MTAT.03.227 Machine Learning, Session XII, maximum score 10p, deadline: 1st of May

I would like you to know what is PCA, how and when to use it, how to explain it to someone not familiar with the topic, how to plot and interpret the results of the analysis and how to write a nice report using the method. To sum up, I would like you to be able to get a full skillset needed to use PCA in practice.

APPLYING PRINCIPAL COMPONENT ANALYSIS TO SOCIOECONOMIC DATA OF THE BIGGEST AND RICHEST COUNTRIES IN THE WORLD

Understanding the richest and biggest countries in the world is important for many reasons. Often they are responsible for economic recessions; most of the political and economic global decisions are made by them; they spread their values to the rest of the world etc.

For these reasons one might want to study the change and development of these countries; learn from them or just to classify them. However, to accomplish this it might be easier if there were lot less variables to deal with.

Your goal is to apply PCA to the data of the richest and the biggest countries in the world and write a nice report explaining your analysis of the data.

DATA

From the 50 richest countries according to The World Bank (2005), countries with population less than 3 million have been excluded. In addition, Iceland and Singapore are excluded due to the lack of available data. Hence, the data comes from the following countries: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Ireland, Israel, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, UK and USA.

Three quarters of the variables are also based on the Richard Wilkinson and Kate Pickett's book (2010) "The Spirit Level. Why Equality is better for everyone". These variables are Income

inequality, Trust, Life expectancy, Infant mortality, Obesity, Maths and literacy scores, Teenage births, Homicides, Imprisonment (logarithmic scale), Child overweight, Drugs index, Calorie intake, Public health expenditure, Child wellbeing, Maths/education/science score, Child conflict, Foreign aid, Peace index, Maternity leave, Advertising, Social expenditure, Women's status, Patents, Lone parents. The variables are listed in the Appendix Table 1 with descriptions.

In addition, six more variables are chosen: GDP per capita, Press Freedom, Population Density, Electricity Consumption, Internet Users and Index of globalization. The variables are listed in Table 2.

Data is given in the file: PCAdata.xls

YOUR GOAL

Write a very short but nice report where the data is analyzed. Report has an introduction, figures with headings and they are cited in the text, conclusion and all other features which are needed. Follow these steps:

1. Read in all the data (0.5 p)
2. Plot all each of the variable separately. As it is not important, use as few pages as possible. (0.5 p)
3. PCA is linear. If some of the data is exponential then make a logarithmic transformation of the data (0.5 p)
4. If it makes sense, you can multiply some of the variables by -1 to make higher values always better (or worse) (0.5 p)
5. Normalize and standardize all the variables i.e. find the z-score (http://en.wikipedia.org/wiki/Standard_score) (0.5 p)
6. Make a PCA transformation of the data and explain with three sentences for a non-mathematician reader what that PCA is. (1 p)
7. Plot cumulative eigenvalues, explain what they mean and how much variance is explained by two first principal components. (0.5 p)
8. Show the loading of two first principal components i.e. what linear combination is used to produce PC1 and PC2. (0.5 p)

9. Write a short interpretation of the results (1 p)
10. Show the component scores of two first principal components for each country (0.5 p)
11. Plot the countries on the 2D scatter plot using the PCA scores (0.5 p)
12. Interpret the results (1 p)
13. Plot PCA loadings and scores on the same plot (make a biplot) (0.5 p)
14. How useful is the final plot? Interpret the results. (1 p)
15. Write a nice report with an introduction and conclusion (1 p)

Look and feel of the report is important – it can increase or decrease your total sum of points. Any relevant extra work earns you some bonus points but 10 is still a maximum that you can get.

Good luck and looking forward to reading your reports!

APPENDIX

Measure	Description
Income inequality	Average of the 20:20 income inequality published in the United Nations Development Program. Human development reports for years 2003, 2004, 2005, 2006
Trust	Percent of people who respond positively to the statement "most people can be trusted", World Values Survey, 1999-2001. Reverse coded in Index of Health and Social Problems
Life expectancy	Life expectancy at birth for men and women, years, 2004, UN Human Development Report, reverse coded in Index of Health and Social Problems
Infant mortality	Deaths in the first year of life per 1000 live births, 2005, OECD
Obesity	Percentage of the adult population with BMI greater than or equal to 30, averaged for men and women, 2002, International Obesity Taskforce
Maths and literacy	Combined maths and reading literacy scores of 15 year olds, 2003,

score	OECD Programme for International Student Assessment. Reverse coded in Index of Health and Social Problems
Teenage births	Births per 1000 women aged 15-19 years, 1998, UNICEF
Homicides	Homicides per million, period average for 1999-2000, United Nations
Imprisonment	Natural log of prisoners per 100,000, United Nations
Drugs index	Index of opiate, cocaine, cannabis, ecstasy and amphetamine use (average z-scores), 2007, United Nations Office on Drugs and Crime
Calorie intake	Calorie intake per capita per day, population averages over time series, OECD Health Database,
Public expenditure on health care	Public expenditure on health care as a proportion of total spending on health, 2003, World Health Organization
Child wellbeing	Based on the UNICEF index of child well-being in rich countries. We re-calculated the index, removing the % of children in relative poverty, and adding in countries that did not have data on all outcomes
Maths/literacy/science score	Combined maths, reading and science literacy scores of 15 year olds, 2003, OECD Programme for International Student Assessment
Spending on foreign aid	Spending on foreign aid as a percentage of Gross National Income, 2005, OECD
Peace index	Index of militarization with measures of domestic and international conflict, security, human rights and stability, 2007, Vision of Humanity and Economist Intelligence Unit
Maternity leave	Weeks of paid maternity leave, Clearinghouse on International Developments in Child, Youth and Family Policies, Columbia University
Advertising	Proportion of national GDP spent on advertising, 2002, World Advertising Center
Social expenditure	Public social expenditure as a proportion of national GDP, 2003,

	OECD
Women's status	Index of % of women in legislature, male-female income gap, % of women completing higher education, United Nations

Table 1 – Description of variables from the book "The Spirit Level". Source: (The Spirit Level international data, 2010)

Measure	Description
GDP per capita	Gross Domestic Product on a purchasing power parity basis divided by population as of 1 July for the same year, 2004. (The World Factbook 2005)
Press Freedom	Reporters Without Borders compiled this Index of 167 countries by asking its partner organizations (14 freedom of expression groups from around the world) and its network of 130 correspondents, as well as journalists, researchers, legal experts and human rights activists, to answer 50 questions designed to assess a country's level of press freedom. (Press Freedom Index 2005)
Population Density	Population of the country divided by the area in square kilometres. (The World Factbook 2005)
Electricity Consumption	Electricity consumption in kWh divided by the population. (The World Factbook 2005)
Internet Users	Number of users within a country that access the Internet divided by the population (The World Factbook 2005)
Index of Globalization	The KOF Index of Globalization measures the three main dimensions of globalization: economic, social and political. (Globalization Index 2005)

Table 2 – Description of extra variables.