Research Statement

of Yue Wang, Ph.D.

Recent development in high-throughput technologies provides researchers with access to massive biological data (e.g., neuroimaging and multi-omics data). Such biological data are *high-dimensional* (i.e., the sample size is less than the number of variables) and *structured* (e.g., smoothness, sparsity, and network structures). My research is mainly focused on developing *high-dimensional* statistical tools (e.g., graphical visualization tools, prediction models, and inferential tools) for analyzing such massive biological data while efficiently accounting for potential *structures*. In the following, I first briefly introduce a few past projects that I have worked on. Then, based on these past projects, I will discuss my immediate work and my long-term future research plan.

Partial Least Squares (PLS) for functional regression models: Clinical studies of neurodegenerative diseases, such as Alzheimer's disease, often involve using medical images (fMRI, PET) at an earlier time point to predict important disease-related outcomes at a later time point. These medical images are high-dimensional (millions of voxels) and have complex structures (voxels in nearby brain regions may be highly correlated). In joint work with my Ph.D. advisors, I have developed a residual-based alternative partial least squares (RAPLS) estimation framework [1, 2] for a broad class of functional regression models to predict continuous, binary, count, longitudinal, or survival outcomes based on high-dimensional neuroimaging data. The RAPLS addresses the high-dimensionality and strong correlations among voxels by iteratively performing PLS-based dimension reduction, leading to higher prediction accuracy than the state-of-the-art methods.

Generalized Matrix Decomposition (GMD) for high-dimensional structured data: High- dimensional data with additional structures also arise in the emerging "omics" field. For example, genes interact with each other through gene networks, and in microbiome studies, bacteria are evolutionarily connected via the phylogenetic tree that shows their most recent common ancestry. In many scenarios, such additional structures is a priori avail; effectively using these structures may result in more efficient analysis and more meaningful interpretations. I first proposed a new graphical visualization tool for such high-dimensional structured data, called the GMD-biplot [3]. The GMD-biplot efficiently incorporates auxiliary structure information, leading to simultaneous visualization of the sample clustering and important biomarkers contributing to the observed sample clustering. I also developed a high-dimensional prediction tool that incorporates auxiliary structure information, called the GMD regression (GMDR), and a high-dimensional inferential tool, called the GMD inference (GMDI), that assesses the significance of individual variables within the context of GMDR [4]. The GMDI, to the best of my knowledge, is the *only* existing high-dimensional inferential tool that can detect *non-sparse* signals and account for highly structured/correlated variables. The GMD-biplot, GMDR, and GMDI have been successfully applied in microbiome, genomics, and neuroimaging applications.

Since I joined Arizona State University in 2020, I have been focusing on developing statistical methods for integrative analysis of multi-omics data and neuroimaging data to have a more comprehensive understanding of the genetic mechanism of human diseases. I also plan to devote my next 3-5 years to this research area. Below I discuss my immediate work and longer-term research plan.

Joint Matrix Decomposition (JMD) for supervised multiomics data integration: Diagnosis and treatment of human diseases require joint interpretation of molecular variations at multiple levels. For example, while the human microbiome has been proved associated with host health, limited knowledge is known about the underlying mechanism of these associations. One potential mechanism is microbial metabolism, which may affect host metabolic processes. Thus, an integrative analysis of the associations between the multi-omics (microbiome and metabolome) and host health may shed light on the biological mechanism of microbiome-host interactions. However, this analysis is statistically challenging because (i) it involves two potentially high-dimensional data sets, and (ii) it involves complex correlation structures, including between-microbe correlations, between-metabolite correlations, and microbe-metabolite correlations. As an immediate project, I plan to overcome these two challenges by proposing a high-dimensional data integration framework based on the joint matrix decomposition,

called the JMDR, to examine associations between multi-omics data and a broad class of outcomes, including continuous, binary, and count outcomes. The JMDR will be able to examine (i) whether the multi-omics variations have joint effects on the outcome, and if yes, which biomarkers contribute most to such joint effects; (ii) whether each omics variation has individual effects on the outcome, and if yes, which biomarkers contribute most to such individual effects. Thus, the JMDR has the potential to greatly facilitate the understanding of the interplay between the multi-omics variations and human health.

Leveraging auxiliary outcomes for more powerful biomarker detection: Recent development in imaging genetics offers unique opportunities to detect associations between genotype and specific brain structures as well as functions, potentially facilitating the diagnosis, prognosis, and treatment of neurodegenerative diseases. However, studies often have compromised powers of detection due to small sample sizes, especially when (high-dimensional) polygenic models are used. Thus, new statistical tools are needed for more powerful and reliable detection of brain-associated genes. Consider an example where it is of interest to detect genes associated with the olfactory cortex, a portion of the cerebral cortex critical for the perception of odor. In this case, one may improve the detection power by borrowing information from other brain regions strongly correlated with the olfactory cortex, such as the hippocampus, amygdala, and orbitofrontal cortex. As an immediate project, I plan to propose a high-dimensional regression model that can improve the power of signal detection by leveraging informative auxiliary outcomes, called the LASSO-Aux. The LASSO-Aux will enjoy broad flexibility as it will allow high-dimensional covariates (i.e., the numbers of genes considered in the model exceeds the sample size) and non-sparse signals (i.e., many regression coefficients are nonzero).

Statistical methods for investigating the gut-brain axis: My longer-term future research will focus on developing statistical methods for investigating the gut-brain axis (GBA). The GBA consists of bidirectional interactions between the gastrointestinal tract and the central nervous system. Recent advances have suggested the importance of the gut microbiome in influencing these interactions. Nonetheless, the mechanism by which the gut microbiome impacts the GBA remains largely unknown. Thus, statistical modeling of how microbes interact with each other and the brain functions is essential for a comprehensive understanding of the GBA. Despite recent progress in methodology for investigating the GBA, outstanding statistical and computational challenges remain. First, similar to imaging genetics, statistical methods for investigating the GBA must scale to two high-dimensional data sets, i.e., many microbes and high-dimensional neuroimaging phenotypes. Second, microbiome datasets generated by high-throughput sequencing are compositional because they have an arbitrary total imposed by the instrument. How to efficiently account for the compositional feature within the high-dimensional framework is a huge statistical challenge. Third, microbiome data have excessive zeros. Ignoring these zeros or simply replacing the zeros with pseudo-counts is ill-suited to zero-inflated microbiome data. Lastly, the discovery of novel microbiome biomarkers associated with neuroimaging phenotypes benefits from incorporating auxiliary omics data. Accordingly, there is an urgent need to develop powerful data integration tools for investigating the GBA to model multiple omics levels simultaneously. The overall objective of the project is to develop efficient statistical methods to investigate the GBA while addressing the above methodological challenges. I plan to attain the overall objective by pursuing the following aims: (i) Develop methods for inferring differential microbial networks to understand how the interactions among gut microbiomes differ between healthy population and patients diagnosed with neuro-degenerative diseases; (ii) develop high-dimensional models for discovering novel brain-associated microbiome biomarkers; (iii) develop methods for integration of microbiome, neuroimaging and other meta-omics data. Successful completion of these aims will generate critical biological insights into the mechanism underlying the GBA and thereby establish a statistical foundation for the development of microbiome-based precision therapies of neurodegenerative diseases.

References

[1] **Wang, Y.**, Ibrahim, J. G., & Zhu, H. (2020). Partial least squares for functional joint models with applications to the Alzheimer's disease neuroimaging initiative study. Biometrics, 76(4), 1109-1119.

[2] **Wang, Y.**, Ibrahim, J. G., & Zhu, H. (2021+), "RAPLS: Residual-based Alternative Partial Least Squares for Functional Regression Models.", *Under Revision, Biometrika*.

[3] **Wang, Y.**, Randolph, T. W., Shojaie, A., & Ma, J. (2019). The Generalized Matrix Decomposition Biplot and Its Application to Microbiome Data. Msystems, 4(6), e00504-19.

[4] **Wang, Y.**, Shojaie, A., Randolph, T. W., & Ma, J. (2021+). Generalized Matrix Decomposition Regression: Estimation and Inference for Two-way Structured Data. arXiv preprint arXiv:2104.08408. *Under Review, JRSSB*.