STAT 548 PhD Qualifying Course Papers

Keegan Korthauer

Updated: December 1, 2020

Choosing a paper

In this document, I provide a list of suggested papers and projects I am interested in supervising. However, I am open to discussing papers and projects from outside this list that relate to my research interests. I am broadly interested in methodology for the statistical analysis of high-dimensional biological data. In particular, I develop and apply methods for inference of the genetic/epigenetic basis of disease, as well as mechanisms of gene regulation, from sequencing data (e.g. from DNA, RNA, or epigenome). I am also interested in multiple testing problems in high-dimensional biology. My preferred computing language is R, but we can discuss the possibility of using other languages if necessary.

Report[1]

Summary (max 5 pages): The first section of the report should provide a summary (max 5 pages) of the problem the paper addresses in the context of previous work, limitations of previous work, the solution technique, important results, why they are important, and limitations of the paper. The paper you select may have multiple contribution areas (theory, modeling, computation, biological insights), and the summary should reflect that. The goal of this portion of the report is to show that you can take a complex body of work (one paper and earlier relevant work), digest it, and present a concise summary of the important points.

Project (no page limit): The remainder of the report (no page limit) will be devoted to a paper-specific mini project. This mini project should build upon the paper in some way. This could involve identifying a problem with the method and proposing an improvement, or extending the method for application in a different context. Your approach might use simulation studies and/or publicly available data. Some ideas for suggested projects are included in the list of papers below (you should feel free to stray from these ideas). Please schedule a meeting with me to discuss these and/or other ideas you may have. Note that your grade will not be affected by how good the results look, whether your approach improves on past work, or whether you achieve the initial goal of the project. Your grade for this section will instead be based upon how you approach the problem and how you evaluate and communicate the results[2].

[1] Adapted from: Trevor Campbell and Marie Auger-Méthé

[2] See also https://www.stat.ubc.ca/phd-qualifying-course

Project organization

All results in your report should be reproducible. Reproducible research requires that your project is organized and well-documented. The ultimate goal is that, given your code and report, another researcher (or your future self!) can and obtain and comprehend the same results.

All code used in this project should be housed in a GitHub repository (for which you should add me as a collaborator before submitting the final report). You are encouraged to use git to commit regularly. This helps me keep track of your progress, but also serves as a backup in case you need to revert to an earlier version.

The directory structure should be organized to separate any raw data from derived results. For example, here is a suggested directory structure to organize your project files[3]:

```
/README.md          <- The top-level README with summary of project
/data
   /raw             <- The original, unchanged data
   /interim         <- Intermediate data that has been transformed
   /processed       <- The final processed data sets for downstream analysis
/references         <- Articles and manuals used for reference
/report             <- Latex files and generated PDF of report
/figures            <- Generated graphics and figures to be used in report
/src                <- Source code: subdirectories will vary by project type/scope
   /01_data         <- Source code to obtain / generate data
   /02_process      <- Source code to process data for downstream analysis
   /03_analysis     <- Source code for downstream analysis
   /04_report       <- Source code to generate final figures/tables in report
```

Your report should be written in Latex. Please make sure to submit all necessary files (.tex, .bib, any style files) necessary for me to compile, as well as the pdf.

Resources on reproducible research

- Karl Broman's guide to reproducible research: https://kbroman.org/steps2rr/
- Jenny Bryan's git manual (for those new to Git/GitHub): https://happygitwithr.com

Available papers (last updated December 1, 2020)

Please send me an email if you have trouble accessing any of these papers, and I can send you a PDF.

1. **Inference of differential methylation for sequencing data**
   Paper: Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016 May 15;32(10):1446-53.
   Link: https://academic.oup.com/bioinformatics/article/32/10/1446/1743267
   Project idea: How (if at all) is type I error controlled for differentially methylated region (DMR) detection? Use simulation to evaluate the False Discovery Rate (FDR) for DMR detection using DSS.

[3] Adapted from https://drivendata.github.io/cookiecutter-data-science/#directory-structure

2. **Evaluation metrics for binary classification (TAKEN)**
   Paper: Cao C, Chicco D, Hoffman MM. The MCC-F1 curve: a performance evaluation technique for binary classification. arXiv preprint arXiv:2006.11278. 2020 Jun 17.
   Link: https://arxiv.org/abs/2006.11278
   Project ideas: (1) Does the severity of imbalance affect some metrics more than others? Investigate analytically and/or using simulation. (2) Discussion point: Should the positive/negative balance of a dataset influence which classifier is considered optimal, or can one classifier be determined to be universally better than another?

3. **Covariate-weighted False Discovery Rate control**
   Paper: Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nature methods. 2016 Jul;13(7):577.
   Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930141/
   Project ideas: (1) Evaluate performance relative to number of tests - suggest rationale for any variation seen and discuss possible improvements. (2) Suggest possible extensions to handle multiple covariates

4. **Normalization of single-cell RNA-seq data (TAKEN)**
   Paper: Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome biology. 2016 Dec;17(1):75.
   Link: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0947-7
   Project idea: Explore the impact of choice of clustering method on normalization of heterogeneous datasets - summarize your findings with recommendations.

5. **Inference of differential expression for single-cell data**
   Paper: Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome biology. 2016 Dec;17(1):222.
   Link: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1077-y
   Project idea: Select and evaluate an alternate metric of distributional distance.