3

High-dimensional statistics with QIIME2

3.1. Software overview

We introduce software for solving high-dimensional statistics problems with an application to microbiome data. The initial component of our software, denoted as q2-gglasso, facilitates microbial network inference by solving the General Graphical Lasso problems. The subsequent plugin, named q2-classo, employs log-contrast models for solving regression and classification tasks. Both of these plugins have been integrated within the QIIME2 framework - an open-source bioinformatic platform which offers diverse user interfaces, tracks data provenance, and has a wide range of other plugins enabling comprehensive downstream analysis of microbiome data. Figure 3.1 illustrates the plugins functionality, including sample covariance estimation, inverse covariance (precision) estimation, principal component analysis (PCA), various regression and classification tasks, as well as summary statistics and interactive visualizations depicting the results. The plugins can be installed from QIIME2 or Docker Hub.

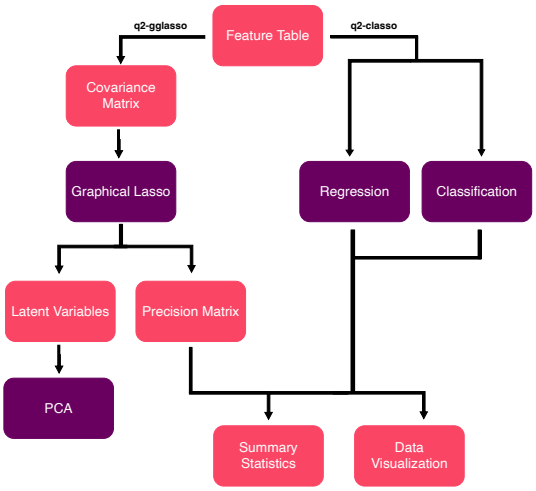

Figure 3.1: High-dimensional statistics with QIIME2.

Both plugins have been previously implemented in Python. For a better understanding of the specific problem formulation and more examples, we encourage you to visit gglasso and classo documentation.

Key features: graphical lasso; model selection; sparse log-contrast regression and classification.

3.2. Atacama soil microbiome data

The Atacama Desert, known for its exceptional aridity, experiences minimal precipitation, with certain regions receiving less than a millimeter of rainfall per decade. However, despite these extremely dry conditions, microbial organisms thrive within the soil. In our example, we showcase the application of our high-dimensional statistics QIIME2 plugins using the Atacama soil microbiome dataset described by Neilson et al. [21]. With q2-gglasso we solve various graphical lasso problems [25] to identify microbial associations which we later assess by fitting sparse log-contrast models [26] implemented in q2-classo.

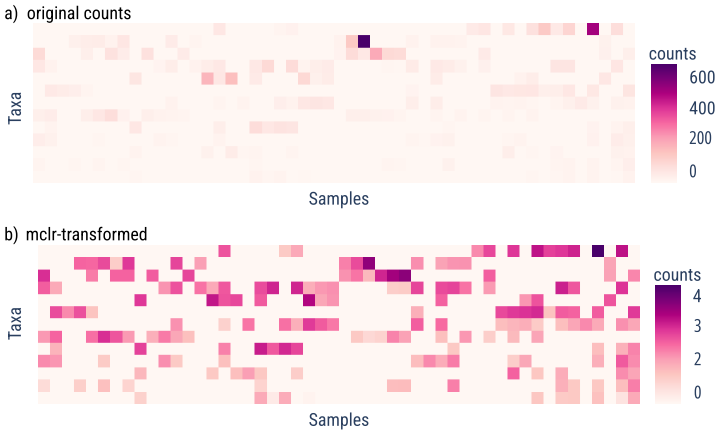

Figure 3.2: Atacama soil microbiome a) before and b) after mclr-transformation.

Microbiome data typically represents the relative abundances of different microbial taxa within a sample. However, these relative abundances are constrained because the total sum of relative abundances adds up to a constant [12]. This compositional constraint can cause statistical issues and distort the interpretation of the data. To address these limitations, we apply the mclr-transformation (Figure 3.2) which converts the compositional data into an unconstrained data space [31]. After the initial pre-processing step, our data consists of 13 distinct microbial taxa and 50 individual samples. In addition, we select covariates related to the soil microbiome data, including pH, elevation (m), average soil temperature (t°), and humidity (%). The covariates live on a different scale, so we perform a standard scaling procedure from scikit-learn [22], where the data is also centered before scaling to unit standard deviation (Figure 3.3). Thus, we ensure that all the covariates are transformed to a comparable scale.

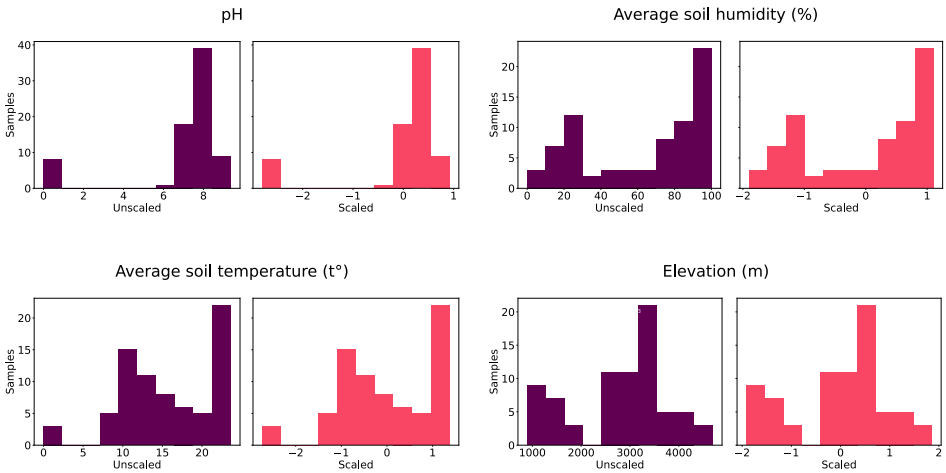

Figure 3.3: Scaled covariates associated with Atacama soil microbiome.

3.3. Graphical Lasso models

Figure 3.4a depicts the empirical correlation matrix between taxa which are clustered with the average linkage method [20]. The matrix reveals distinct patterns, where certain taxa exhibit closer relationships forming two cohesive blocks and we would like to investigate those relationships further. By assuming the underlying inverse covariance matrix is sparse, we solve graphical lasso problem [11] to estimate the inverse correlation matrix which represents a conditional independent relationship between taxa.

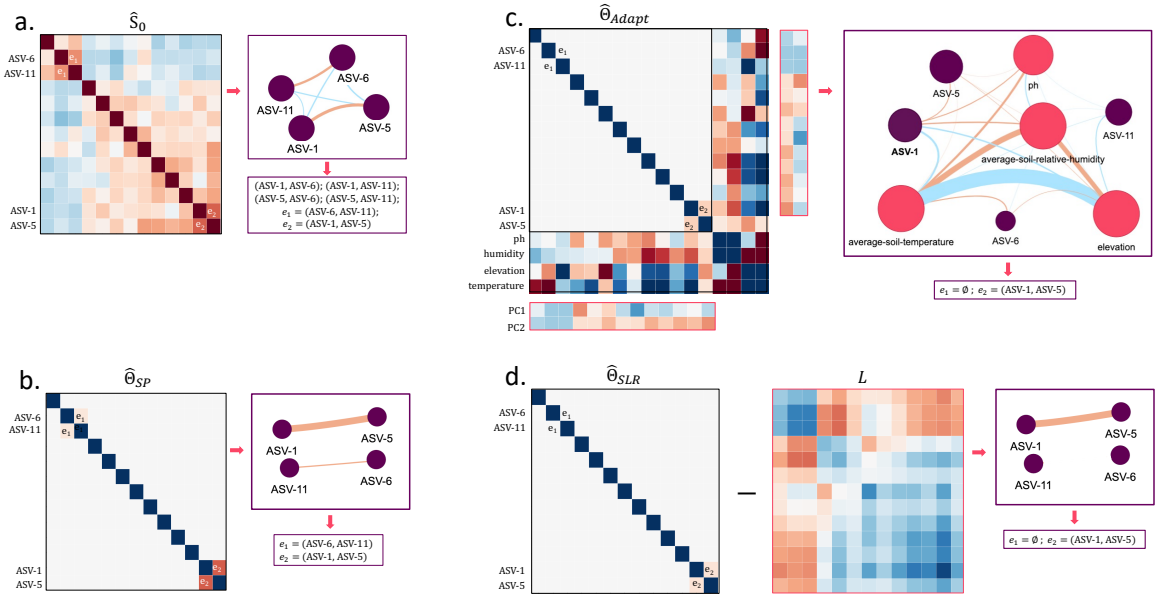

Figure 3.4: General Graphical Lasso: a) empirical correlation; b) sparse; c) adaptive d) sparse + low-rank models.

We start our analysis by looking at the empirical correlation matrix $\hat{S}_0$ and corresponding network. In Figure 3.4a, we observe two edges $e_1 = (ASV_6, ASV_{11})$ and $e_2 = (ASV_1, ASV_5)$. Also, we see associations between ASV_1, ASV_5, ASV_6 and ASV_{11} making this sub-graph fully connected. In a larger graph, capturing these individual associations would be challenging. So, under the sparsity assumption, we proceed to solve a single graphical lasso problem [11]. In Figure 3.4b, we observe the persistence of the positive edges e_1 and e_2, and the sparse model has eliminated the rest of the edges.

It is well-known that taxa correlate with specific environmental factors, such as pH, soil temperature, humidity, and elevation. We estimate the partial correlation between taxa and these covariates jointly by applying penalization procedure element-wise which is referred to as an adaptive graphical lasso solution $\hat{\Theta}_{Adapt}$. In Figure 3.4c, we see the remaining positive edge e_2, and the previously identified edge e_1 can be attributed to the temperature since ASV_6 and ASV_{11} no longer correlate with each other.

In scenarios where covariates are missing or corrupted, it becomes challenging to capture the effects of hidden confounders. To address this issue, we solve a single graphical lasso with latent variables also known as the sparse + low-rank model [8]. Figure 3.4d shows solution $\hat{\Theta}_{SLR}$ where e_2 remains, but e_1 disappears due to the low-rank component L. The network of $\hat{\Theta}_{SLR}$ is more similar to $\hat{\Theta}_{Adapt}$, which incorporates the available covariate information. We compare the first two principal components (PCs) of L and compare them with the estimated covariates from the adaptive model. Notably, we observe that PC_1 likely captures the effect of pH, while PC_2 captures the effect of the average soil temperature.

In a related study [16], an alternative approach utilizing the sparse + low-rank model was explored. The authors used the low-rank part of the model to perform a robust principal component analysis (PCA) [7] of the microbial counts. In their analysis, they discovered that some correlations could be attributed to latent factors such as zero counts, compositional artifacts, and batch effects. In Figure 3.5, we replicate their example using the Atacama soil microbiome dataset. We assume that the top-left block of the empirical correlation matrix $\hat{S}_0$ is driven by a latent factor which is not taken into account. In the sparse solution $\hat{\Theta}_{SP}$, most of those spurious correlations are eliminated, but edge e_1 is still present and claims the direct association between two taxa. In contrast, the top-left block of $\hat{\Theta}_{SLR}$ does not exhibit any direct associations, as the low-rank component L effectively captures the influence of latent

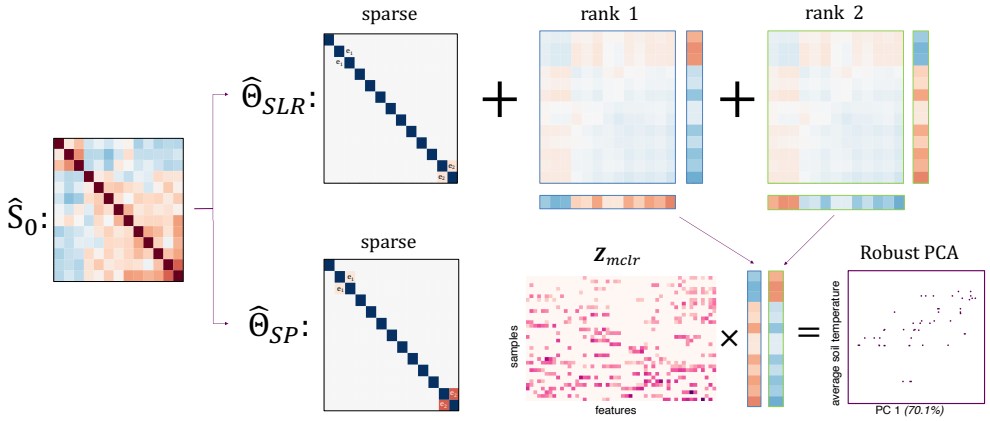

Figure 3.5: Example of using sparse + low-rank model for robust PCA.

confounders. This can be verified by projecting each sample from Z_{mclr} onto the principal components of L and examining the correlation between these PCs and the covariates. Figure 3.6 demonstrates the correlation between PC_1 and the average soil temperature and elevation. We observe that microbial compositions increase as the temperature rises, while they decline with an increase in elevation.

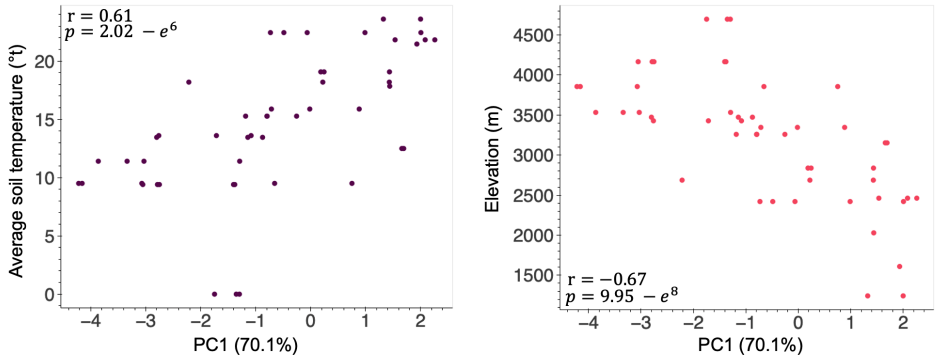

Figure 3.6: Relationship between PC1 and the average soil temperature (left), elevation (right).

To sum up, we present three distinct scenarios for using q2-gglasso, namely sparse, adaptive, and sparse + low-rank models. The results showcased that the sparse + low-rank model was able to effectively replicate a widely acknowledged biological phenomenon: the dependence of microbial compositions on environmental factors such as temperature and elevation. In line with the assumptions outlined in [8], we propose a framework which captures the underlying structure and interactions within the microbial community while promoting sparsity in the estimated network.

3.4. Sparse log-contrast models

Classification and regression are classical machine learning tasks, with numerous frameworks designed for their solution. However, microbiome data is compositional [12], and employing standard regression or classification models may yield inaccurate outcomes. Therefore, we recommend the adoption of sparse log-contrast models implemented in q2-classo and taking into account the data compositionality.

$$\min_{\beta \in \mathbb{R}^d} \|X\beta - y\|^2 + \lambda \|\beta\|_1 \qquad \text{subject to:} \quad C\beta = 0 \tag{3.1}$$

Extending our previous analysis of the Atacama soil microbiome, now our objective is to predict the average soil temperature based on microbiome compositions and additional covariates. We address the problem outlined in Equation 3.1, which entails a linear model. Notably, this model incorporates an extra constraint on coefficients (β) to adjust for the compositionality. Figure 3.7 shows the solution ($R^2 = 0.707$) at the chosen parameter $\lambda = 0.035$. The framework facilitates a stability selection procedure

with a threshold probability ($P = 0.7$) corresponding to the likelihood that ASV_4, ASV_5, ASV_6, ASV_7, ASV_8, ASV_9, ASV_{11} and ASV_{12} have been consistently chosen throughout the model selection process.

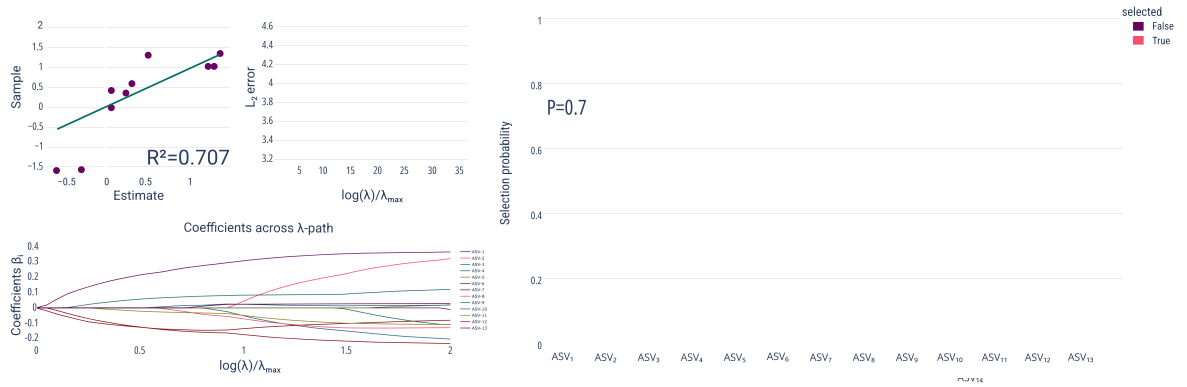

Figure 3.7: Standard constrained Lasso regression with stability selection.

Similarly, we can solve a classification task to predict a binary vegetation variable. We use the same set of covariates and microbial compositions, but the problem is now defined by Equation 3.2. In this example, we would like to use cross-validation as the model selection procedure, an option provided by q2-classo which complements the previously described stability selection approach.

$$\min_{\beta \in \mathbb{R}^d} \max\left(0, 1 - y^t X\beta\right)^2 + \lambda\|\beta\|_1 \qquad \text{subject to:} \quad C\beta = 0 \tag{3.2}$$

Figure 3.8 shows the misclassification rate of the model as well as the refitted coefficients of β after cross-validation with $\lambda = 0.376$. From the summary statistics of the solution, we can compute standard metrics to assess and evaluate the performance of the model. The observed accuracy (0.9) signifies that 90% of the model's predictions are accurate. Precision (1) implies an absence of false positives and recall (0.8) suggests a comparatively low incidence of false negatives. F1-score (0.889) is the harmonic mean of precision and is particularly useful in our case with uneven class distribution. Overall, these metrics suggest a reasonably good model. The most substantial contributions to vegetation prediction come from ASV_6 and ASV_9. Notably, ASV_6 is a taxon previously linked to average soil temperature.

Figure 3.8: Standard constrained Lasso classification with cross-validation.

q2-classo also supports a combination of log-contrast regression and tree-aggregated regularization models [3]. The goal of these models is to utilize the taxonomic information from microbiome data (Figure 3.9). Our analysis, when employing trac-models, revealed marginally superior results in a regression context ($R^2 = 0.8$) but did not demonstrate any enhancement in the classification setting. Nevertheless, we recommend exploring trac-models, as it is common for the inclusion of taxonomic information to enhance model performance and interpretability.

Figure 3.9: Taxonomic tree.

3.5. Results

Microbiome data represents the relative abundances of microbial taxa, encounters compositional constraints [12], leading to statistical challenges and potential misinterpretation. In our analysis of the Atacama soil microbiome data [21], we applied two new QIIME2 plugins for high-dimensional statistics. We employed q2-gglasso to solve graphical lasso problems [25] and subsequently assessed microbial associations through sparse log-contrast models [26] implemented in q2-classo. While we provide only a handful of illustrations demonstrating the application of our framework, we encourage users to search the documentation for more in-depth insights into these two plugins. In summary, our analysis employed statistical approaches, unveiling microbial associations, highlighting the influence of environmental factors, and demonstrating the versatility of regression and classification models for microbiome data.

References

[1] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[2] Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. “Adaptive linear step-up procedures that control the false discovery rate”. In: *Biometrika* 93.3 (2006), pp. 491–507.

[3] Jacob Bien et al. “Tree-aggregated predictive modeling of microbiome data”. In: *Scientific Reports* 11.1 (2021), p. 14505.

[4] Supinda Bunyavanich et al. “Early-life gut microbiome composition and milk allergy resolution”. In: *Journal of Allergy and Clinical Immunology* 138.4 (2016), pp. 1122–1130.

[5] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis”. In: *The ISME journal* 11.12 (2017), pp. 2639–2643.

[6] Benjamin J Callahan et al. “DADA2: High-resolution sample inference from Illumina amplicon data”. In: *Nature methods* 13.7 (2016), pp. 581–583.

[7] Emmanuel J Candès et al. “Robust principal component analysis?” In: *Journal of the ACM (JACM)* 58.3 (2011), pp. 1–37.

[8] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. “Latent variable graphical model selection via convex optimization”. In: *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2010, pp. 1610–1613.

[9] Robert C Edgar. “Updating the 97% identity threshold for 16S ribosomal RNA OTUs”. In: *Bioinformatics* 34.14 (2018), pp. 2371–2375.

[10] Bradley Efron. *Local false discovery rates*. 2005.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Sparse inverse covariance estimation with the graphical lasso”. In: *Biostatistics* 9.3 (2008), pp. 432–441.

[12] Gregory B Gloor et al. “Microbiome datasets are compositional: and this is not optional”. In: *Frontiers in microbiology* 8 (2017), p. 2224.

[13] Jean-Jacques Godon et al. “Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis”. In: *Applied and environmental microbiology* 63.7 (1997), pp. 2802–2813.

[14] Rolf Holle et al. “KORA-a research platform for population based health research”. In: *Das Gesundheitswesen* 67.S 01 (2005), pp. 19–25.

[15] Anna Klindworth et al. “Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies”. In: *Nucleic acids research* 41.1 (2013), e1–e1.

[16] Zachary D Kurtz, Richard Bonneau, and Christian L Müller. “Disentangling microbial associations from hidden environmental and technical factors via latent graphical models”. In: *bioRxiv* (2019), pp. 2019–12.

[17] Joseph J Lee et al. “More powerful multiple testing in randomized experiments with non-compliance”. In: *Statistica Sinica* (2017), pp. 1319–1345.

[18] Huang Lin and Shyamal Das Peddada. “Analysis of compositions of microbiomes with bias correction”. In: *Nature communications* 11.1 (2020), p. 3514.

[19] Henry B Mann and Donald R Whitney. “On a test of whether one of two random variables is stochastically larger than the other”. In: *The annals of mathematical statistics* (1947), pp. 50–60.

[20] Daniel Müllner. “Modern hierarchical, agglomerative clustering algorithms”. In: *arXiv preprint arXiv:1109.2378* (2011).

[21] Julia W Neilson et al. “Significant impacts of increasing aridity on the arid soil microbiome”. In: *MSystems* 2.3 (2017), e00195–16.

[22] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[23] Christian Quast et al. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools”. In: *Nucleic acids research* 41.D1 (2012), pp. D590–D596.

[24] Jessica H Savage et al. “A prospective microbiome-wide association study of food sensitization and food allergy in early childhood”. In: *Allergy* 73.1 (2018), pp. 145–152.

[25] Fabian Schaipp, Oleg Vlasovets, and Christian L. Müller. “GGLasso - a Python package for General Graphical Lasso computation”. In: *Journal of Open Source Software* 6.68 (2021), p. 3865. DOI: 10.21105/joss.03865. URL: https://doi.org/10.21105/joss.03865.

[26] Léo Simpson, Patrick L Combettes, and Christian L Müller. “c-lasso–a Python package for constrained sparse and robust regression and classification”. In: *arXiv preprint arXiv:2011.00898* (2020).

[27] Emmanuel Stephen-Victor and Talal A Chatila. “Regulation of oral immune tolerance by the microbiome in food allergy”. In: *Current opinion in immunology* 60 (2019), pp. 141–147.

[28] Qiong Wang et al. “Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy”. In: *Applied and environmental microbiology* 73.16 (2007), pp. 5261–5267.

[29] Amy Willis and John Bunge. “Estimating diversity via frequency ratios”. In: *Biometrics* 71.4 (2015), pp. 1042–1049.

[30] Amy D Willis and Bryan D Martin. “Estimating diversity in networked ecological communities”. In: *Biostatistics* 23.1 (2022), pp. 207–222.

[31] Grace Yoon, Irina Gaynanova, and Christian L Müller. “Microbial networks in SPRING-Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data”. In: *Frontiers in genetics* 10 (2019), p. 516.

[32] Grace Yoon, Christian L Müller, and Irina Gaynanova. “Fast computation of latent correlations”. In: *Journal of Computational and Graphical Statistics* 30.4 (2021), pp. 1249–1256.