**Project Description**

* These scripts were developed to facilitate hypothesis tests described in Chapter 7.3 of Modeling and Analysis of Compositional Data by Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015). All equations listed here are taken from the chapter text.
* My motivation was to facilitate application of these tests to compositional datasets I have generated using the electron microprobe. In principle, they are applicable to any population of compositional data.
* To use the equations, you will need to define the following variables:
  + nx = population size (subscript x identifies which population)
  + μx = population mean expressed in isometric log-ratio coordinates
  + Σx = population covariance matrix

These may be calculated using any standard arithmetic method or statistical package (in the case of population covariance matrix).

* Three hypotheses are tested against a so-called “general” hypothesis as described on page 137 of the text. All four hypotheses are concerned with comparing the means and covariance matrices of two hypothetically linked populations of compositional data. They are written as, quoting the authors:

1. There is no distributional difference between both populations
2. The covariance structure is the same, but the centers are different
3. The centers are the same, but the covariance structures are different
4. The populations differ in their centers and covariance structure

Hypothesis 4 is termed the “general” hypothesis, and hypotheses 1:3 are compared against it as null hypotheses.

**Written Explanation of the Code**

Three files are included in this repository:

* *common\_mean.R* calculates the common mean as defined in the text following this equation:

using the built in *solve()* function to invert the covariance matrices.

To call this function, the following syntax is used

μc = common.mean(n1,n2,Σ1,Σ2,μ1,μ2)

* *test\_covariance.R* iteratively calculates a maximum likelihood estimate for the third hypothesis test described below. It first calculates a common mean using the initial covariance matrices and then recalculates covariance matrices using the following equation:

where h simply designates the next iteration of the covariance structure and T the transpose operation(*t()* in R). This step is repeated until the following ratio is achieved:

where |Σ| is the determinant of a given covariance matrix (*det()* in R). For simplicity and in the interest of efficiency, ratio values are rounded to the 6th decimal place. Once this ratio equals one, the function ceases (a break is executed when this condition is met) and the function returns the variances of each group with respect to the common mean. The number of iterations needed to reach convergence is printed in the console.

To call this function, the following syntax is used

comm\_vars = test.cov(n1,n2,Σ1,Σ2,μ1,μ2) = (Σ1h, Σ2h)

* *hypothesis\_test.R* does most of the work. It follows the precise procedure for the maximum likelihood ratio tests covered on pages 136-142 of the aforementioned text. The function may be broken into four sections. The first section calculates the pooled and combined sample statistics, which are used in testing hypotheses 1 and 2 against the general hypothesis (sections 2 and 3, respectively). The final section utilizes the functions described on the previous page to estimate a test statistic. Each hypothesis test includes calculation of a unique set of test parameters, called “restrictions” by the authors, which are used as input to calculate the Chi-Square critical value. Each restriction value is printed during computation.
  + The first section calculates the pooled covariance matrix (Σp) and the combined sample estimates (μco and Σco):
  + The second section performs the first hypothesis test:

The restrictions are calculated as such:

where D is the calculated degrees of freedom for the dataset (the number of elements in the mean vector minus one). The test statistic is approximated as:

After calculating the value, the returned p-value is compared against χ2(c1), and if Q1 vs 4 > χ2(c1), then we fail to reject the null hypothesis and the function prints “H1: FAILURE” in the console. Else, “H1: PASS” is printed.

* + The third section performs the second hypothesis test:

Restrictions are calculated this time as:

and the test statistic for is calculated as:

Again, this p-value is compared to critical value calculated under the Chi-Square distribution, and if it is greater than that critical value, then we fail to reject the hypothesis, and “H2: FAILURE” is printed. Else, “H2: PASS” is printed.

* + Finally, the fourth section performs the third hypothesis test:

Restrictions are:

The common mean (μc) and variances of each group with respect to the common mean (Σ1h and Σ2h) are then used to calculate the test statistic:

The test is performed, and either “H3: FAILURE” or “H3: PASS” is printed in the console. Also printed is the number of iterations needed to reach convergence during the *test\_covariance* computation.

**Cautioning Remarks and Tips**

* These hypothesis tests may be performed for any sort of compositional data. They do not consider any effects of thresholding, screening, or outlier presence. We caution against using this code as a “black box” and prescribe no thorough methodology for preparing the data for hypothesis testing.
* I prescribe no additional interpretation using these hypothesis tests. The maximum likelihood ratio test simply indicates whether a given null hypothesis may be rejected or not (fail to reject).
* Manipulating the significance used to calculate Chi-Square critical values (i.e., the quantile) may be done by pulling up the *hypothesis\_test.R* script and looking for lines 20, 30, and 44. The first value inside the parentheses in these lines is the given level of significance. Changing this value will change the output critical value. You may also change which tail of the distribution is sampled in these same lines.
* The full citation for Pawlowsky-Glahn et al. (2015) is as follows:

Pawlowsky-Glahn, V., et al. (2015). Modeling and analysis of compositional data, John Wiley & Sons.

I recommend trying to find this text and reading at least the section on hypothesis testing about two populations. It explains the equations reproduced here and provides example calculations.