Research Experience

Aditya Mishra

My academic research focuses on developing novel statistical methods and computational packages to solve the challenging data-driven problems in the domain including, but not limited to, biological science. In past decades, advances in experimental capabilities have enabled the rapid collection of an unprecedented amount of data. In addition to using the existing statistical/computational methods for the data analysis, many biological problems pose a methodological challenge that requires developing new information retrieval techniques. As a Flatiron Research Fellow, I have developed techniques to analyze microbiome data: a) a robust regression model with compositional covariates, b) a negative binomial factor regression model, and c) an embedding model to learn microbial association network. In my Ph.D. dissertation work at the University of Connecticut, I have developed computationally efficient procedures for the multivariate analysis: a) sequential co-sparse factor regression model with an application on genomics data, b) generalized co-sparse factor regression with an application on aging data. The estimation procedures for all the methods are implemented, tested, validated, and made publicly available on my GitHub page (`https://github.com/amishra-stats/`) as user-friendly R packages.

At the Flatiron Institute, I am also associated with the Simons Collaboration on Computational and Biological Modeling of Marine Ecosystem (CBIOMES) project that brings together a multi-disciplinary group of investigators from oceanography, statistics, ecology, biogeochemistry, and remote sensing with the aims to understand marine ecology and underlying biogeochemical processes. My ongoing methodological research is inspired by the data problems originating in the CBIOMES collaboration as well. In the following, I will discuss my application-driven methodological contributions.

Microbiome Data Analysis

Robust Regression with Compositional Covariates

Biological data collected from high-throughput sequencing of 16S rRNA gene fragments, single-cell RNA-seq, metagenomes, or metatranscriptomes are intrinsically compositional, i.e., only relative abundances or compositions are measured and not absolute quantities. For instance, a typical microbiome study collects samples from the host in a habitat (such as ocean, soil, and human), and process it to obtain the relative abundance data of microbiome and host-associated features. Inferring parsimonious and robust statistical relationships between the abundance data and these additional host-associated measurements is often a first important step in the exploratory data analysis. The log-contrast regression provides an efficient framework to model a continuous response in terms of microbial abundance data as compositional covariates. However, outliers in the observations have a serious effect on the parameters estimate. In the current work on robust regression with compositional covariates (RobRegCC) [9], we extend the log-contrast regression framework to simultaneously identify the outliers and model the underlying association in terms of a subset of compositional predictors via a sparse estimate of the coefficient; see Figure 1 for workflow. The approach combines the idea of the mean shift model in linear regression with robust initialization and the regularized log-contrast regression [1]. We estimate the parameters of the over specified model using a regularization approach which solves the optimization problem with sparsity inducing penalty. Theoretical properties to depict a finite sample behavior of the estimator are shown in terms of a non-asymptotic error bound on the prediction and estimation error. We have implemented the procedure and make it available in the R package, RobRegCC.

Recently, we have developed a path-based algorithm to enhance the computational efficiency for solving the optimization problem in RobRegCC. In addition, we have also extended the algorithm to solve the robust classification problem with compositional covariates in which the objective function

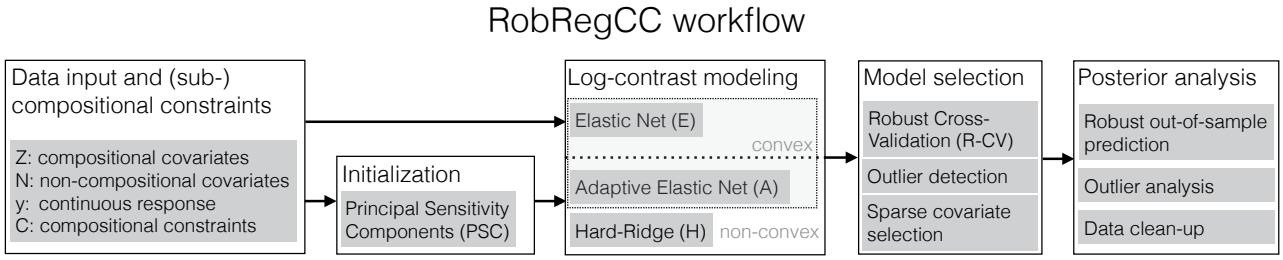

Figure 1: The RobRegCC workflow for robust regression with compositional covariates.

is designed using a squared hinge-loss function. With the computational efficiency, we plan to develop an exploratory analytic technique to learn a robust microbial association network. The method will extend the sparse inverse covariance estimation for compositional data [3] where an association network is obtained using the neighborhood selection framework. Motivated by the approach, the proposed procedure will require solving a node-wise RobRegCC problem.

Microbiome Embedding Model

Microbial abundance in the ocean is regulated by factors such as biogeochemical features, spatio-temporal conditions, and species-species interaction. With the aim to study the structure and function of the global microbiome, the Tara ocean expedition collected ocean water samples from distinct geographical locations across the globe [11]. The microbial abundance data is obtained using the high-throughput sequencing of the samples. Our interest is to learn about the association among species by jointly modeling the abundance in terms of host-associated environmental factors and species-species interaction components. To this end, we propose a probabilistic model using the parametric framework of the negative binomial distribution for the overdispersed count outcomes of the microbial abundance data [10]. The underlying mean parameters of the model are formulated in terms of interpretable components that drive the observed microbial abundance pattern, i.e., biogeochemical factors, spatio-temporal (location, depth and time) indicators, and species-species interaction. In particular, the model characterizes each of the species and spatio-temporal indicators in terms of latent vectors, and use it to design the species-species interaction component effect and spatio-temporal components effect. We use the variational inference framework to approximate the otherwise intractable posterior. In terms of the species-specific latent variables, the model provides an efficient framework to find out a similar set of species after discount the effect of biogeochemical and spatio-temporal components. We have used the probabilistic programming language Stan to obtain the variational posterior under the mean-field assumption. In the CBIOMES project, the model will have several applications to analyze the ocean microbiome data. Currently, we are working on setting up an analysis pipeline notebook in Google Colab to make it available to the domain experts.

Multivariate Analysis

In the multivariate regression, one of the major problems of interest is to model related responses using correlated predictors. Such problems are routinely required and formulated in various scientific investigations. For example, in the yeast cell cycle data, our interest is to understand the role of transcription factors in regulating the gene expression observed at several time points. In the high-dimensional setting, several studies have shown that a sparse singular value decomposition (SSVD) of the coefficient matrix is appealing for achieving dimension reduction and facilitating better model interpretation; see Figure 2 for the model. However, the problem of estimating the required SSVD structure remains challenging due to the presence of orthogonality constraint and co-sparsity (both left and right singular vectors are sparse) regularization, especially in large-scale problems or in the presence of incomplete data.

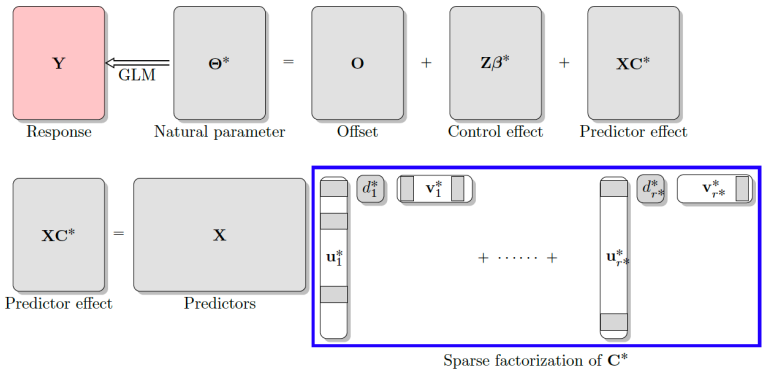

Figure 2: The multivariate model for the related responses matrix $\mathbf{Y}$ in terms of a correlated/high-dimensional predictor matrix $\mathbf{X}$ and a control variable matrix $\mathbf{Z}$ such that the low-rank coefficient matrix $\mathbf{C}^*$ have sparse singular vector components.

Sequential Co-sparse Factor Regression

We explore the connection between reduced rank regression and factor analysis and then develop a computationally efficient sequential estimation procedure to recover the required SSVD components of the coefficient matrix, named as sequential factor extraction via co-sparse unit-rank estimation (SeCURE) [7]. In a unit step of the sequential procedure, we suitably deflate the response matrix using the parameters estimated from previous steps and then solves a constrained unit-rank estimation (CURE) problem to estimate the subsequent co-sparse unit-rank component. The procedure completely bypasses the orthogonality requirements. Each of the estimated co-sparse singular vectors results in constructing a latent factor as a linear combination of a subset of predictors that affects only a subset of responses. The proposed algorithm is guaranteed to converge, and also works well to estimate the parameters in case of incomplete responses. Theoretically, we have shown that our sequential estimators are consistent and enjoy the oracle property asymptotically. The computational procedure is implemented in the R package *secure*.

Generalized Co-sparse Factor Regression

SeCURE provides an efficient framework for modeling the related continuous responses in terms of correlated predictors. The model assumes the underlying distribution of outcomes to be Gaussian types. However, several multivariate analysis problems require us to model the related responses that can be of mixed types, i.e., continuous, binary, and count. Depending on the type of outcome, one can assume a suitable underlying distribution from the exponential dispersion family, for example, Gaussian for continuous, Bernoulli for binary, and Poisson for the count, and estimate the parameters by maximizing the likelihood for each outcome separately. In a high-dimensional setting, the approach ignores the fact that some responses may be interrelated and some predictors may be either correlated or unimportant. Building upon the recent advances in the joint modeling of mixed outcomes [5], we extend SeCURE and proposes generalize co-sparse factor regression (GOFAR) model that also encode the underlying dependency through a low-rank and sparse coefficient matrix [8]. GOFAR proposes to estimate each of the unit-rank components of the coefficient matrix using either sequential or parallel extraction procedures. A unit step of both the approach solves a CURE problem with a suitably updated offset matrix that is designed to account for the effect due to other available unit-rank components of the coefficient matrix. Our approach integrates mixed and partially observed outcomes belonging to the exponential dispersion family, by assuming that all the outcomes are associated through a shared low-dimensional subspace spanned by the features. By combining the ideas of alternating convex search and majorization-minimization, an efficient algorithm with a monotone descending objective function is developed to solve the CURE problem. We have implemented the procedure in the R package *gofar*.

Because of the lack of suitable selection criteria, each of the unit-rank estimation problems chooses the tuning parameter using k-fold cross-validation. The estimation procedure is computationally intensive for large-scale problems. Authors in [2] proposed a computationally efficient stagewise learning procedure to extract unit-rank components sequentially in the case of SeCURE. Motivated by the research, one of my ongoing research focuses on extending the procedure for solving the constrained unit rank estimation problem in the case of GOFAR.

Negative Binomial Factor Regression Model

Microbiomes have developed a symbiotic relationship with their hosts over thousands of years of evolution. Several microbiome studies aim to understand this relationship with the host using the abundance data and the host-associated features data related to health, dietary intake, and demography. On considering the abundance data of count type as outcomes and host-associated features as covariates, GOFAR can provide an efficient framework to model the underlying dependency under the assumption that each of the responses comes from a Poisson distribution. However, the framework is not suitable to model the overdispersed count microbial abundance data. To this end, we propose a negative binomial reduced rank regression (NB-RRR) and a negative binomial co-sparse factor regression (NB-FAR) [6]. We encode the underlying dependency through a structured coefficient matrix. NB-RRR does so via a rank constrained coefficient matrix whereas NB-FAR does so via an SSVD of the coefficient matrix. NB-FAR mainly follows the sequential estimation strategy [8] as suggested in GOFAR to extract unit-rank components of the coefficient matrix. The optimization problems to estimate the parameters in each of the sequential steps of NB-FAR and the parameters of NB-RRR are non-convex. We propose an iterative procedure that updates the parameters in blocks via a surrogate majorizing the objective function. Using the data from the American Gut Project, we have applied NB-FAR to understand the association of microbial abundance to features related to health, dietary intake, and demography. The proposed procedure is implemented in the R package *nbfar*.

CBIOMES project: Software and statistical method developments for ocean sciences

Recent advances in technology have enabled the large-scale collection of biological, biogeochemical, and imaging data of the ocean on a global scale. In the CBIOMES project, the Simons CMAP database hosts multiple datasets from the ocean-related field study, remote sensing satellite, and forward model outputs. We have developed an R package "cmap4r" that allows users to download, analyze, and visualize data from the database. Analyzing the marine data is challenging due to several factors, including the presence of outliers, missing entries, different spatial and temporal resolutions, spatio-temporal dependencies, high-dimensionality, and the absence of absolute count of the ocean microbiome due to experimental limitations. This presents a unique opportunity for both the development and the application of novel statistical methods for the analysis of marine data. To answer some of these research questions and to help bridge the gap between statistics and oceanography I am currently focusing on applying network inference, sparse regression methods, and matrix factorization techniques. My ongoing projects with the CBIOMES group are focused on the application of existing methods and the development of novel statistical methods to analyze the data. For example, we are applying RobRegCC to find out microbial species that regulate primary productivity in the ocean and NB-FAR to associate microbial abundance to environmental factors.

My ongoing project with the group is to develop a workflow for marine data analysis. The project will demonstrate the usage of "cmap4r" to download marine data, and then apply a wide range of existing statistical tools and methods to perform various analysis tasks including regression, classification, network analysis, clustering, and change-point detection for time series.

References

[1] John Aitchison and John Bacon-Shone. Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330, 1984.

[2] Kun Chen, Ruipeng Dong, Wanwan Xu, and Zemin Zheng. Statistically guided divide-and-conquer for sparse factorization of large matrix. *arXiv preprint arXiv:2003.07898*, 2020.

[3] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol*, 11(5):e1004226, 2015.

[4] Wei Lin, Rui Feng, and Hongzhe Li. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288, 2015.

[5] Chongliang Luo, Jian Liang, Gen Li, Fei Wang, Changshui Zhang, Dipak K Dey, and Kun Chen. Leveraging mixed and incomplete outcomes via reduced-rank modeling. *Journal of Multivariate Analysis*, 167:378–394, 2018.

[6] Aditya Mishra, Andreas Buja, and Christian Müller. Negative binomials factor regression model for microbiome data analysis. *In preparation*, 2020.

[7] Aditya Mishra, Dipak K Dey, and Kun Chen. Sequential co-sparse factor regression. *Journal of Computational and Graphical Statistics*, 26(4):814–825, 2017.

[8] Aditya Mishra, Dipak K Dey, Yong Chen, and Kun Chen. Generalized co-sparse factor regression. *Computational Statistics & Data Analysis*, 157:107127, 2020.

[9] Aditya Mishra et al. Robust regression with compositional covariates. *arXiv preprint arXiv:1909.04990*, 2019.

[10] Aditya Mishra, Christian Müller, Jesse McNichol, and David Blei. Embedding model for learning microbial associations. *In preparation*, 2020.

[11] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237), 2015.