

Stat 666
Mini-Project #3
Due TBA

Mini-Project #3 (to be done in groups of two...choose your own partner, but it must be someone different than your partners from Mini-Projects #1 and #2):

In recent years, there has been increasing interest in quantitative methods for identifying authorship of written texts. Peng and Hengartner (*The American Statistician*, **56**, 175-185) noted, “It is often recognized that authors have inherent literary styles which serve as ‘fingerprints’ for their written works. Thus, in principle, one should be able to determine the authorship of unsigned manuscripts by carefully analyzing the style of the text.” In this project, we are interested in grouping 1000 selected texts into coherent groups (or clusters) based on only 18 quantitatively-measured aspects of writing style. The file `collins.txt` contains 18 stylistic variables for each of 1000 texts. The first column is simply a text number (from 1 to 1000), and the next 18 columns (FirstPerson through ShiftingEvents) contain measurements of different facets of literary style. You can treat the 1000 texts as if it is a random sample from the total population of texts about which we are interested. More information about the variables is available at tofu.byu.edu/docs/files/stat666/assignments/Summary18RhetoricalCategories.pdf and even more at <http://tofu.byu.edu/docs/files/stat666/assignments/DissertationOn18RhetoricalCategories.pdf>.

In this project, we wish to identify the ways in which texts differ from one other in terms of literary style. We wish to begin by completely ignoring the Genre of writing noted in the dataset. You should not use the Genre variable until Step 3. You will address this research question using an array of multivariate methods. For each question or task being considered, carefully justify the methods you choose to use.

Step 1. Use the 18-dimension measure of writing style to create several clusters of texts. The number of clusters you choose should be between 3 and 7 clusters. Create a “best 3-cluster solution,” a “best 4-cluster solution,” etc., up to 7 clusters.

Step 2. In order to choose which set of clusters you will use for the remainder of the analysis, carry out a follow-up analysis of the clusterings created in Step 1. Are the cluster means significantly different from each other when using $k = 3$? What about $k = 4$? $k = 5$? $k = 6$? $k = 7$? If we were to predict the group membership for a new written text, which set of clusters would yield lowest misclassification rates? (Note that MANOVA alone will not be sufficient to help you choose a reasonable set of clusters. Additional methods will also be necessary.)

Step 3. After selecting a set of clusters in Step 2, describe their differences. Are cluster means collinear or diffuse in the s -space containing these means? In what ways are clusters different from each other? That is, can you characterize each cluster in some way—can you give a name or interpretation to each cluster? (At this point you may use the Genre labels to assist you in describing your natural clusters if that is helpful.) Note that the genres in the Genre column of the spreadsheet are as follows:

1. Press: Reporting
2. Press: Editorial
3. Press: Reviews (theatre, books, music, dance)
4. Religion
5. Skills and Hobbies
6. Popular Lore

7. Biography, Memoirs, Belles Lettres, etc.
8. Official Communications (government documents, reports, catalogs, etc.)
9. Learned (scholarly journals)
10. General Fiction
11. Mystery and Detective Fiction
12. Science Fiction
13. Adventure and Western Fiction
14. Romance and Love Story
15. Humor

Using a 2-way table or some other device, describe how your natural clusters align with the genres for these texts. Compare your best 5-cluster solution with the super-genre variable defined by

1. Press (genre = 1, 2, or 3)
2. Non-press Nonfiction (genre = 4, 5, or 6)
3. Biography (genre = 7)
4. Scholarship and Official Documents (genre = 8 or 9)
5. Fiction (genre = 10, 11, 12, 13, 14, or 15)

We expect the 5 cluster groups to be more statistically distinct than the 5 super-genres, but does the improvement in group separation warrant the use of less intuitive group definitions? I.e., are we better off if we compare different writing style groups simply using the super-genres?

Unlike your other clients this semester, your client Dr. Y has an M.S. degree in Statistics, but needs additional assistance because she was unable to take a course in Multivariate Statistical Methods. With the exception of her lack of experience with Multivariate Statistical Methods, you may assume that Dr. Y is completely competent in the fundamentals of statistics. Consequently, she will be able to discern whether or not you know what you are talking about! Despite her relative competence with statistics, you will still want to make sure that any new concepts are defined in a manner that is both technically sound (using formulas) and intuitively explanatory.

Using the standard M.S. project format, you should turn in a write-up of no longer than **5 pages** (11 pt font, 1 inch margins, single spaced), including all tables and graphs. Your reasonably-well-commented computer code (R, SAS, or C) for your analyses should be attached to the back of your report. (The code in the Appendix will not be counted against your page limit.) You are not allowed to discuss the results of your analysis with other members of the class (except for your project partner). You may, however, discuss the project with me.

Because of the amount of work required in this project, this project will get 1.5 times the weight of the first project.