Robust inverse normal transformation-based tests under linear mixed effects models

Sonja Friesen, University of Guelph

Many biomedical applications generate vast amounts of data that cannot be managed under human oversight alone, necessitating automated analysis pipelines. These pipelines require robust statistical methods to handle frequent assumption violations in biological settings. Our research is motivated by the International Mouse Phenotyping Consortium (IMPC), which examines mammalian gene function using genetically edited knockout mice. The IMPC employs a linear mixed effects model (LMM) that considers both fixed effects, such as gene, sex, and weight, and random effects that may arise from shared experimental batch or litter. Using simulations based on the IMPC experiment setup, we assess the robustness of the LMM under violations of the normality assumption for error and random effects. We further investigate whether rank-based inverse normal transformations could enhance the detection power of gene effects in these scenarios.

A Closed-Form Correction for Attenuation Bias in Imputed-Genotype Association Studies

Ayisha COK, University of Guelph

Genotype imputation is now standard in genome-wide association studies, yet using imputed genotypes in place of true genotypes in association testing has statistical consequences that are easy to overlook. Whether the imputed genotype is summarised as a dosage or as a best guess, the estimated genetic effect is attenuated toward zero and the residual variance is inflated, so both the effect estimate and its standard error, and hence any inference based on them, are biased when imputed genotypes are analysed directly. In this work, we derive a bias-corrected estimator of the genetic effect and its standard error using imputation quality metrics and genotype variance under Hardy-Weinberg equilibrium, both available from standard imputation workflows. Across a range of effect sizes, allele frequencies, and imputation qualities, our numerical studies show the corrected estimator recovers the true genetic effect, with the attenuation almost entirely removed. Association studies usually focus on power, but our goal is unbiased estimation, so the reported effect reflects the true genetic signal rather than an artefact of imputation. The same attenuation appears wherever an error-prone measurement substitutes for a true quantity. This makes it a single statistical problem shared across the interconnected domains of One Health, from host genetics to the microbiome and environmental surveys.

Flexible Tree Ensembles with Data Adaptive Splitting Rules

Heshani Mendis, University of Manitoba

Accurate split selection in regression trees and random forests is critical for predictive performance. In regression trees, splits are selected by minimizing within-node variances weighted by node proportions. These weights directly influence the impurity reduction criterion. We modify the standard weighting scheme by applying power transformations to node proportions and examine how this modification influences split selection and predictive accuracy. Based on the proposed weighting framework, we develop four algorithmic variants that introduce flexibility at the node, tree, and forest levels, along with a randomized weight-selection strategy. The proposed framework is evaluated via simulations across a range of data-generating mechanisms, including varying signal strengths, noise levels, functional relationships, and covariate dependence structures and distributions. Performance is assessed using relative prediction error against the default weighting scheme, and the results demonstrate notable improvements, with the tree-level flexible method achieving improved or comparable predictive performance relative to the standard random forest in the majority of scenarios considered. To illustrate practical applicability, the proposed framework is applied to real data, including a Parkinson's disease telemonitoring dataset.

Prediction and Model Assessment of Regression Models with Circular Data

Nimsara Dissanayaka, University of Guelph

Angular or circular data arise in many applications, such as environmental studies, where observations are recorded as directions or angles. Examples include wind direction, animal movement, bird migration, and ocean current direction. In such cases, statistical analyzes must properly account for the circular nature of the observations. In this work, we study the consequences of ignoring circularity by comparing the predictive performance of parametric circular regression models with that of traditional linear models in settings with circular responses and/or circular covariates. The results of numerical studies indicate a substantial loss in prediction accuracy when circularity is ignored. We further evaluate parametric circular regression models by comparing them with their nonparametric counterparts under various circular distance metrics and predictive accuracy criteria. The analysis is illustrated using wind direction data to demonstrate the importance of accounting for circularity and its potential applicability to other directional data, such as animal movement.