

STATISTICS USRA PROJECTS 2018

1. Drug Classification using Statistical/Machine Learning

The project relates to drug discovery in connection with a team of scientists at Vancouver Prostate Centre. Statistical/machine learning methods are being explored for classification of drugs as inactive/active, non-toxic/toxic, etc. In addition, new variables to characterize the chemical structures of drug compounds, such as pixels from images, are being explored for utility in classification.

The work will explore the above methods, comparing them in a systematic way. Facility with (1) scripting in R, Matlab, or python and (2) classification and validation at the level of STAT 406 or equivalent is essential. The research also requires an organized approach, to keep multiple analysis strategies transparent and reproducible.

Interested students should contact Dr. Will Welch at will@stat.ubc.ca, and provide a cover letter and unofficial transcripts.

2. Forest Products and Lumber Strength

BACKGROUND: Wood is one of the Earth's sustainable resources and it will be around for a long time after other resources have been depleted. Thus there is much interest these days on Forest Products including their structural engineering and other such properties. But as wood is an organic material, these properties are highly variable making the need for statistical methods for analysis much in demand. The Department of Statistics has engaged in a long time collaborative research project with industrial research lab on campus (but not part of UBC) and over the past few years has developed a lot of modern tools for handling such analyses. For example, we have helped develop a long-term monitoring network for assessing the strength of lumber in Canada and we are on the verge of completing the development of a high tech method for classifying sawn lumber by its strength property, one that can be used in saw-mills that are largely operated these days by computers.

THE USRA: ASDA is seeking one or more UBC statistics undergraduate majors or honors students to work during the summer period of 2018 (May 1 to Aug 31) on a research team that has a number of Statistics faculty and grad students involved. As a member of the team he/she will: help review and write some relevant background material about the strength properties of lumber; help analyze data about lumber properties; explore earlier work about the statistical properties of lumber and apply extreme value models using the R package to analyze such data. The USRA would be expected to write a brief report at the end of the summer that describes his/her work on the team and some of the findings-you might even end up being a co-author of a research paper! As you would be working not only with statisticians but as well with engineers and wood scientists. you should have good English language communication skills. You will also need a reasonably good background in computing and statistics.

BENEFITS. The student will be paid at the standard USRA rate. You will develop the skills associated with handling real data, its management as well as its analysis. In particular, you will enhance your programming and statistical analytic skills and learn a lot about collaborative research and teamwork.

APPLICATION: Please email Dr. Jim Zidek at jim@stat.ubc.ca with any questions you may have or to express an interest in joining the team. In the latter case please provide a brief resume that describes your background, notably the courses you have taken and the grades you received in these courses. You will also need to provide an unofficial copy of your transcripts. All material will be treated as strictly confidential. An interview will be arranged for shortlisted candidates.

3. Probabilistic Programming Languages

Probabilistic programming languages (PPL) allows rapid prototyping of complex models and is a current hot topic in computational statistics and Bayesian data science. Our lab is developing a PPL based on new developments in the field of non-reversible Monte Carlo methods. The student could get involved in a variety of aspects of this projects ranging from designing inference engines, to creating probabilistic models and doing data analysis.

I am looking for students with particularly strong programming skills. Previous exposure to JVM languages such as Java, Scala or Xtend highly recommended.

Interested candidates should contact Dr. Alexandre Bouchard-Côté at bouchard@stat.ubc.ca, and provide a cover letter and unofficial transcripts.

4. Deep Learning

Determining the binding locations of transcription factors (TF) is important for advancing our understanding of genome regulation. Technology such as ChipSeq enables locating specific TF binding sites, but requires a large biological sample, which is sometimes not feasible. In contrast, ATACseq requires a much smaller biological sample, but can only identify open chromatin regions. We are thus interested in building deep learning (DL) models for predicting TF binding sites from DNA sequences and other features of ATACseq peaks that signifies open regions. In addition, we are interested in deciphering the properties of open regions that give rise to their binding affinity, i.e. devising methods for interpreting DL models.

Interested students please contact Dr. Mostafavi: saram@stat.ubc.ca and provide a cover letter and unofficial transcripts

5. Manifold Regression

Standard regression models assume the input variables live in the Euclidean space. However, many data types come in other forms, e.g. brain connectivity is typically represented as covariance matrices. We are interested in developing models and hypothesis testing procedures for the case where responses are covariance matrices, and evaluating whether modeling the manifold nature of the data would enhance our analysis. Our targeted application is to associate genetic variants to large-scale functional networks. Although a few recent methods can handle covariance matrices as responses, the required computational time restrict the number of genetic variants that can be tested. Also, the large number of genetic variants presents a huge multiple comparison burden, which limits the statistical power. Hence, more efficient methods that permits testing of association between variant-sets and covariance responses are necessary.

Interested students please contact Dr. Mostafavi: saram@stat.ubc.ca and provide a cover letter and unofficial transcripts