University of Michigan, Dept of Statistics

Stat 503, Instructor: Long Nguyen

Homework Assignment 1
(due Thursday, January 25, 2018)

1. **Bivariate Gaussian.** Please work out the example on page 30 of the lecture note `linalgreview-lec2.pdf`

2. **Conditional independence.** Given a zero-mean multivariate Gaussian vector $X = (X_1, X_2, X_3)$ for which the covariance matrix is

$$\Sigma = \begin{pmatrix} -0.1382271 & 0.98935598 & 0.04547539 \\ 0.1565976 & 0.06717134 & -0.98537567 \\ -0.9779420 & -0.12908426 & -0.16421565 \end{pmatrix}. \quad (1)$$

   Show that X_1 is independent of X_2 given X_3, i.e., the conditional densities of these variables satisfy identity $P(x_1, x_2|x_3) = P(x_1|x_3)P(x_2|x_3)$.

3. **Whitening and standardizing.** Turn in your figures and codes for the following.

   (a) Load the height/weight data from `heightWeightData.txt` in Canvas. The first column is the gender label (1 for male and 2 for female), the second coumn is height, the third weight. Extract the height/weight data corresponding to the males. Fit a 2-dim Gaussian to the male data, using the empirical mean and covariance. Plot your Gaussian distribution as an ellipse, superimposing on your scatter plot of data points, each which should be labeled by its index number (ranging from 1 to 210).

   (b) *Standardizing* the data means ensuring the empirical variance along each dimension is 1. This can be done by computing $\frac{x_{ij}-\overline{x}_j}{\sigma_j}$, where σ_j is the empirical std of dimension j, $\overline{x}_j$ the empirical mean. Standardize the data and replot.

   (c) *Whitening* or *sphereing* the data means ensuring its empirical covariance matrix is proportional to identity matrix, so the data is uncorrelated and of equal variance along each dimension. This can be done by computing $\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{x}$ for each data vector $\mathbf{x}$, where $\mathbf{U}$ are the eigenvectors and $\mathbf{\Lambda}$ the eigenvalues of the covariance matrix $\mathbf{X}^T\mathbf{X}$. Whiten the data and replot. Note that whitening rotates the data, so people (data points) move to counter-intuitive locations in the new coordinate systems.

4. **PCA warmup.** On the class website there is a data set *fa-data.txt* consisting of 500 data points in 7 dimensions. The data are believed to lie mostly near a 2-dim linear submanifold. (Please see *fa-gendata.m* for the Matlab code I used to generate the data set).

   (a) Produce a few visualizations of the data set by plotting only 3 dimensions selected randomly. Do you see that the data indeed lie near a 2-dim subspace?

   (b) Write your own code of PCA to identify the principal components and the projections of the data set on to the 2-dim principal subspace.

   (c) What is the proportion of total variance that is explained by PCA's two principal components?

5. **PCA.** In this exercise you are welcome to use an existing PCA package. The data set (`auto-mpg.data` on Canvas) concerns city-cycle fuel consumption in miles per gallon (mpg) and other attributes collected for 398 vehicle instances. The variables are: mpg, cylinders, displacement, horsepower, weight, accleration, model year, origin and car name. Perform exploratory data analysis on this dataset including PCA and write a report summarizing your data analysis. In particular:

(a) Describe the data and present some initial pictorial and numerical summaries, such as scatterplots, histograms, etc.

(b) Consider which variables should or should not be included in PCA on this dataset. Compare PCA on covariances and correlations (you may choose one of them to proceed with for the subsequent questions).

(c) Comment on the percentage of variance explained and number of principal components to retain. Include a scree plot.

(d) Comment on variable loadings and their potential interpretations.

(e) Make a plot of the data projected on the first two PCs. Comment on any interesting features, including potential outliers, if any. By a visual display of PC scores, can you detect a categorical (discrete) attribute which is the most distinguishable (i.e., data are most separated according to the attribute values)?

(f) Compute a bootstrap confidence interval for the percentage of variance explained by the first k PCs, where k is the number of PCs you recommend retaining for this dataset.

(g) Make a PCA biplot and comment on any interesting features.

Instructions The solution to problem 1 and 2 can be either typed up or written by hand. The solution to data analysis questions (Problems 3–5) may be written as a data analysis report. The report needs to be clear, concise, and to the point. There should be no graphs or tables that are not commented on in the text. Please include your R code (at least the main parts) in a *separate* appendix at the end of the report.