Data Analysis & Statistics for Earth Scientists
Nanjing University, Spring 2026

Lab 3 Assignment

1. The US government has exploded a large number of nuclear explosive devices underground at its Oasis Valley Test Site (Nevada), many at or below the water table. Farnham et al. (2000) provide a table of analyses for 19 trace elements collected from 22 wells at this location. These data are provided in the Oasis Valley datafiles, along the locations of the sampled wells and springs.

    a. Determine whether similarities exist among the trace element compositions of these water sources using principal components analysis (PCA).

        i. Decide whether the data need to be transformed prior to analysis. Explain the logic behind your decision.

        ii. Decide what the appropriate basis matrix for your analysis is, given the decision you made above. Explain your logic.

        iii. Decide how many principal components (eigenvectors) are necessary to represent the major features of these data. Explain your logic.

        iv. Do natural trace-element groupings of water bodies exist in these data? If so, define these groups and characterize both their similarities and differences in terms of the original variables. Show your plots and the equations of the eigenvector axes.

        v. Which of these groupings (if any) reflect the geographic location of the sampled locations? Summarize the data you used to answer this question on a geographic location map. Note: some localities may be represented by multiple samples.

2. Different sedimentary depositional environments may be able to be "typed" by their characteristic grain size distribution. Krumbein and Aberdeen (1937) tabulated grain-size data from a set of 50 samples collected from different near-shore depositional environments in Barataria Bay, Louisiana (see the Barataria Bay dataset). The problem here is to explore which representation of these data is most informative when it comes to typing these environments using grain-size data, which consist of the percentage sand in each of seven grain-size categories. Compare and contrast the following representations of these data:

    - mean category grain size,
    - pooled sample PC-1 score,
    - percent sand in each sand grain-size category.

        i. Decide whether the data need to be transformed prior to analysis. Explain the logic behind your decision.

        ii. Decide what the appropriate basis matrix for your PCA analysis is, given the decision you made above. Explain your logic.

        iii. Decide how many principal components (eigenvectors) are necessary to represent the major features of these data. Explain your logic.

        iv. Do natural grain-size groupings of water bodies exist in these data? Show your plots and the equations of the retained eigenvector axes.

        v. The ratio of the sum of squares among groups (SS_A) to the total sum of squares (SS_T) can be used as a measure of how tightly a set of data is clustered and separated from other groups. Determine which representation is most effective for characterizing the different depositional environments. Show your calculations for each analysis along with your interpretation.

3. A collection of ten trilobite taxa has been made from localities in the Great Basin (Utah) for the purpose of determining whether these species can be identified using simple length measurements of the cephalon and pygidium. There data are provided in the Great Basin Trilobites dataset

Variable	Description
x1	Border length
x2	Brim length
x3	Palpebral lobe length
x4	Glabella width
x5	Fixed cheek length
x6	Genal spine length/free cheek length
x7	Pygidial axis width
x8	Pleural axis width
x9	Pygidial axis length
x10	Pygidium length

In order to standardize for size differences among individuals all cephalon lengths were divided by the length of the glabella and all pygidum lengths were divided by the width of the pygidium. Use multivariate dimensionality reduction methods to summarize these data and determine whether representation of the associations between variables or distances between individuals performs best in recovering the standard taxonomy.

   i. Decide whether the data need to be transformed prior to analysis. Explain the logic behind your decision.

   ii. Decide what the appropriate basis matrices are for your analysis. Explain your logic.

   iii. Decide how many eigenvectors are necessary to represent the major features of these data. Explain your logic.

   iv. Identify the variables and species most closely aligned with each of your eigenvector axes.

   v. Show the comparative ordination plots that support your (written) interpretation.

4. The Swedish government was interested in evaluating the potential for exploiting heavy mineral deposits in the heavily forested northern part of that country. As airborne magnetometer surveys proved to be of limited value a program of ground sampling from stream sediments was undertaken. The Sweden Sediments HM dataset contains measurements (in ppm) of 13 heavy-metal concentrations for 46 locality samples divided into three groups: 1 - samples from known productive regions, 2 - samples from known non-productive regions and 3 - samples from regions whose productivity was unknown. Perform a multivariate discriminant analysis on the data from groups 1 and 2.

   i. Decide whether the data need to be transformed prior to analysis. Explain the logic behind your decision.

   ii. Identify the variables most closely aligned with your discriminant function.

   iii. Show the comparative ordination plots that support your (written) interpretation.

   iv. Test your results for statistical significance.

   v. Use your result to predict the potential for mining operations to be successful in the regions of unknown productivity.

5. The carbonate sediments of the Upper Jurassic Smackover Fm. on the eastern Gulf Coast of the USA are extensively dolomitized with the source of dolomitization varying regionally. Among the known mechanisms are seawater seepage (SS), reflux of evaporite brines (EB), shallow burial alteration by freshwater seepage (SF), shallow burial alteration by mixed connate and freshwater (SM) and deep burial (secondary) alteration (DB). The Smackover Dolomites dataset contains a variety of isotopic, elemental and physical variables collected from 97 samples for which the dominant mechanism of dolomitization has been established based on independent criteria. However, since these variables can be assessed from sediment cores we would like to know the extent to which these variables reflect different dolomitization environments. Use multivariate dimensionality reduction analysis to assess the potential use of these measurements'

    i. Decide whether these data need to be transformed prior to analysis. Explain the logic behind your decision.

    ii. Select an appropriate data-analysis method for use in assessing these data. Justify your decision.

    iii. Decide how many dimensions are necessary to represent the major features of these data. Explain your logic.

    iv. Identify the variables most indicative of each dolomitization mechanism.

    v. Show the ordination plot(s) that support your (written) interpretation.

    vi. If appropriate, test the reliability and/or statistical significance of your results.