

Math 543: Midterm/Homework 5.**Due Wednesday March 12**

Analyze the global tuberculosis data set (GlobalTB.txt extracted from <http://www.who.int/globalatlas/>). These data contain TB rates per 100,000 people by country, for the years 1982 to 2005, though some countries have been excluded because of lack of data. As this problem is somewhat more challenging than the previous homework problems, you are encouraged to work in groups of 2 and 3 and submit a single report for your group. The analysis breaks down into two main tasks.

1. Characterize annual trend in TB rates, and in particular, identify countries that are suffering from large TB rates and also those that have experienced substantial *recent* increases in TB rate. In addition, characterize the pattern of variation in rates by region, or if the region variable does not help in this task, identify groups of countries that are alike with respect to TB rates.

Some points to be aware of.

- (a) After omitting the countries with missing data, the data file is 113 rows (countries) by 26 columns (country name, region, and 24 years of rates). It's useful to think of the data both as a 113×24 matrix \mathbf{X} (so that the variables are annual rate), and perhaps also as a 24×113 matrix $\mathbf{Y} = \mathbf{X}^T$, so that the variables are countries.
 - (b) Unlike the previous applications of principal components analysis that centered on correlation structure, and hence used centered data (so that every variable had mean 0), it is important *not* to center the data when analyzing these data, since the magnitude of the TB rate variable is of central interest. A centered data matrix is one in which every country has the same 24-year average rate (namely, 0). So, an eigenvector analysis of the primary sources of trend (in contrast to sources of variation) should be based on the un-centered moment matrix $\mathbf{M} = \mathbf{X}^T \mathbf{X}$.
 - (c) The analysis of country-to-country differences (or similarities) may be carried out via a multidimensional scaling approach, or one utilizing $\mathbf{Y} = \mathbf{X}^T$. Again, one should be careful in pursuing correlation analysis since correlation analysis is fundamentally an analysis that uses centered data.
2. Estimate the annual rate of change in TB rate for each country and project an estimate of the expected rate in 2010 A.D. Summarize your results in a table where the rows are countries and the columns are the 1. estimated rate of change, 2. a confidence interval for the true rate, 3. an estimate for 2010 A.D., and 4. a confidence interval for the true rate. Remark on any countries with particularly unusual or unpredictable rates.

Two strategies are suggested: 1. Fit models for each region while accounting for differences among countries within region, if necessary. However, regional models may not suffice for some countries, in which case, country-specific models may be necessary for countries. 2. Automatically fit simple models to those countries that are experiencing linear trend, and manually fit more complicated models to a few countries that are experiencing unusual or non-linear patterns of trend.