MSPR 4 PCA and Probabilities (Due: 4.10.2015, 12 p.m. (noon))

Dr. Hendrik Purwins, Assistant Professor
Jan Stian Banas, TA
Dept. Architecture, Design and Media Technology, Aalborg University Copenhagen
A.C. Meyers Vænge 15, DK-2450 Copenhagen SV, Denmark

1. (Feedback) Please give us feedback on the last lecture and homework: `http://goo.gl/forms/Xdi5XjkxOR` Thanks!

2. Analyze the `adult` dataset. Use features age, education-num, sex, capital-gain, capital-loss, hours-per-week, and income ('>50k','<=50k'). Convert the categorical variables sex and income into a number (0,1), using the Matlab function `strcmp`.

   (a) Perform an eigenvalue decomposition of the covariance matrix of covariances between the first 6 features. Plot the cumulant relative eigenvalues. How much percentage of the variance is explained by the eigenvector with the second largest eigenvalue alone? How many eigenvectors seem to be enough to represent the data? (10P)

   (b) Which features dominate the eigenvector with the highest eigenvalue? (10P)

   (c) Reconstruct the adult data using just the scores on the eigenvector with highest eigenvalue. Plot the features capital-gain vs age for the reconstructed data and the original data. (10P)

   (d) Calculate the variance value of each feature and comment the reconstruction plot. Discuss how an eigenvalue decomposition on the correlation matrix could change the situation. (10P)

3. Recap Statistics. (no hand-in required)

   (a) Recapitulate what the the encircled questions on the pdf `field_hole_p274.pdf` are about (from Field/Hole: how to Design and Report Experiments , Sage 2003, p. 274-275.)

   (b) If the terms are unclear to you, watch the video `http://laugefelix.dk/medieology/hlecture4-1.` read the book (if you have it), or check on some of the terms in wikipedia.

4. Self Assessment: Check the exercises that you have seriously worked on.

2 a	2b	2 c	2 d