Leveraging large language models in a composable system to design and simulate experiments

Emi Tanaka
The Australian National University

Abstract

Designing and simulating experiments requires consideration of a number of factors. We present a composable system that leverages large language models (LLMs) to streamline the end-to-end workflow for designing and simulating experiments. Through a modular, composable interface built on the edibble R package, experimenters can assemble, modify, and reuse experimental components enabling rapid prototyping, customization, and replication of complex experimental designs. The extended system, built on a LLM, automates simulation of outcomes, leverages historical data and information, and provides context to support informed experimental design decisions. I demonstrate this system to explore potential genetic gains, and accelerate cultivar development by enhancing rigor, adaptability, and discovery in plant breeding research.

Title:
Scalable finite mixture of regression models for community ecology

Abstract:
When modeling species communities in ecology, clustering species with similar responses to the environment can facilitate a more parsimonious understanding of the community, and improve prediction by borrowing strength across species within the same cluster. One statistical method for achieving the above is species archetype models (SAMs), a type of finite mixture of regression model where species are clustered according to the **shape** of their environmental response.

In this talk, we introduce approximate and scalable SAMs or asSAMs, which overcomes some of the current computational drawbacks in fitting SAMs. We demonstrate how asSAMs promotes fast estimation and uncertainty quantification via bootstrapping. Simulation studies and an application to presence-absence records of over 230 species from the Great Barrier Reef Seabed biodiversity project demonstrate asSAMs can achieve similar to or better estimation, selection, and predictive performance than several existing methods in the literature.

Speaker:
Francis K.C. Hui (joint work with Patricia Menendez, Scott Foster, and Skipton Woolley)

Sparse and Multi-Dimensional Fréchet Clustering for Longitudinal Trajectories

Soon-Sun Kwon

Department of Mathematics, Ajou University

Abstract

Multi-dimensional longitudinal data in clinical research often exhibit irregular and asynchronous measurements, making clustering and interpretation challenging. We introduce two novel extensions of the Fréchet distance: Multi-dimensional Fréchet K-means for Longitudinal Data (MFKmL) and Sparse Fréchet K-medoids (SFKmL), the latter enabling functional variable selection. These scalable methods effectively group irregular trajectories while identifying variables with the highest discriminative power. Simulation studies confirm their robustness, and applications to longitudinal thyroid cancer data from South Korea reveal clinically relevant patient subgroups and key prognostic biomarkers.

Keywords: Functional data analysis, clustering, Fréchet distance, thyroid cancer.

Copula modelling for flood risk assessment in Johor River Basin, Malaysia

Shariffah Suhaila Syed Jamaludin
UTM Centre for Industrial and Applied Mathematics
Universiti Teknologi Malaysia
Johor Bahru, Johor.
Malaysia

Flood risk assessment in regions like Malaysia's Johor River Basin demands a comprehensive understanding of the complex relationships among multiple flood characteristics, such as peak flow, duration, volume, and time to peak. Traditional univariate frequency analysis often fails to capture these interdependencies, especially under non-linear and extreme conditions. Copula modelling provides a flexible statistical framework that separates the modelling of marginal distributions from the dependence structure, enabling a more accurate joint probabilistic assessment. This talk introduces a brief theory and application of copulas for flood risk assessment, beginning with the fundamentals of Sklar's theorem and dependence measures, and progressing from bivariate to trivariate and vine copula models. Using flood data from the Johor River Basin, the methodology is demonstrated to quantify joint probabilities and estimate return periods under different dependency structures.