factors associated with compositional data. Thus in microbiome analyses, our package is agnostic to the sequencing approach used to quantify microbial samples. In addition to compositional data, the method requires a set of covariates collected for each subject and a tree object that can be read by the R package ape [21]. Before analysis, we recommend standardizing continuous covariates and reparameterizing categorical covariates using indicator variables. Standard for any Bayesian approach, our algorithm requires the specification of various hyperparameters in the model. While we have set default values for each of the hyperparameters, the vignette contains details of their function in the algorithm as well as recommendations for their adjustment. Technical details of the model can be found in the Supplementary Material.

Once the algorithm has run, a list of MCMC samples for each of the parameters' posterior distributions is outputted. This list includes MCMC samples for intercept terms, covariates' respective regression coefficients, and latent inclusion indicators for covariates, which take on values of zero or one, corresponding to exclusion or inclusion in the model. Inclusion in the model is determined if the marginal posterior probability of inclusion (MPPI), calculated as the average of the MCMC inclusion indicator samples for each covariate-branch combination, is ≥ 0.50 [22]. An alternative inclusion threshold can be obtained using a Bayesian false discovery rate, which controls for multiplicity [23]. In addition to the functions provided in the package to draw posterior inference, the output can easily be transformed into a format that is readable by the coda package in R for further summaries, plotting, and diagnostics [24].

Application

To demonstrate the functionality of our software, we apply it to a benchmark data set collected to study the relation between dietary intake and the human gut microbiome [15]. Previously, Wang and Zhao [9] proposed a penalized DTM regression model to identify dietary intake covariates associated with genus-level operational taxonomic units (OTUs) on a subset of these data. For comparison, we apply our software to the same data. Briefly, the data used in this analysis consist of 28 genera-level OTU counts obtained from 16S rRNA sequencing and a corresponding set of 97 dietary intake covariates derived from diet information collected using a food frequency questionnaire on 98 subjects.

In this analysis, the model was run on these data using a DTM regression model. The phylogentic tree used in this analysis is presented in Fig. 1. We assumed a non-informative

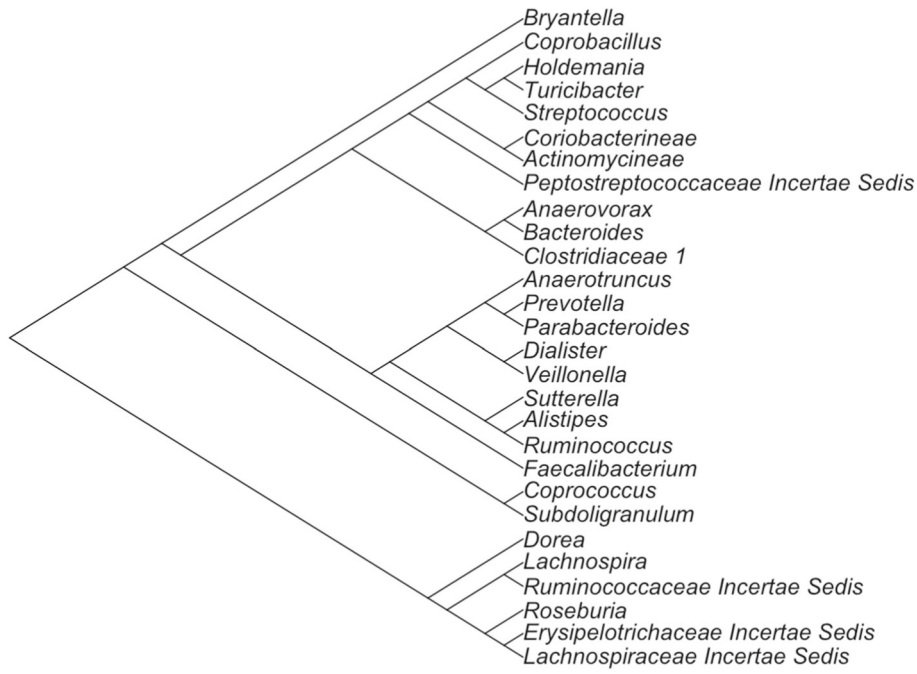

Fig. 1 Phyologentic tree for application data

Beta-Binomial prior for inclusion indicators ($a = b = 1$). The MCMC algorithm was run for 150,000 iterations. After a burn-in of 75,000 samples, inference was drawn from the remaining 75,000. Visual inspection of the trace plots for the number of active covariates in the model and the log posterior distribution indicated good convergence and mixing. A covariate's inclusion in the model was determined using a Bayesian false discovery rate of 0.01, corresponding to a MPPI $\geq$ 0.89. Additionally, we ran the method of [9] with penalty parameter $\gamma = 0.25$, corresponding to a sparse grouped lasso prior, over a grid of λ values, similar to their analysis. For the penalized approach, the best model was then chosen by minimizing the Akaike information criterion [25].

Results and discussion

We identified 232 dietary factor-branch associations with our Bayesian variable selection method for DTM regression models, whereas the penalized approach identified 271 associations overall. See Figs. 2 and 3 for a network representation of the associations identified by each model. Figure 4 captures the associations that our proposed method found that the penalized approach excluded. We observed that the penalized approach tended to identify similar dietary factors across taxa. These results may reflect the structure imposed by the sparse grouped lasso penalty used in the penalized approach. While the Beta-Binomial prior for inclusion indicators does not impose any structural relations between covariates, the MicroBVS package can be specified with graph-based inclusion priors, similar to [26, 27]. See the vignette for details regarding inclusion indicator prior specification.

Similar to our approach, Wang and Zhao's method identified factors associated with each branch of the phylogenetic tree. To summarize association results at the genus-level, they reported the most frequently selected dietary intake covariates along the paths from

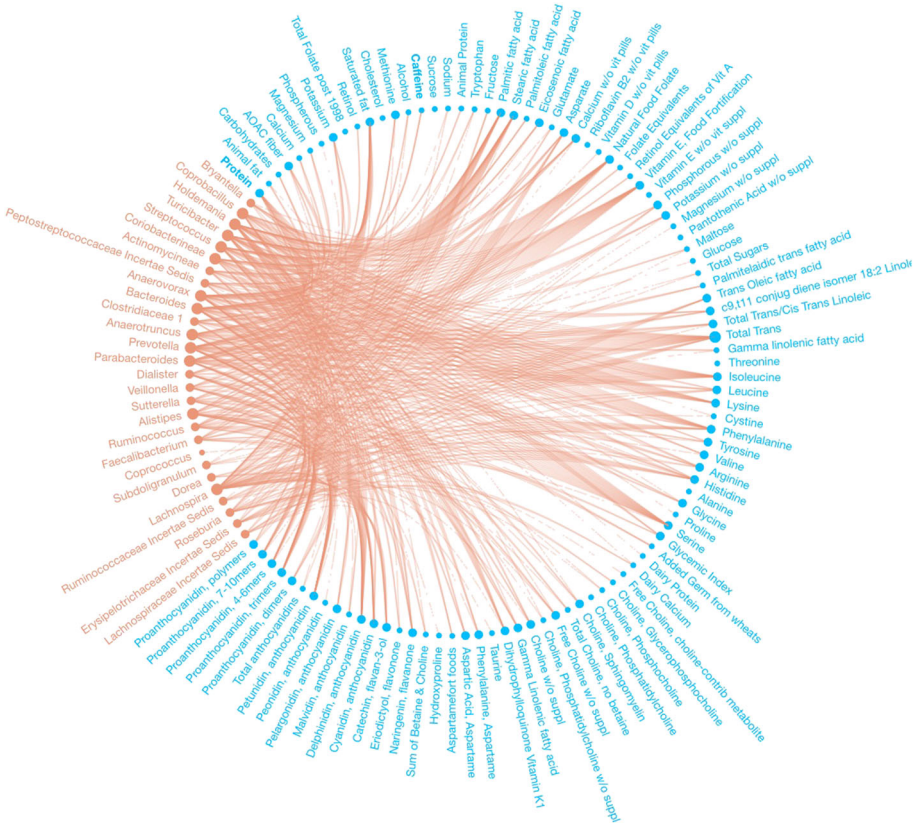

Fig. 2 Network of associations found using the proposed DTM MCMC algorithm. Identified associations are represented by edges between microbial taxa (red) and dietary factors (blue)

the root node of the phylogenetic tree to the leaf nodes representing two genera previously used to define enterotypes of the human microbiome [15, 28], Bacteroides and Prevotella, across 100 randomly split testing and training data sets. For comparison, we present a network graph of the dietary intake covariates identified by our model, but not the method of [9], along these same paths using the full data set (Fig. 5).

As in Wu et al. [15], we found associations between Bacteroides and various amino acids and fatty acids. Relations between amino acids and Bacteroides were also confirmed in [9]. Both [9] and [15] found Prevotella to be associated with a carbohydrate-based diet. Similar to [9], we identified *Naringenin, flavanone* and *Total Trans/Cis Trans Linoleic* as associated with Prevotella. Additionally, we identified relations between Prevotella and *Methionine*, *Phenylalanine*, *Total Choline, no betaine*, and *Sum of Betaine and Choline*, similar to [15]. Compared to [6], who proposed a penalized likelihood approach for a DM model, we also found relations between Bacteroides and *Animal fat, Eriodictyol, flavonone*, and *Maltose* as well as between Prevotella and *Choline, Phosphatidylcholine*.

Bayesian variable selection methods for regression models have shown better selection performance than penalized approaches [5, 29, 30]. However, these approaches are typically computationally less efficient. For the DTM regression models of this paper, the dimension of the model space grows dramatically as a function of the number of covariates, number of leaf (or root) nodes, and complexity of the phylogenetic tree. Specifically

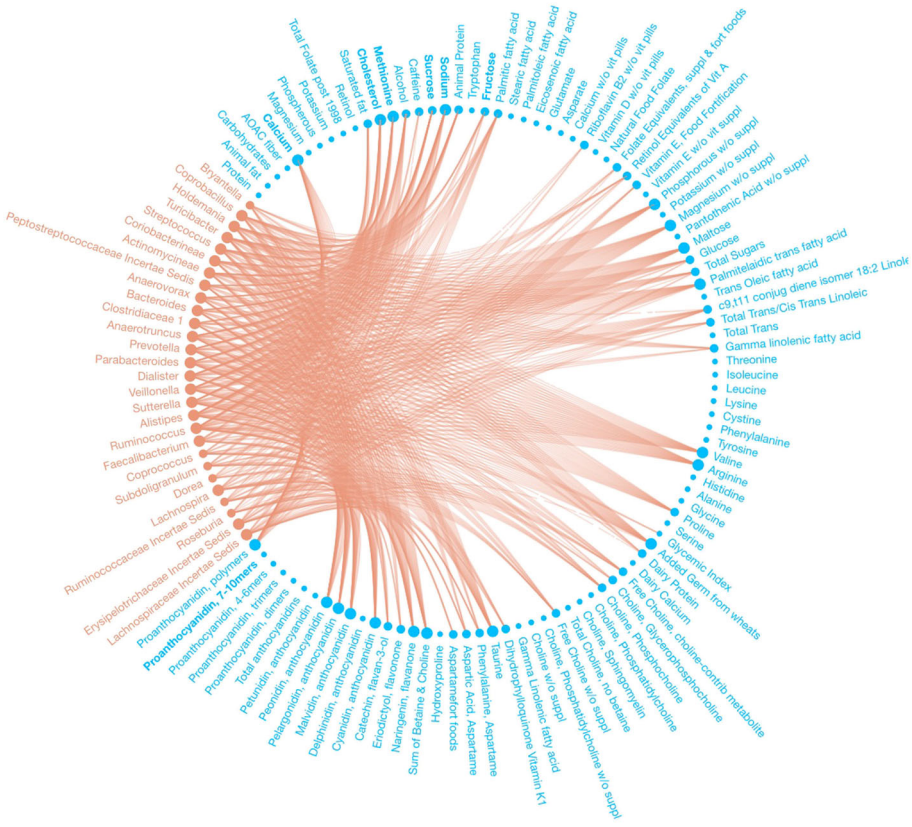

Fig. 3 Network of associations found using the method of [9]. Identified associations are represented by edges between microbial taxa (red) and dietary factors (blue)

for B branches and P covariates, there are $2^{B \times P}$ potential models to choose from. In addition to large parameter spaces, convergence of the model is highly dependent on the correlation structure between covariates and count data, as well as the sparsity level of the model. For the analysis of this paper, the DTM model took around 9 hours to run 150,000 iterations on a 2.5 GHz dual-core Intel Core i5 processor with 8 GB RAM. To maintain reasonable computation times and selection performance, we recommend applying the Bayesian DTM model to small-to-medium sized microbiome data sets, that is, with less than 100 compositional components and moderate-to-large tree-structures when $B \times P >> n$. Larger data sets might be analyzed by employing the Dirichlet-multinomial regression model of Wadsworth et al. [5], which does not incorporate the phylogenetic tree. This option is available within the MicroBVS software.

Our software implementation includes some of the most commonly used inclusion indicator priors. In practice, researchers are often interested in identifying higher-order terms, such as interactions, or grouped covariates. Future developments of the software may include functionality to handle these type of settings following [31]. Additionally, we assume that all of the covariate relations in the model are linear, which may not be realistic. Alternative priors for regression coefficients are available that can handle non-parametric relations (e.g., Dirichlet process priors). As the dimension of the model grows, inference becomes challenging. In addition to the posterior inference tools we provide in this version of the R package, more advanced visualization tools may permit a deeper

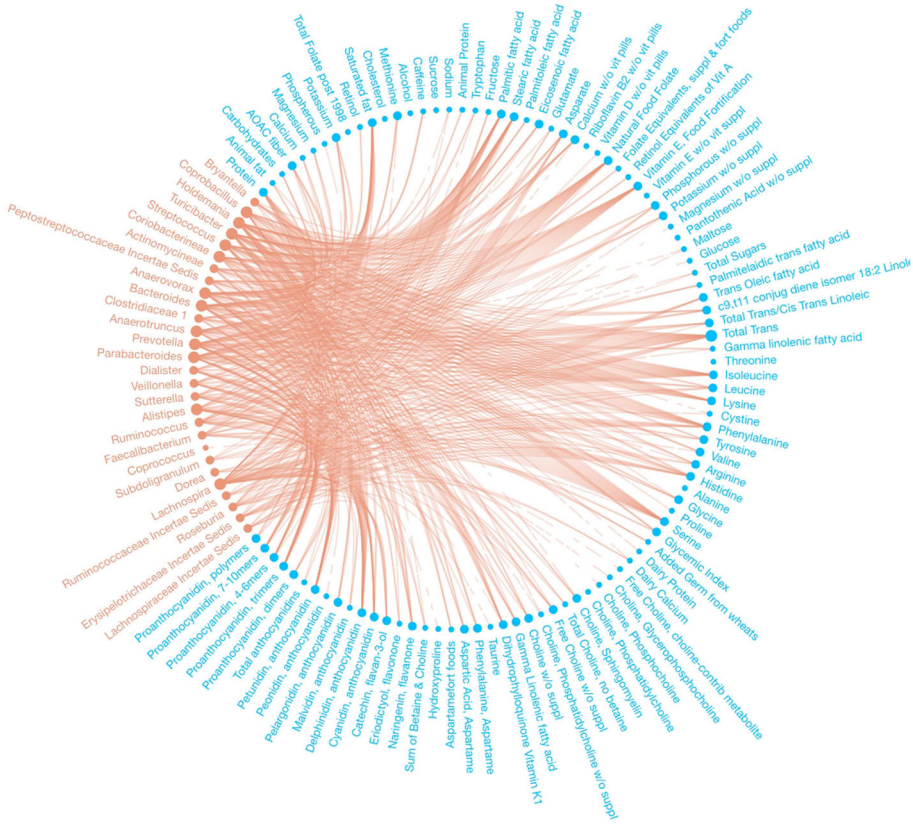

Fig. 4 Network of associations found using the proposed DTM MCMC algorithm and not the method of [9]. Identified associations are represented by edges between microbial taxa (red) and dietary factors (blue)

understanding of the model's results in applications. While using a fully Bayesian MCMC algorithm for posterior inference accommodates both parameter estimation and model selection uncertainty, our approach may not scale as well as approximate Bayesian methods, which may underestimate model uncertainty, to extremely large data sets. For DM and negative binomial regression models, [32] devised an efficient, variational Bayes variable selection approach via spike-and-slab priors. In future work, we aim to incorporate a variational alternative for DTM regression models, as well as extend our package to handle other data structures commonly found in microbiome research (e.g., zero-inflated counts, negative binomial distributions).

Conclusions

This software package provides a general Bayesian approach for identifying factors associated with compositional data that may have known tree-like structure. Additionally, the package is accompanied by a detailed vignette that contains a step-by-step tutorial demonstrating how to use the package in practice. Together, our user-friendly package enables researchers to investigate heterogeneity in compositional data potentially explained by a set of covariates. While we demonstrate our package in the context of human microbiome data, it can be applied to various research settings.

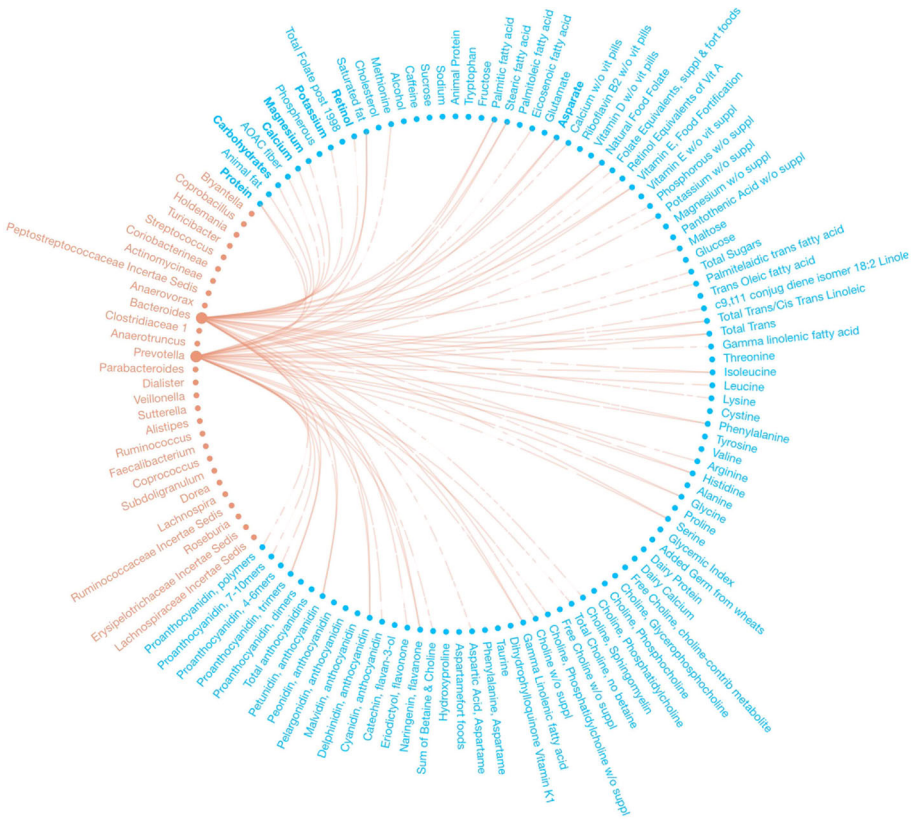

Fig. 5 Network of associations found for Bacteroides and Prevotella using the proposed DTM MCMC algorithm and not the method of [9]. Identified associations are represented by edges between microbial taxa (red) and dietary factors (blue)

Availability of data and requirements

Project name: MicroBVS

Project home page: https://github.com/mkoslovsky/MicroBVS

Operating system(s): Linux, Mac OS, Windows

Programming language: R and C++ Other requirements: R Rcpp RcppArmadillo ape MCMCpack mvtnorm ggplot2 GGMselect devtools ape igraph

License: MIT

Any restrictions to use by non-academics: None.

Data Availability: All simulated data can be generated using the R package. Data analyzed in the Case Study are available in the R package [15].

Abbreviations

DM: Dirichlet-multinomial; DTM: Dirichlet-tree multinomial; GB: Gigabyte; GHz: Gigahertz; KEGG: Kyoto encyclopedia of genes and genomes; MCMC: Markov chain Monte Carlo; MPPI: Marginal posterior probability of inclusion; OTU: Operational taxonomic unit; RAM: Random access memory

Acknowledgements

We thank Tao Wang, Hongyu Zhao, and Hongzhe Li for providing access to the case study data found in [15].

Authors' contributions

MV and MK conceived the method. MK developed the R package and drafted the manuscript. All authors read and approved the final version of the manuscript.

Funding
Matthew Koslovsky is supported by NSF via the Research Training Group award DMS-1547433. The funders had no role in the design of the study, collection, analysis and interpretation of data and in the preparation of the manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 2 January 2019 Accepted: 2 July 2020
Published online: 22 July 2020

References

1. Sanz Y, Olivares M, Moya-Pérez Á, Agostoni C. Understanding the role of gut microbiome in metabolic disease risk. Pediatr Res. 2014;77(1-2):236–44.
2. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu Rev Stat Appl. 2015;2: 73–94.
3. Zhang Y, Zhou H, Zhou J, Sun W. Regression models for multivariate count data. J Comput Graph Stat. 2017;26(1): 1–13.
4. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, Sodergren E, Weinstock G, Shannon WD. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PloS ONE. 2012;7(12): 52078.
5. Wadsworth WD, Argiento R, Guindani M, Galloway-Pena J, Shelburne SA, Vannucci M. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. BMC Bioinformatics. 2017;18(1):1–12.
6. Chen J, Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. Ann Appl Stat. 2013;7(1):418–42.
7. Dennis III SY. On the hyper-Dirichlet type 1 and hyper-Liouville distributions. Commun Stat-Theory Methods. 1991;20(12):4069–81.
8. Minka T. The Dirichlet-tree distribution. 1999. https://www.microsoft.com/en-us/research/publication/dirichlet-tree-distribution/.
9. Wang T, Zhao H. A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. Biometrics. 2017;73(3):792–801.
10. Tang Y, Ma L, Nicolae DL, et al. A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. Ann Appl Stat. 2018;12(1):1–26.
11. Tang Z-Z, Chen G, Alekseyenko AV, Li H. A general framework for association analysis of microbial communities on a taxonomic tree. Bioinformatics. 2017;33(9):1278–85.
12. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016;32(17):2611–7.
13. McMurdie PJ, Holmes S. Waste not, want not: Why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10(4):1003531.
14. Garcia TP, Müller S, Carroll RJ, Walzem RL. Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. Bioinformatics. 2013;30(6):831–7.
15. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334(6052):105–8.
16. Koslovsky MD, Hoffman KL, Daniel CR, Vannucci M. A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. Ann Appl Stat. 2020. In press.
17. Brown PJ, Vannucci M, Fearn T. J R Stat Soc Ser B Stat Methodol. 1998;60(3):627–41.
18. George EI, McCulloch RE. Approaches for Bayesian variable selection. Stat Sin. 1997;7(2):339–73. JSTOR.
19. Eddelbuettel D, Sanderson C. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. Comput Stat Data Anal. 2014;71:1054–63.
20. Eddelbuettel D, François R, Allaire J, Ushey K, Kou Q, Russel N, Chambers J, Bates D. Rcpp: Seamless R and C++ integration. J Stat Softw. 2011;40(8):1–18.
21. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–90.
22. Barbieri MM, Berger JO, et al. Optimal predictive model selection. Ann Stat. 2004;32(3):870–97.
23. Noecker C, Eng A, Srinivasan S, Theriot CM, Young VB, Jansson JK, Fredricks DN, Borenstein E. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. MSystems. 2016;1(1):13–5.
24. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R News. 2006;6(1):7–11.
25. Akaikei H. Information theory and an extension of maximum likelihood principle. In: Proc 2nd Int Symp on Information Theory; 1973. p. 267–81.
26. Peterson CB, Stingo FC, Vannucci M. Joint Bayesian variable and graph selection for regression models with network-structured predictors. Stat Med. 2016;35(7):1017–31.

27. Stingo FC, Chen YA, Tadesse MG, Vannucci M. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. Ann Appl Stat. 2011;5(3):1978–2002.
28. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, et al. Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174–80.
29. Kyung M, Gill J, Ghosh M, Casella G, et al. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal. 2010;5(2):369–411.
30. Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann Appl Stat. 2011;5(3):1780–815.
31. Chipman H. Bayesian variable selection with related predictors. Can J Stat. 1996;24(1):17–36.
32. Miao Y, Kook J, Lu Y, Guindani M, Vannucci M. Scalable Bayesian variable selection regression models for count data. In: Yanan F, Smith M, Nott D, Dortet-Bernadet J, editors. Flexible Bayesian Regression Modelling. Elsevier; 2020. p. 187–219.

Publisher's Note