Sparse Canonical Correlation Analyses of Multimodal Omics Data*

Kejun He[1], Xiaoning Qian [5], Jianhua Huang [1], Sharon M. Donovan[4], Robert S. Chapkin[3], Ivan Ivanov[2]

[1] Statistics, TAMU, kejun@stat.tamu.edu, jianhua@stat.tamu.edu
[2] Veterinary Physiology & Pharmacology, TAMU, IIvanov@cvm.tamu.edu
[3] Integrative Nutrition & Complex Diseases, TAMU, r-chapkin@tamu.edu
[4] Food Science & Human Nutrition, U of Illinois, sdonovan@illinois.edu
[5] Electrical & Computer Engineering, TAMU, xqian@ece.tamu.edu

Keywords: omics data integration, canonical correlation analysis

There have been an increasing number of applications of sparse Canonical Correlation Analysis (sCCA) to genomic data during the past several years. Most of the research in this area has focused on the relationships between gene expression levels and phenotype variations. However, as multimodal omics data becomes available there is a need to integrate these data modalities into a framework that allows for simultaneous data analyses, thereby providing novel insights for various fields in the life sciences. The pioneering work of [1] used the classical Canonical Correlation Analysis (CCA) to provide an integrative approach to the analysis of host gene expression and microbiota composition data from neonates with different feeding types. Although promising, the proposed approach has serious deficiencies. First, the statistical interpretation is problematic because the involved two-stage analysis makes the results sensitive to the variations of data and the original interpretation of CCA is lost. Second, the associated computational cost is tremendous, $O(n^3)$ where n is the numer of variables involved in the analysis. Thus, we developed a methodology based on the sCCA to overcome these problems. The performance of our approach is compared to that of [1] and to the sparse Principal Component Analysis (sPCA) on a large synthetic data set with the subsequent application to a multimodal omics data (gene expression, microbiota composition, and metabolites) from neonates with two different feeding types.

[1] Schwartz, Scott et al., *A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response*, Genome Biol., 13,(4):r32, 2012.

*Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under grant number P30ES023512. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health