Zero-inflation in multinomial principal component analysis for microbiome data

Name, affiliation, and contact information for the supervisor and co-supervisor

Supervisor: Kevin McGregor, Department of Mathematics and Statistics, York University.

Email: kevinmcg@yorku.ca. Website: https://kevmcgregor.com/

Co-supervisor: Maxime Turgeon, Department of Statistics, Department of Computer Science, University of Manitoba.

Email: max.turgeon@umanitoba.ca. Website: https://maxturgeon.ca/

Co-supervisor: Saman Muthukumarana, Department of Statistics, University of Manitoba

Email: saman.muthukumarana@umanitoba.ca. Website:
https://www.samanmuthukumarana.com/

Abstract

Dimension reduction techniques are among the most essential analytical tools in the statistical analysis of genomic data. Generalized principal component analysis is an extension to standard principal component analysis (PCA) for non-Gaussian data and has been used in the analysis of data from genomic platforms such as single-cell sequencing and microbial/metagenomic sequencing. In particular, multinomial PCA has been adapted for use in these contexts. In microbiome data, however, there is often an abundance of zero counts, which is not accounted for in the multinomial PCA framework. In this project, the postdoctoral fellow will develop novel statistical methodology related to zero-inflated multinomial PCA, and to explore options for fitting the model through variational Bayes methods. The fellow will also have the opportunity for collaboration with Prof. Aleeza Gerstein from the University of Manitoba on the analysis of microbiome data coming from patients with recurrent vulvovaginal candidiasis infections (yeast infections).

Interdisciplinary/applied experience

The Post-doctoral Fellow (PDF) would work collaboratively with Prof. Aleeza Gerstein in the Departments of Microbiology & Statistics, University of Manitoba. Prof. Gerstein is undertaking a research project investigating the role of the microbiota in vulvovaginal candidiasis (yeast infection). The project involves collecting yeast isolates from women with recurrent infections and examining vaginal yeast and bacterial communities. There will be opportunities for the PDF to work with and present their work to individuals from a variety of backgrounds including other microbiologists in Prof. Gerstein's research group as well as clinical collaborators including OB/GYN, research nurses, and other health researchers.

In addition to the methodological work inherent to this application, the PDF will also assist Prof. Gerstein and her collaborators with data analysis and provide statistical advice in their own research. This will allow the PDF to develop skills in communicating statistical ideas to researchers in non-quantitative fields. Additionally, the PDF will be exposed to real-world datasets from a variety of platforms in the biomedical sciences.

Teaching/training/education

The successful applicant will have the opportunity to teach a one-semester course in each year of the program. At York University there are many opportunities for postdoctoral fellows to teach introductory courses in statistics and probability theory for statistics majors, as well as service courses such as probability and statistics courses intended for engineering or business majors. Similarly, the University of Manitoba has several open positions for sessional lecturers every year. These positions target first- and second-year courses in statistics, probability, as well as service courses. Moreover, the PDF may have the opportunity to teach a third- or fourth-year topics course. All these opportunities will provide valuable teaching experience to the PDF.

Mentoring of the postdoctoral fellow

As part of Prof. Turgeon's and Prof. Muthukumarana's research groups at the University of Manitoba, the PDF will join an interdisciplinary group working at the intersection of statistics and computer science. This will provide them with the opportunity to develop and hone their computational skills, which are in high demand both in industry and academia. The successful applicant will also interact with two larger groups of researchers: the Complex Data Lab, comprised of faculty members in statistics and their students interested in statistical learning and computational statistics; and the Machine Learning Interest Group, comprised of researchers in statistics, computer science, engineering, and biomedical sciences. These interactions will provide opportunities for collaboration, networking, and for presenting their work.

At York University, Prof. McGregor is actively involved in the Statistical Consulting Service (SCS) housed in the Institute for Social Research. The PDF will be involved in the SCS at York. This will entail doing several hours of consulting work every month. The PDF will attend SCS group meetings where various issues regarding consulting and statistical analysis are discussed.

Profs. McGregor, Turgeon, and Muthukumarana will hold weekly meetings with the PDF to discuss the project details and provide advice. Additionally, the PDF, supervisor and co-supervisors will hold monthly meetings with Prof. Gerstein in order to make sure the direction of the project takes Prof. Gerstein's research needs into consideration. Prof. McGregor has extensive experience in the development of multinomial models in biological sequencing data and will therefore provide guidance on the applied aspect of the project. Prof. Turgeon has extensive experience in dimension reduction techniques in biological data and will provide guidance on the theoretical aspect of the project. Prof. Muthukumarana has extensive experience with Bayesian methods and computation for complex models which integrate multidisciplinary applications

York University offers many opportunities for Equity, Diversity, and Inclusion (EDI) training. Examples of EDI workshops include *Introduction to the Centre for Human Rights* and *Creating & Maintaining Positive Space,* and *Challenging Unconscious Bias and Microaggressions*. The PDF will be expected to attend several such workshops during their first year at York University.

Proposed schedule for postdoctoral fellowship

Schedule for year 1: The PDF will spend the first year of the fellowship at York University in Toronto. The PDF will have the following objectives for the first year:

1. Gain familiarity with specifics of microbiome/metagenomic sequencing platforms as well as theory behind gPCA.
2. Develop methodological techniques for the zero-inflated multinomial PCA problem.
3. Write an R package to employ zero-inflated multinomial PCA.
4. Submit results on zero-inflated multinomial PCA to a statistics journal by the end of the first year.
5. Teach an introductory statistics course in the Mathematics and Statistics department at York University.

Schedule for year 2: The PDF will spend the second year of the fellowship at The University of Manitoba in Winnipeg. The PDF will have the following objectives for the second year:

1. Integrate variational Bayes techniques for zero-inflated multinomial PCA.
2. Run the developed zero-inflated multinomial PCA model on vaginal yeast and bacterial microbiome sample data.
3. Submit results of variational Bayes techniques in zero-inflated multinomial PCA to a statistics journal by the end of the second year.
4. Assist Prof. Gerstein's research group in the analysis of collected data.
5. Teach a statistics course in the Statistics department at University of Manitoba.

A list of qualifications of suitable candidates

The successful applicant will have a PhD in Statistics, Biostatistics, or in a health-related field with a strong quantitative background. The applicant will also have demonstrable computational experience, including programming in R or Python. Prior knowledge of microbiome and/or sequencing platforms is not required but will be considered an asset.