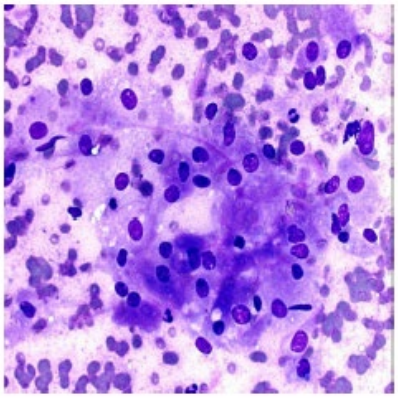
Instructions for the Group Exercise on Principal Component Analysis

We will be working with data from Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Cancer Letters, 77(2-3), 163-171. Back in 1994 (sic!) this paper used machine learning to predict whether breast tissue biopsy was cancerous or benign. Here is an example of the image data they worked with:



Black dots are cell nuclei. Irregular shapes or highly variable cell sizes can mean cancer, but it's tricky.

The sample contains 212 cancer patients and 357 healthy individuals (variable *cancer_yn*). Columns 1 through 30 of the table *cancerwdbc* contain 30 other aggregate characteristics of each patient's biopsy:

1	radius
2	texture
3	perimeter
4	area
5	smoothness
6	compactness
7	concavity
8	concave points
9	symmetry
10	fractal dim
11	radius_std
12	texture_std
13	perimeter_std
14	area_std
15	smoothness_std
16	compactness_std
17	concavity_std

18	concave points std
19	symmetry std
20	fractal dim std
21	radius extreme
22	texture extreme
23	perimeter extreme
24	area extreme
25	smoothness extreme
26	compactness extreme
27	concavity extreme
28	concave points extreme
29	symmetry extreme
30	fractal dim extreme

The names of these features are listed in the *feature_names* variable.

Assignment 4 (cancer data):

- Download the file cancer_wdbc.mat and load it into Matlab using > load cancer_wdbc.mat (be sure to save the file in your current Matlab directory)
- Data in the table *cancerwdbc* (569x30). The first 357 patients are healthy. The remaining 569-357=212 patients have cancer. This information is contained in the variable *cancer_yn*
- Carry out the PCA of the 30 variables describing biopsies. For this first calculate Z-scores of the variables in cancerwdbc (see the Matlab command zscores). Then carry out the PCA of the matrix of Z-scores. Matlab also has a dedicated pca command (read the manual)
- Which variables give the strongest positive or negative contributions to the 1st, 2nd, and 3rd largest eigenvalues?
- Plot the scores (Score=Z*V) of the 1st vs 2nd eigenvalues for healthy and cancer patients separately. Can these PCA scores be used to separate cancer from normal patients? Spoiler alert: they cannot, but you need to plot them separately for healthy and cancer patients. How about other pairs of PCA scores: 1 vs 3, 2 vs 3?