Fast log-ratio selection for high-dimensional compositional data

E. Gordon-Rodriguez[1,*], T. P. Quinn[2], and J. P. Cunningham[1]
[1]Columbia University, New York, USA; **eg2912@columbia.edu*
[2]Deakin University, Geelong, Australia

Abstract

Log-ratio selection tools such as *selbal* (Rivera-Pinto et al., 2018) are becoming increasingly popular in the context of high-dimensional compositional data, such as those obtained from high-throughput sequencing technologies. These tools allow the user to automatically identify an (approximately) maximally predictive log-ratio, given a set of compositional input features. In turn, downstream models based on these log-ratios inherit attractive theoretical and empirical properties, including scale-invariance, subcompositional coherence, and subcompositional dominance. However, the space of possible log-ratios grows combinatorially in the dimension of the data, and as a result, existing selection algorithms are slow to run on high-dimensional datasets that are becoming commonplace in bioinformatics.

The *CoDaCoRe* algorithm and software package solves this computational bottleneck by deploying gradient-based optimization to the log-ratio selection problem (Gordon-Rodriguez et al., 2022; Quinn et al., 2021). At the heart of the algorithm is the use of ***Co****ntinuous* ***Re****laxations*, inspired by deep learning. The computational cost of this optimization technique scales linearly in the dimension of the data. As a result, CoDaCoRe is able to perform balance selection in a matter of seconds on very high-dimensional datasets (with tens of thousands of input features). Notwithstanding the reduced computational cost, the log-ratios obtained by CoDaCoRe are on par with *selbal* and other competitors in terms of predictive accuracy and sparsity. Here we review the algorithm and showcase its usage through domain applications to microbiome data. In addition, we describe some extensions to the framework, including unsupervised log-ratio selection and multi-omic analyses. To the best of our knowledge, CoDaCoRe is the first balance selection tool that scales to truly high-dimensional compositional data.

Key words: Log-ratio selection, balances, high-throughput sequencing.

References

Gordon-Rodriguez, E., T. P. Quinn, and J. P. Cunningham (2022). Learning sparse log-ratios for high-throughput sequencing data. *Bioinformatics 38*(1), 157–163.

Quinn, T. P., E. Gordon-Rodriguez, and I. Erb (2021). A critique of differential abundance analysis, and advocacy for an alternative. *arXiv preprint arXiv:2104.07266*.

Rivera-Pinto, J., J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle (2018). Balances: a new perspective for microbiome analysis. *MSystems 3*(4), e00053–18.