1 Quantification of Impact of New Cancer Therapy

Consider that you are working as a data scientist at a biological data science consultancy. The company you work for has been contracted by a pharmaceutical company, TheraTech, to investigate the effect of a new drug they are developing on a type of cancer. TheraTech has provided the following scientific background information.

1.1 Scientific Background

The new drug TheraTech has developed is believed to have a long-term impact on an overall reduction in cancer in patients. However, the form of cancer that this drug treats presents itself through a variety of factors. These factors range from macroscopic measurements, such as those obtained from a biopsy (removing a piece of tissue so that it can be tested in a laboratory), to the concentration of specific chemical species in a patient's urine. The direct measurement of the overall effect of the treatment is difficult.

The working hypothesis from TheraTech is that their new drug leads an impact on a total of **six** cancer markers. The clinicians at TheraTech are struggling to identify a quantitative relationship between any of the markers alone. However, the overall effect on the patients indicated a reduction in cancer. Due to concerns regarding intellectual property, the TheraTech team cannot share specifics of the cancer markers with you. Instead, they have shared the raw numbers as a series of `.csv` files with headings of `marker_*`, where the `*` indicates a value between 0 and 5.

In total, two data files have been provided by TheraTech:

- `pre_cancer_markers.csv`: The cancer markers **before** the patients have been treated.
- `post_cancer_markers.csv`: The cancer markers **after** the patients have been treated.

Each file contains anonymous data from 2000 patients that were studied. The team at TheraTech highlighted the **first sample** in the `post_cancer_markers.csv` dataset presented a significant overall improvement in cancer symptoms. TheraTech has stated that these data have been normalised with respect to appropriate controls, and therefore, changes observed should be a result of the treatment alone.

1.2 Program of Work

In order to understand the impact of the new drug on the provided cancer markers, it is believed that some **coherent** effect is present. This coherent effect will likely be correlated across the different cancer markers, involving some mix of different amounts of the markers. Additionally, the complex dataset that TheraTech provided has been challenging to discuss in detail with clinicians.

Assessment: You have been tasked with the development of a quantitative model to interpret the measured data. In particular, TheraTech would like to know if it is possible to quantify how many patients receiving the drug presented an overall reduction in their symptoms. If possible, what percentage of patients improved? Additionally, they would like a simple analysis dashboard to help understand the model that you develop. This dashboard should:

- Allow the TheraTech team to visualise the complex datasets as two-dimensional plots.
- Include interactive widgets that show the different dimensions of your analysis.
- Enable your team to communicate your results to the clinical team.

TheraTech would like a report that describes in detail your findings and outlines these results should be presented with your dashboard.