# Epi-on-the-Island

## Multivariate Visualization and Analysis

## 18-22 June 2018

**Tentative Schedule**

| Day | Time | Lecture | Laboratory |
|---|---|---|---|
| Mon | 8:30 - 10:00 | Introduction to the course; Introduction to visualization | |
| | 10:30 - 12:00 | Exploration of visualization tools | Hands-on demonstrations |
| | 1:30 - 3:00 | Introduction to Python, key libraries and data structures | |
| | 3:30 - 5:00 | | Using Python to visualize multivariate data lab |
| | 5:00 - 6:00 | Get familiar with participants' data, and assist with data import | |
| Tues | 8:30 - 10:00 | Multivariate distance measures, multidimensional scaling and unsupervised hierarchical clustering | |
| | 10:30 - 12:00 | | Hierarchical clustering lab |
| | 1:30 - 3:00 | Partition-based clustering, and other clustering problems (e.g. dynamic time warping) | |
| | 3:30 - 5:00 | | Partition-based and other clustering lab |
| | 5:00 - 6:00 | Discuss data structure/quality with participants in relation to their study | |
| Wed | 8:30 - 9:15 | Introduction to dimension reduction approaches, especially principal components analysis | |
| | 9:15 – 10:00 | | Principal components analysis lab |
| | 10:30 - 12:00 | Multiple correspondence analysis | |
| | 1:00 - 2:00 | | Multiple correspondence analysis lab |
| | 2:00 - 3:00 | Discussion of normalization, rotation, and data requirements for dimension reduction approaches | |

| Day | Time | Lecture | Laboratory |
|---|---|---|---|
| | 3:30 - 5:00 | | Advanced dimension reduction techniques lab |
| | 5:00 - 6:00 | Participants free to work on own data | |
| Thur | 8:30 - 10:00 | Introduction to classification based on categorical outcomes, including regression and support vector machine techniques, as well as dimension reduction for predictors | |
| | 10:30 - 12:00 | | Classification with predictor dimension reduction lab |
| | 1:30 - 2:30 | Classification and regression trees and random forests, and other tree-based and ensemble methods | |
| | 3:00 - 4:00 | | Comparings results from a range of classification methods lab |
| | 4:00 - 5:00 | Measures of 'success' for classification and multivariate methods in general | |
| | Evening | Course dinner | |
| Fri | 8:30 - 10:00 | | Participants work on own or provided data |
| | 10:30 - 12:00 | | Participants work on own or provided data |
| | 1:30 - 3:00 | Presentations by participants | |
| | 3:30 - 5:00 | Presentations by participants; Course wrap-up | |

## *Course Information*

### Objective:

To expose participants to a range of techniques for visualising complex multivariate data, based on a suite of modern and classical multivariate analysis approaches. The aim of the exposition is to convey an understanding of the methods without necessarily going into heavy mathematical detail, to the level where participants will be able to apply such methods to their own data and critically assess the results.

To carry this out within, and thereby introduce participants to, the major 'data science' open source platform that currently competes with R (i.e. Python) so as to increase the likelihood of finding novel/alternate ways of thinking about analysing complex data sets, while providing supplementary pointers to similar implementations (when existing) within standard statistical software.

### Text:

The course will not follow a single textbook (because standard texts do not include the topics covered and/or at the targeted level of mathematical detail), and handouts will be provided for lectures and labs. However, the following texts cover some of the topics (and are recommended as good sources for further study for the advanced learner):

- Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*, 2nd ed, 2009. (freely available at https://web.stanford.edu/~hastie/ElemStatLearn/)
- Izenman AJ: *Modern Multivariate Statistical Techniques*, 2008.

### Software

The primary software used in the course will be the **Python 3** programming environment, with heavy use of relevant libraries, in particular numpy and scikit-learn. All participants will be expected to work on their own laptops and to have set up an appropriate Python working environment on these computers. If participants are already familiar with Python then the main set-up task will be to ensure that they are using Python 3.x (we will be using 3.6) and have the relevant libraries loaded. For those not familiar with Python, and even for some who are, we would advise using the Anaconda data science distribution, which includes all the scientific packages as well as the Jupyter Notebook IDE (https://www.anaconda.com; make sure to get the Python 3.x version).

Around a month prior to the course, participants will be sent more detailed instructions as to how to go through the installation process and also given links to some optional tutorial material for those who wish to develop a little familiarity with the Python environment before the course begins.

As far as possible, the laboratories and demo code used within the course will also be provided in a format to run within the R environment. We will not be using R within the course but realise that a number of participants may be more familiar with the R programming framework and as such plan to provide this 'parallel' set of resources for participants to explore in their own time and take away from the course.

## *Course Preparation*

In order to get the maximum value out of the multivariate visualization course, we encourage participants to bring their own data with them to the course. There will be time during the course to work on your own data, and we will endeavour to have lots of help available in the lab sessions to expedite this process. The following actions are recommended:

**1. Prepare a 1 page description of your data / problem** using the template attached (last page). These will be copied at the beginning of the course and may be distributed to all course participants.

2. **Prepare your data (if you are bringing some)**: If you have data of your own which you would like to work on during the course, please give some thought regarding the dataset that you are bringing with you. Some suggestions for preparing the dataset include:

(a) where possible use a standard data worksheet layout with observational units as rows and variables recorded per observational unit as columns;

(b) ensure that each observation can be uniquely identified (e.g. herd_id, cow_id, sample_number). This is particularly important if you plan to link multiple datasets together (i.e. to have primary and 'foreign' keys defined for each set);

(c) identify the variables of main interest and think about bringing a dataset with just these variables in it (rather than bringing the whole dataset if it is very large);

(d) you can bring the data in any digital format, but when importing such datasets the general rule is that the simpler the better. Thus a simple .csv export may work best in many settings. However, if your data are more complex and you have these in a database (e.g. MySQL, SQLite) or JSON format this will also work;

(e) the most important dataset-linked task is probably to ensure that you bring a 'clean' dataset which you understand in detail, so that you can spend the exercise time working effectively with these data rather than wasting time on cleaning/re-structuring/etc.

# Multivariate Visualization Project

| |
|---|
| Name: |
| Project Title: |
| Background: (provide a brief description of the background to your study) |
| Hypothesis: (what is the most important hypothesis you want to investigate) |
| Expectations: (what are your expectations in terms of results ... based on literature or previous work) |
| Levels of Organization: (list the observational unit on which the variables are recorded, as well as any additional levels in the hierarchy of your data) |
| Key Dependent Variables: (describe the most important dependent (outcome) variables in your study; make sure to indicate the variable type (e.g. quantitative continuous or categorical ordinal)) |
| Key Predictor(s): (describe the most important independent (predictor) variables in your study; also here indicate the variable type) |

Copies of these sheets may be distributed to all course participants.