PR414 / PR813 — Principal Component Analysis (PCA)
Lecture 1 — Linear Discriminant Analysis (LDA)

This document is also available in PDF[1] format.

Purpose: Getting used to high-dimensional feature spaces. Introduction to PCA and LDA.

Material: LECTURE NOTES[2] on PCA/LDA, Devijver & Kittler chapter 9, Fukunaga, Therrien, Peebles.

General: To start off the course we are going to investigate the manipulation of (typically) multi-dimensional feature vectors. More specifically, we will use PCA to decorrelate such feature vectors, resulting in a system of (possibly) lower dimension.

Topics:
- General overview of a PR system.
- Orthonormal projection.
- Selection vs extraction of features.
- Basic PCA (KLT)
- LDA (class-based KLT).
- Using PCA to indicate the relative importance of the original features.

Project: (To be completed by the next lecture)
The directory **http://www.dsp.sun.ac.za/pr813/data/** contains various data sets that will be used in this course. Use the data as indicated below to do the following:
- Explain the differences between PCA based on correlation vs covariance matrices.
- The file `simvowel.tar.gz` contains the files `klasxa.txt` to `klasxu.txt`. These represent 5 different sets of simulated feature vectors. In each file, each row represents a single (4-dimensional) feature vector. Find a reduced subspace for this data, using both PCA and LDA. In the case of PCA, pool all the data together to form a single data set.
- The file `timit.tar.gz` contains data useful for speaker recognition. Each final subdirectory indicates a specific speaker. The `.cep` files are cepstra calculated from 16 ms frames of speech. See the `read.me` file in the timit directory for file formats. Use all the `/train/dr1` (training set, dialect region 1) speakers and determine the appropriate subspaces via PCA and LDA. In the case of PCA, pool all the data together to form a single data set.

  OR

  The file `faces.tar.gz` contains a database of face images from Tom Mitchell's website[3]. The files are in PNG format, which can be loaded into Matlab with the `imread` command. Each final subdirectory indicates a specific person. Convert each 30x32 image into a 960-dimensional feature vector by reshaping the matrix. Use all the "straight" faces and determine the appropriate subspaces via PCA and LDA. In the case of PCA, pool all the data together to form a single data set. Plot the global mean face as well as the principal component faces ("eigenfaces").
- How would you go about generating multi-dimensional Gaussian feature vectors with specified (non-zero) covariances between the components? (Hint: Have a look at the Choleski decomposition).
- Can you use the Choleski decomposition to decorrelate feature vectors?
- *(Bonus marks)* What is the result of a zeroth-order PCA of a data set?
- *(Bonus marks)* What is the relationship between the Karhunen-Loéve transform, the Fourier transform and the discrete cosine transform (DCT)?
- **Tip: Use the "test yourself" data provided in the notes to make sure your code is working.**

---

[1] `http://www.dsp.sun.ac.za/pr813/lectures/1_pca/1_pca.pdf`
[2] `http://www.dsp.sun.ac.za/pr813/lectures/lecture01/pr813_lecture01.pdf`
[3] `http://www-2.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/faces.html`