MATH RESEARCH AT UNC CHARLOTTE 2024

Project 1: Zero-inflated models for sparse data

Mentor: Dr. HeeCheol Chung

Project description. In biological data analysis, high-throughput sequencing offers an unmatched level of precision in transcript quantification, although it comes with the trade-off of amplifying the influence of technical noise. Achieving a reliable reduction in the random background noise while capturing biologically meaningful signals remains a persistent challenge.

Data from these sequencing techniques often illustrate highs skeweness and contain an excess number of zeros. While some scientists see these zeros as signs of very low or absent gene activity and others see them as missing information (Jiang et al., 2022), dealing with these zeros properly is essential for trustworthy statistical analysis. To tackle this, different probabilistic models including zero-inflated Poisson (Witten, 2011), zero-inflated negative binomial (Dong et al., 2016), Hurdle (McDavid et al., 2019), truncated Gaussian (Ma, 2021), and latent Gaussian copula models (Yoon et al., 2020), have been proposed. Nevertheless, extreme skeweness and excessive number of zeros in sequencing datasets often exceeds the tolerance of these models, and thus, figuring out which model works best in different situations becomes crucial.

REU Students' role and start-up topics: This project invites students to delve into diverse zero-inflation models using both simulated and real datasets. We seek students with a background in probability distributions and statistical computing to contribute to this endeavor. Proficiency in fundamental statistics, statistical simulation concepts, and R programming is essential.

During the initial 2-3 weeks, students will acquire foundational knowledge of zero-inflation models and develop simulation codes. Subsequently, their focus will shift towards comparing these models using simulated and real datasets, exploring their efficacy in handling zero-inflated data.

References

Dong, K., H. Zhao, T. Tong, and X. Wan (2016). NBLDA: negative binomial linear discriminant analysis for RNA-seq data. *BMC Bioinformatics 17*(1), 1–10.

Jiang, R., T. Sun, D. Song, and J. J. Li (2022). Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome biology 23*(1), 1–24.

Ma, J. (2021). Joint microbial and metabolomic network estimation with the censored gaussian graphical model. *Statistics in biosciences 13*(2), 351–372.

McDavid, A., R. Gottardo, N. Simon, and M. Drton (2019). Graphical models for zero-inflated single cell gene expression. *The annals of applied statistics 13*(2), 848.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics 5*(4), 2493–2518.

Yoon, G., R. J. Carroll, and I. Gaynanova (2020). Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika 107*(3), 609–625.