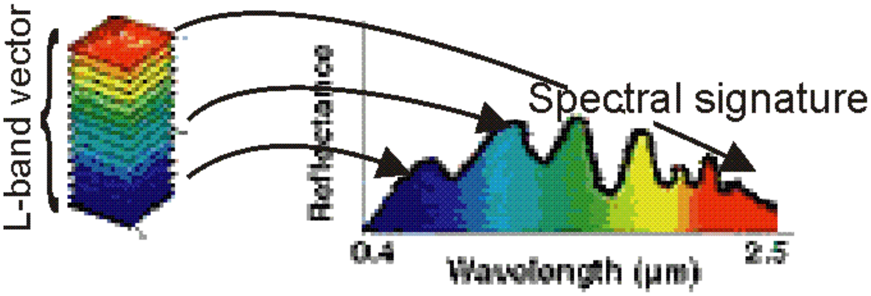
“Clustering Algorithms”

Project

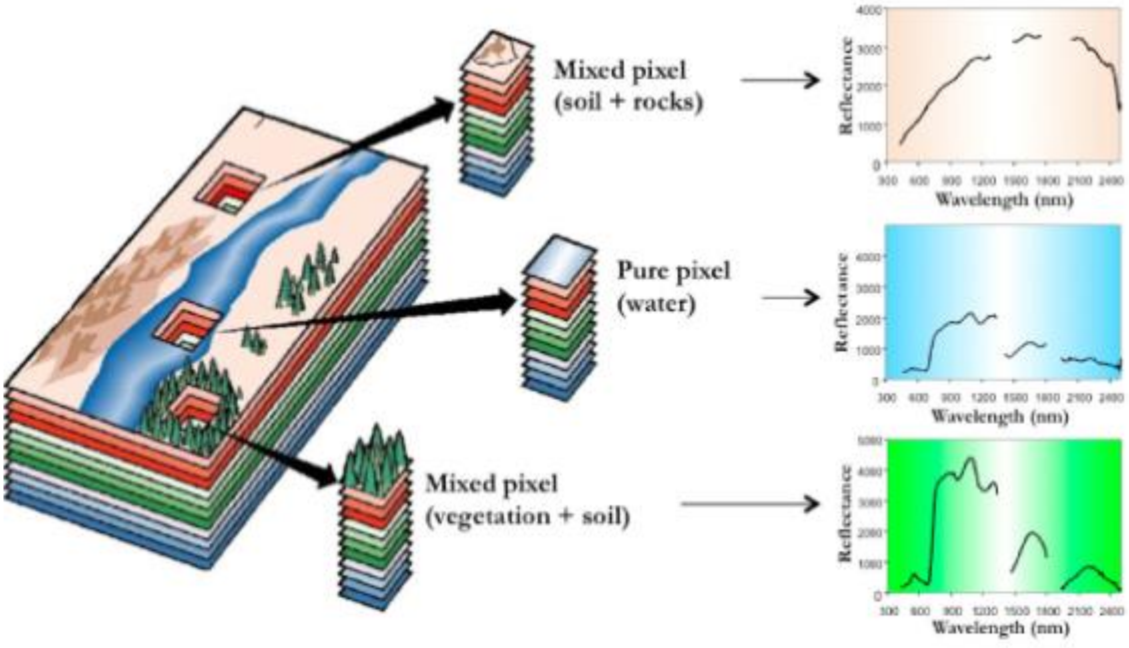
Introduction

This project is devoted to Hyperspectral Images (HSI) processing.

Hyperspectral images (HSIs): An HSI depicts a specific scene at several (*L*) narrow continuous spectral bands (actually, they visualize the reflectance of the depicted scene in various spectral bands). It can be represented by a *M*x*N*x*L* three-dimensional cube, where the first two dimensions correspond to the spatial information, while the third corresponds to the spectral information. Thus, the (*i,j*) pixel in such an image, *i*=1,…,*M*, *j*=1,…,*N*, is represented by an *L*-dimensional vector (constituted by the corresponding spectral bands), called the ***spectral signature*** of the pixel.



In several remote sensing applications, the HSIs (taken from satellites) that depict specific scenes of the earth surface at a specific spatial resolution (that is, a single pixel may represent an area from 3x3 m^2, to, e.g., 100x100m^2 or more). That is, each pixel is likely to depict more than one materials depicted in the corresponding area of the scene. Such pixels are called ***mixed pixels*** and they are the vast majority of the pixels in the image. On the other hand, there are (usually) a few pixels that depict a single material. These are called ***pure pixels***.



Processing in HSIs: The usual processing procedures in HSIs follow two main directions, namely, the ***spectral unmixing***[1] and the ***classification*** (supervised, unsupervised). In the sequel, we focus on the unsupervised classification – clustering.

[1] **Spectral unmixing (SU):** The problem here is stated as follows: Assume that a set of m spectral signatures corresponding to the pure pixels in the HSI under study is given[1]. For a given pixel in the image, the aim is to determine the percentage (***abundance***) to which each pure material contributes in its formation. It is clear, that SU provides ***sub-pixel information*** for a given pixel. Speaking in mathematical terms, let

(i) $\mathbf{y}$ be the (column L-dimensional) spectral signature of the pixel under study,

(ii) $\mathbf{x}_1$,… $\mathbf{x}_m$, be the spectral signatures (column L-dimensional vectors) of the pure pixels in the image (each one corresponding to a pure material met in the image) and

(iii) $\boldsymbol{\theta}$, the m-dimensional ***abundance vector*** of the pixel (its q-th coordinate corresponds to the percentage to which the q-th pure pixel contributes to the formation of the pixel under study).

Adopting the ***linear spectral unmixing hypothesis***, the above quantities are related as follows

$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ is an L-dimensional i.i.d., zero mean Gaussian noise vector. Note that, physically, the entries of $\boldsymbol{\theta}$ should be nonnegative and (ideally) they should sum to one.

A classical way to estimate $\boldsymbol{\theta}$ is to utilize the Squared Error criterion, using the non-negativity constraint on the coordinates of $\boldsymbol{\theta}$ and (possibly), in addition, the sum-to-one constraint (coordinates of $\boldsymbol{\theta}$ are required to sum to one).

All questions in this project refer to the so called “Salinas” HSI, which depicts an area of the Salinas valley in California, USA. It is a 220x120 spatial resolution HSI and consists of 204 spectral bands (from 0.2μm – 2.4μm) and its spatial resolution is 3.7m (that is, the HSI is a 220x120x204 cube). Thus, a total size of $N = 26400$ sample pixels are used, stemming from seven ground-truth classes: **Class 1**: “*grapes*”, **Class 2**: “*broccoli*”, **Class 3**: “*fallow 1*”, **Class 4**: “*fallow 2*”, **Class 5**: “*fallow 3*”, **Class 6**: “*stubble*”, **Class 7**: “*celery*”. Note that there is no available ground truth information for the dark blue pixels.

Aim of this project: Identification of homogeneous regions in the Salinas HSI.

NOTE: Only the **pixels** with **nonzero class label** will be **taken into consideration** in this project.

The data that will be used are (a) a 220x120x204 three-dimensional matrix named “***salinas_cube***” (the Salinas hypercube, in the file ***Salinas_cube.mat***) and (b) a 220x120 two-dimensional image named “***salinas_gt***” (the class label for each pixel, in the file ***Salinas_gt.mat***).

Description of the project

The goal here is compare the performance of (a) cost function optimization clustering algorithms (the k-means, the fuzzy c-means, the possibilistic c-means and the probabilistic (where each cluster is modelled by a normal distribution) clustering algorithms) on the one hand and (b) the hierarchical algorithms (Complete-link, WPGMC, Ward algorithms) on the other hand, in finding homogeneous regions in the Salinas HSI, focusing ONLY on the pixels for which the class label information is available. To this end:

(a) **execute** the above algorithms for various combinations of their parameters (e.g. the number of clusters etc) in order to identify the homogeneous regions in the image, reporting any problems you may spot for anyone of them.

(b) **verify qualitatively** the results obtained by the each algorithm, in terms of (i) the label information for the pixels and (ii) information that can be gathered by the image (for example, by examining the Principal Components of the image).

(c) **verify quantitatively** the results, in terms of the labeling information.[2]

[2] We will discuss some ways to perform this procedure in the class.

(d) **Comment** on the differences in the performance of the above algorithms.

NOTE: You can use any of the codes provided in the class for the cost function optimization algorithms. For the hierarchical algorithms you can use the MATLAB commands “linkage”, “cluster”, “crosstab” and set their parameters appropriately. Also, you can use parts of the attached MATLAB file. Finally, you can utilize the MATLAB function “pca_fun”, which determines the principal components of a data set.