Biostats VHM 802 Course
Winter 2021/2024, Atlantic Veterinary College, PEI
Henrik Stryhn

Additional Exercises for Multivariate Methods

Exercise MV.1
Univariate and graphical analysis for Egyptian skulls data
Example 1.2 of the Manly textbook describes measurements on 30 skulls from each of five periods. Among the questions of interest being discussed for analysis of the data are:

1) How are the measurements related (not physically, but based on the data)?

2) Are there important differences (between periods) in the sample means or standard deviations for the variables, and if so do such differences reflect gradual changes with time?

The aim of this exercise is to use univariate analyses (i.e., for one measurement at a time) and graphical displays to explore these questions. It is suggested to start with graphical display(s) that can assist in determing both the distributions and relationships between the measurements, thereby helping to clarify what assumptions may be reasonable for subsequent analysis. Comparisons between periods (based on suitable statistical analyses) should be displayed graphically, in order to allow assessment of whether any differences/changes are gradual over time. Time may be represented either as periods $1-5$ or the actual years when the measurements were taken (this information is however not included in the dataset).

Additionally, explore the use of Chernoff faces to display mean differences between the periods. Try also to construct faces for individual samples, thereby potentially allowing a visual assessment of how consistent differences between periods are among the individual samples.

Exercise MV.2
Univariate and graphical analysis for European employment data
Example 1.5 of the Manly textbook describes socioeconomic data for 26 European countries obtained in the early 1990s, consisting of percentages of the population employed in nine (mutually exclusive) different sectors. Some questions of potential interest for analysis of the data are:

1) Can groups of countries with similar employment patterns be identified, and if so, do such patterns align with a division of the countries into the four political groups provided?

2) Do some countries show stark differences to other countries, both within and across the political groups?

3) What patterns exist between the employment percentages? (primarily when seen across the political groups because of the limited replication within some of the groups)

The aim of this exercise is to use summary statistics and graphical displays to explore these questions. It is suggested to start with graphical display(s) that can assist in determing both the distributions and relationships between the measurements, thereby helping to clarify what assumptions may be reasonable for subsequent analysis (in particular, whether some of the variables show serious deviations from a normal distribution). It may be worthwhile to establish whether the percentages add up to 100%. Additionally, explore the use of Chernoff faces to display differences between the countries, both in determining general patterns and in identifying countries that are outlying on one or several parameters.

Exercise MV.3

Graphical exploration of Iris data

A famous dataset for multivariate methods, in particular classification and linear discriminant analysis, consists of measurements for 50 samples from each of three species of Iris (*Iris setosa*, *Iris virginica* and *Iris versicolor*). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. The data were collected by the American botanist Edgar Anderson to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula, "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus" (Anderson (1935), The irises of the Gaspé Peninsula, *Bulletin of the American Iris Society* **59**, $2-5$). The data owe their fame to analyses by the British statistician, eugenicist, and biologist Ronald Fisher.

Use graphical tools to explore how the four measurements provide information to separate the three Iris species from each other. In particular, the data are well suited for 3-d plots (using three of the four variables). Based on your exploration, can you come up with an approximate rule to identify the species from a set of four measurements?

Exercise MV.4

Cluster analyses for Prehistoric dogs data

The Manly textbook shows results of a hierarchical cluster analysis for these data (Chapter 9). Use your statistical software of choice to try to reproduce these results, and interpret the resulting dendrogram carefully. Next, try some other linkages than the single/nearest neighbour linkage used in the textbook. Do you notice any differences in the configurations obtained?

Continue the analysis by exploring the K-Means algorithm for these data, with K set at 2 and 3. Do the resulting partitions agree with the clusters obtained from the hierarchical cluster analysis? Make sure to assess the stability of your results from different starting values and/or versions of the algorithm.

Exercise MV.5

Distance-based methods for European employment data

The Manly textbook uses the European employment data as the primary example for cluster analysis. Try again to reproduce these results in your statistical software of choice, and interpret the resulting dendrogram. You may find the four political groups among the 30 countries helpful for interpretation. If interested, explore some other linkages than the single/nearest neighbour linkage used in the textbook.

The analysis in the textbook uses the standardized variables. Discuss the pros and cons of this approach, and compare some of the results with corresponding results for untransformed data. Another alternative analysis is for a subset of the data with certain countries excluded; it was previously discussed (Exercise 2) that some of the values were strongly outlying and/or possibly errors. Without aiming to cover all options, try to explore whether such issues affect the results substantially.

Finally, explore also how multidimensional scaling works for these data (effectively the Chapter 11 exercise of the Manly textbook). Do graphical representations of distances between the countries show similar or different patterns than those obtained from the clustering algorithms?

Exercise MV.6

Exploration of distance-based methods for a large(r) dataset

Two options are offered for this exercise: the botanical Steneryd dataset described in Exercise 1 for Chapter 9 in the Manly textbook, and the NCIarray data already discussed briefly in the lecture (but treated in more detail (as the NCI60 data) in Chapter 10 of the SL textbook). For the Steneryd data, it may be of interest to cluster both species and plots, as suggested in the Manly exercise. Two versions of the data are provided: `steneryd.csv` is formatted with plots as observations and includes additional environmental variables for the plots, whereas `stenerydxp.csv` is formatted with species as the observations (and therefore cannot include the environmental variables). It could be of interest to explore whether clusters obtained for the plots are related to the environmental variables.

For the NCIarray data, classical multidimensional scaling with the first three dimensions might be of interest because they also correspond to the first three components extracted from a principal components analysis (shown in the SL textbook example).

Exercise MV.7

Distance-based methods to explore shape

The Manly textbook (3rd/4th edition) describes a dataset with six measurements on each of 25 pottery goblets excavated from prehistoric sites in Thailand (data provided by Professor C.F.W. Higham of the University of Otago, New Zealand). Figure 6.3 (accessible at the Moodle site of the course) shows the typical goblet shape and how the measurements (simply labelled as $x_1, \ldots, x_6$) were defined. The interest is in describing similarities and differences between the 25 goblets.

Use multidimensional scaling and cluster analysis methods to develop distance-based graphical representations that can be used to answer the questions of interest. Interpret the results carefully.

The Manly textbook notes that any differences between goblets may be due to size or shape (or both). It is of interest to carry out an additional analysis focused entirely on shape similarities/differences. Two proposed ways to normalize the measurements to eliminate strong size dependencies are to divide all measurements by either the total height of the goblet or the total sum of all measurements. Carry out a similar analysis as above for at least one of these rescaled sets of measurements. Draw conclusions, and comment on whether or how the impact of size differences can be seen in the comparison of the results from the two versions of your analysis.

Exercise MV.8

Dimension-reduction methods for European employment data

The Manly textbook uses the European employment data as the primary example for both principal components analysis and factor analysis. Try to reproduce the results for both analyses in your statistical software of choice, and interpret the resulting graphical displays, in particular the score and loading plots. You may find the four political groups among the 30 countries helpful for interpretation. If interested, explore also some other rotations than the varimax rotation used in the textbook. Also here it may be of interest to explore the sensitivity of the results to extreme countries and variables; carry out such additional analyses if interested.

Exercise MV.9

Dimension-reduction methods for data with different variable types

For this exercise, we will use the `beef_ultra` dataset supplied with the VER2 textbook. (*Hint:* Using the Stata data format will save you some formatting of variables.) It contains measurements on 487

cattle. The final carcass grade will be used only for display purposes, and our objective here will be to describe the relationship between the other variables.

Determine the variable type for each of the variables, and (briefly) compute suitable descriptive statistics. Discuss and decide which variables to include in a principal components analysis; you should end up with 8 variables. Carry out the analysis and try to interpret the components. Explore also whether any of the first components seem to be associated with carcass grade; it is suggested to use different plotting symbols for the three grades in score plots.

Repeat the analysis for "-choric" correlations, and compare the results with those previously obtained. Make sure to cover both the eigenvalues, the eigenvectors and the resulting scores. You may also want to compare the Pearson and -choric correlation matrices. In summary, which analysis (Pearson or -choric correlations) do you find preferable?

For your preferred choice of correlation matrix, continue your exploration by a factor analysis. It is suggested to start from the principal components, and use a varimax rotation as the first example of a rotation. Does the rotation improve the interpretability of the factors? From here onward, there is room for exploring the data in different ways (different numbers of factors retained, different estimation methods, different rotations), depending on your enthusiasm.

Exercise MV.10

Exploration of properties for the PCA solution

The purpose of this exercise is to verify that the PCA solution provided by statistical software indeed satisfies all the conditions attached to it. In a sense it would be very surprising if it did not, so it is not so much about checking the software calculations but about to understand what the different conditions actually mean. It is suggested to work with the PCA results for the sparrow data presented in the lecture, but other data can be used as well.

Verify each of the following conditions/results (if the condition involves repeated checks, the first one should do):

1) The score vectors for each of the components are computed by linear combinations of the original variables with the stated coefficients/loadings.

2) The score vectors have variances equal to the eigenvalues.

3) The score vectors are uncorrelated.

4) The sum of the eigenvalues equals the trace of the matrix analyzed (either a covariance or a correlation matrix).

5) The sum of the squared (eigenvector) coefficients equals 1.

6) The eigenvectors are orthogonal.

7) The eigenvector and its corresponding eigenvalue satisfy the defining equation for eigenvectors/values.

Exercise MV.11

Dimension-reduction methods for human body measurements

Additional exercise 2.9 is about developing a multiple regression model of bodyfat on age, weight and 11 measurements of body dimensions for a sample of 252 men. In this exercise, we will explore multivariate analyses for these highly correlated values on each subject. The dataset includes also a body density variable, but it is functionally related to bodyfat and therefore one should not include both variables in a multivariate analysis.

Use principal components and factor analyses (i.e., include at least one rotation of the PCA solution) to answer each of the following specific questions:

1) Analyze first the physical body measurements only (i.e., 11 variables). How many components are needed to represent the (majority of the) information? Use graphical display to interpret those components (briefly).

2) In your graphical displays, you should identify two extreme observations. Explore those observations in more detail, and make decisions about whether they should be included or excluded from further analysis. (*Hint:* To make such decisions, you may need to compare the results with and without the observations in question.)

3) Include weight among the variables for your multivariate analysis. Describe how the results and the interpretations of the extracted components change.

4) Same question as 3) for inclusion of either the age variable or the bodyfat variable, of your own choice (one of these should suffice).

Exercise MV.12

Multivariate ANOVA/regression for Egyptian skulls data

The Manly textbook uses the Egyptian skulls data, with its one-way grouping by periods, as the second and most comprehensive example for multivariate ANOVA (MANOVA). Try to reproduce the results in your statistical software of choice, with exception of the multivariate test for equal variances (if interested in variance tests, Box's M-test is easier accessible). Make sure to extract not only the test results, but also the parameter estimates, including the estimated error variance-covariance matrix (represented by the variance, or standard deviation, for each variable and the correlations between variables). If you use Stata, try also the multivariate equivalent of the `regress` command, interpret the listed coefficients and explore the available postestimation commands (e.g., the `margins` command). For the postestimation features you explore, make sure to compare the output with corresponding output following a univariate analysis (whenever relevant). Do you discover any new statistics or new results compared to univariate analysis?

As an alternative to the use of Mantel's test to determine significance for the strength of association between a distance matrix obtained from the multivariate data and the time distances for the periods, use univariate and multivariate methods to model an explicit dependence of the (means of) measurements on time as a quantitative predictor. Make sure to assess whether the relations appear to be linear. Interpret (some of) the regression coefficients, and draw conclusions.

Exercise MV.13

Multivariate ANOVA/regression for mandible measurements of canine species

The Manly text provides an expanded version of the data for the prehistoric dogs from Thailand data for its Chapter 4 exercise. For 5 of the 7 species, a total of nine lower jaw measurements are available for between 10 and 20 individuals. Additionally, for 4 of the 5 species, the sex is available. The full list of variables is:

species	:	one of the 5 species
x_1	=	length of mandible (mm)
x_2	=	breadth of mandible below first molar (mm)
x_3	=	breadth of articular condyle (mm)
x_4	=	height of mandible below first molar (mm)
x_5	=	length of first molar (mm)
x_6	=	breadth of first molar (mm)
x_7	=	length of first to third molar, inclusive (first to second for cuon) (mm)
x_8	=	length from first to fourth premolar, inclusive (mm)
x_9	=	breadth of lower canine (mm)
sex	:	coded as: 0 (unknown), 1 (male), 2 (female)

Consider the following questions from the Manly text (here reworded):

1) Without accounting for sex, compare the 5 species in terms of the mean values and variations for the 9 variables. Include a descriptive analysis, parameter estimates and statistical tests. Specifically, compare the prehistoric Thai dog with each of the other species singly. What conclusion do you draw with regard to the similarity between prehistoric Thai dogs and the other species?

2) Now include the sex in the comparison between species. How does this affect your ability to compare the prehistoric Thai dog with the other species? Discuss the options you might have for dealing with this problem. Make sure to also assess whether sex differences differ between species.

3) Use a suitable graphical method to compare the distributions of the nine variables between the prehistoric and modern Thai dogs.

Exercise MV.14

Simulation-based ANOVA

The purpose of this exercise is to explore simulation-based ANOVA methods (`ANOSIM` and `PERMANOVA`) for multivariate data. Two datasets you could try this for, are the expanded prehistoric dogs data from the previous exercise and the Steneryd data (for plots). The first dataset has two factors, and the second one has four quantitative predictors (the environmental variables); note that `ANOSIM` does not allow quantitative predictors. Explore at least two different distance measures, of your own choice. Include also for comparison a corresponding analysis by MANOVA (most meaningfully compared with Euclidean distances). Draw conclusions, and discuss any issues you identify with use of the simulation-based methods.

Exercise MV.15
Linear discriminant analysis for European employment data
The Manly textbook uses the European employment data as its second example for discriminant function analysis (Example 8.2). The results shown for the first example, for the Egyptian skulls data, do not appear to be easily reproducible, not in Stata, nor with the R code supplied with the textbook. Therefore the second example is more interesting for our purpose here.

Try again to reproduce the results from the Manly text in your statistical software of choice (Minitab is not recommended for this exercise). You should be able to reproduce the canonical discriminant function coefficients, their correlations with the original variables (Table 8.5) and the plots of these functions (Figure 8.1). Interpret the loadings and scores for the 3 discriminant functions. Note that because the percentages sum to 100% (up to a very small approximation error), one of the variables needs to be omitted from analysis.

Manly notes the discrimination obtained between Western and Eastern European countries to be better than what is obtained by the first two components of a principal components analysis; do you agree with this (visual) assessment? Explain also why one would certainly expect to do better with LDA than with PCA. Finally, compute classification tables (or confusion matrices) in different settings: with uniform and data priors, and with/without leave-one-out cross-validation, and interpret the results.

Exercise MV.16
Comparison of classification algorithms for the Iris data
We used the (classical) Iris data to illustrate the ability of clustering methods to detect well-known clusters. Its "fame" derives however for its use in development and exploration of classification algorithms. Explore how well linear and quadratic discriminant analysis, logistic classification and the kth nearest algorithm (with different choices for k) perform for these data, when evaluated by leave-one-out cross-validation. Investigate the flowers that are difficult to classify correctly, and try to understand why the different algorithms fail on those samples.

If you are interested in trying out a more challenging data set, Section 4.6.6 of the *Introduction to Statistical Learning* text describes and briefly discusses some results for a dataset involving 85 predictors on 5822 individuals. The interest is in predicting whether people will purchase a caravan insurance (yes/no); only 6% of responses were positive. The data can be accessed through the book's website (link in lecture 7−L).

Exercise MV.17
SaTScan analyses of Vietnam data
The VER textbook includes a dataset on highly pathogenic avian influenza serotype H5N1 in North Vietnam constructed from surveillance records during 2004-2006. The 2296 records at the "commune" level indicate whether each commune had an outbreak of H5N1 during the study period. A commune is a third-level administrative unit (below municipality/province and cities/towns); Wikipedia gives the total number of communes at the end of 2008 as 9111, but the divisions into communes vary over time.

Try to reproduce the results shown in VER Example 26.8 from a SaTScan analysis based on a purely spatial Bernoulli model for these data. In the collection of datafiles included with the textbook, `ver2_data_ch25_26.zip`, you should use the Chapter 26 files named `viet_commune_centroid.*`. The collection already includes the ready-to-go files for SaTScan (`.cas`, `.ctl` and `.geo`), but you could also try to create those files yourself from the dBase file (`.dbf`). Once you get the software to

display the same clusters as in the textbook, interpret the results carefully. Experiment also with different settings of the software/algorithm, such as:

- elliptical shapes instead of circular,
- expanding the analysis to include both high risk and low risk areas,
- different options for secondary clusters.

Another possible analysis from the data provided can be done at the district level, based on the file `viet_district_centroid.csv`. The data have apparently been aggregated from commune to district level, for a total of 117 districts. The variables `no_commune` and `all_ai` contain the total number of communes and the number of communes with H5N1 cases, respectively. The datafile however also includes two additional variables: `duckdens` and `avg_height`. It is of interest to explore (and compare) cluster detection with and without accounting for these variables. One way to do this is to fit suitable Poisson regression models to the counts of H5N1 cases and use the predicted counts from those models as the expected counts in a purely spatial Poisson model for the SaTScan analysis. It may be helpful to briefly review SaTScan Tutorial #1, for such a model. Some of the districts have missing values for `duckdens` which will cause problems with the process, so it is suggested to drop those districts from the data (or drop the `duckdens` variable). If you manage to get analyses with and without accounting for the predictors to work, compare the results and their interpretations.