University of Maryland
CENTER FOR ENVIRONMENTAL SCIENCE

Multivariate Analysis in Environmental Science

Number of credits

MEES ***
Semester?

Course Objectives / Overview

A previous course in statistics or some familiarity with probability in statistics, as well as some experience using R is recommended though not requried. Students without a statistical background should consult the instructor prior to regestering for the course.

This course will teach techniques in exploring and anylizing highly multidimensional enviornmental datasets.

A common challenge in environmental science are data sets in which there hundreds of variables, and in which the number of variables exceeds the number of samples. This challenge is particularly common when employing molecular -omics techniques, as well as with other approaches that assemble and compare many parameters. This course teaches analystical and statistical techniques that scientists can use generate and test hypotheses in such highly multidimensional datasets. We will explore statistical tests for dissimilarity between multivariate data (such as anosim and permanova), visualize and quantify relationships between multivariate datasets and key environmental parameters using ordination approaches, and explore patterns in associations between many pairs of variables using association network analysis. We explore challenges in dealing with spurious associations that appear in all such datasets, as well as challenges of compositionality and uneven read depth that often appear in -omics data. The course will use R-programming with version control in github, and will improve students practice with producing reproducable and well commented code, and ability to collaborate on coding challenges.

Expected Course Learning Outcomes

1. Students will learn approaches to explore, generate hypotheses about, and test hypotheses regarding multivariate environmental datasets.

2. Students will learn analytical techniques including the following:

2a. Statistical techniques for testing for differences between groups including Analysis of Similarity (ANOSIM), Permutational Multivariate Analysis of Variance using Distance Matrices (PERMANOVA).

2b. Ordination techniques including metric and non-metric multidimentional scailing (MDS, nMDS), and canonical correspondence anslysis (CCA) to identify patterns and their relationships to key variables.

INSTRUCTOR DETAILS:
Jacob Cram
jcram@umces.edu
(410)221-8481

CLASS MEETING DETAILS:
Dates:
Times:
Originating Site:
IVN bridge number:
(*******)
Phone call in number:
(***)
Room phone number:
(*****)

CURRICULUM FULLFILMENT:
MEES *** fulfills a **PD** MEES requirement.

Prerequisites
A previous course in statistics or some familiarity with probability in statistics, as well as some experience using R is recommended though not requried. Students without a statistical background should consult the instructor prior to regestering for the course.

Teaching Assistant
TBD or N/A

2c. Network analysis approaches including spearman and graphical lasso approches, and an understaning of measurements of network properties (such as density, and clustering coefficients).

2. Students will develop coding and collaboration skills by working in pairs on in-class assignments using the R programming language and Github.

3. Students will conduct a final data analysis submitting well commented code, and and share results with the class through an oral presentation.

Course Assessment / Grading

In class coding assignments 40%

Class participation 10%

Research project proposal (1-2 paragraphs, 2 page maximum + references) 10%

Well documented Rstudio notebook describing research findings and sharing code 30%

Final oral presentation on a term paper (10 min) 10%

In class assignments (40%)

Most classes will have an in class exercise. Students will be expected to work in pairs during the time allotted to complete at least some of the exercise. Students who miss class will be expected to spend a comperable amount of time working on the assignment. Points will be awarded for engaging with, and in some cases completing the material. Grades will not be assigned for accuracy.

Participation (10%)

Students will be encouraged to actively participate in class discussion which includes raising and answering questions, and working with their piers on in class assignemnts.

Research project proposal (10%)

Students will conduct a mini data-analysis-based research project. A large database, with many variables such as the Tara Ocean Dataset, will be explored to address different ecological questions. Students will have access to the database, some example code that analyzes that database, and hands-on instruction in learning about what the code does and modifying the code does. By the mid-term of class, students will be expected to come up with a short research project proposal (<two pages, one page is acceptable) which includes the rationale, questions, hypothesis, and a brief plan.

The proposal (10 points) will be graded on the following breakdown.

Rationale, Experimental Questions and Hypothesis \5

Quantity of research plan /5

<u>Project report (30%)</u>
Students will continue to work on the research project following the research proposal they submitted. Instructors will engage with students and provide guidance to students. Some class time, will be spent working on this analysis. Projects may be done in groups. The format of the report will be an Rstudio notebook. The notebook will contain an abstract, introduction, and methods overview. Results and figures will be produced by interweaving code and commentary. A discussion should also be included. The report is due on the last week of class.

The report (30 points) will be graded based on the following breakdown.

Abstract /2
Does the abstract concisely summarize the key parts of the report?
Introduction /3
Does the introduction convey the significance of the research problem?

Methods Overview /2
Is a summary of the methods and a rationalle for their use provided.

Code /4
Is the code necessary to produce the figures provided and commented such that the analytical approach is clear?

Versioning /2
Is the final project available on github, with a proper commit history?

Results /4
Are the results clearly described in text?

Figure /4
Is at least one figure produced? Does it clearly convey research findings? Is a caption provided?

Discussion /4
Is the discussion well integrated with the results and supply a compelling narrative that ties back into the introduction?

Overall clarity, writing quality, commenting, and grammar /5

Tentative Weekly Course Schedule

Week	Topic	Dae	Activities

1	Course Meeting (Re)Introduction to R		Discussion Interactive Excercise
2	Introduction to Github/Github Classroom		Brief Lecture Interactive Excercise
3	Introduction to the San Pedro Microbial Observatory Time-series data-set.		Lecture
4	Data wrangling and data visualization		Interactive Exercise
5	Multidimensional distance & distance matrices		Brief Lecture Interactive Exercise
6	Non metric multidimensional scaling & Analysis of similarities + PERMANOVA Testing		Brief Lecture Interactive Exercise
7	Ordination approaches		Brief Lecture Interactive Exercise
8	False discovery rate adjustment		Brief Lecture Interactive Exercise
9	Lasso Regression		Brief Lecture Interactive Exercise
10	General additive models		Brief Lecture Interactive Exercise
11	Generating Association Networks		Brief Lecture Interactive Exercise
12	Analyzing association networks		Brief Lecture Interactive Exercise
13	Work on final projects		In-class workshop
14	Final Presentations		Final projects due

Required textbooks, reading and/or software or computer needs

The following free to use and open source software and accounts will be required:

Rstudio, R, Git, a Github account, and Slack are all required. Sign up for each is free.
We will use several free resources including
Riffomonas Minimal R by Schloss,
An Introduction to Statistical Learning 2nd Edition by James, Witten Hastie, Tibshirani.
Materials by Gavin Simpson https://fromthebottomoftheheap.net/
Materials produced by Cram and Weissman on Bioinformatics Virtual Coordination Network.

Course Communication

Classwork will be available through github classroom.
Communication will be over a course slack channel.
Classes will all be held over zoom.

Resources

Campus Policies

The University of Maryland Center for Environmental Science has drafted and approved of various academic and research-related policies by which all students and faculty must abide.

Please see especially Policy III-1.00: Policy on Faculty, Student and Institutional Rights and Responsibilities for Academic Integrity.

Course-Specific Policies and Expectations