BSc. + MSc. Thesis opportunities in the Chair of Statistics:
Regression for Non-Euclidean data

Are you curious of how to deal with complex data that goes beyond real numbers? We have 3 Thesis offerings to explore regression models whose response has a complex structure. More info here: Website: Chair of Statistics

1) Bayesian Compositional regression

In this project we study the Bayesian approach to Compositional regression using Liesel. Compositional data refers that represent proportions of a whole. That is a compositional data point can be represented by as a real-valued vector constrained to sum up to 1. Mathematically this is define in a simplex:

$$\Delta^d = \{x = [x_1, x_2, ..., x_d] \in \mathbb{R}^d \mid \sum_{i=1}^{d} x_i = 1 \; ; \; x_i > 0\}$$

This type of data naturally arises from Econometrics (e.g. Household expenditure survey), or biology (Microbiome data) to name a couple. You would then implement and model the data using a probability model define in the simplex Δ^d. Depending on the structure of the data, potential models include: VARs, additive models, or spatio-temporal models, among others. The model then would be implemented in Python using the Liesel framework.

2) Leaf shape heterogeneity analysis

Leaf shape variability is an important characteristic of plant development and health. It is driven mostly by genetic and environmental factors. Studying and modelling such variability can be used to understand and forecast the growth of the tree. Fresh leaves from juvenile beech trees grown in an experiment were harvested at the same time of the entire tree in August-September 2021. For each tree, all leaves were harvested and a sample of 60 to 120 leaves were scanned on a flatbed scanner. Thus, this dataset corresponds to the raw images of the scan. The aim of this project is to extract the shape, size, and average colour of the leaf images and estimate the Fréchet mean and variance via the elastic metric approach. Further, we study the modes of variation through Geodesic Principal Component Analysis (GPCA).

3) Score Matching for Directional regression

In this project we aim at developing and implementing the Score Matching estimator in the directional/circular regression models. Substantial development has been done. Score matching estimators are widely used in directional distribution due to the fact that the normalizing constants are usually computationally intractable. Common applications of directional statistics are Forestry, Ecology, and Palaeomagnetism among many others. In this project we will focus on the Kent distribution, also know as the 5 parameter Fisher-Bingham, due to its flexibility, and nice functional form. To the best of my understanding, score matching estimator has not been implemented in the most popular libraries, both in R, Python, and Julia. Your task would be develop, and implement the estimator in your language of choice. And then test the result using either simulated or real data.

If you are interested, please contact Alejandro Pereira (alejandro.pereira@uni-goettingen.de)