Project Description

Martin Kvisvik Larsen

October, 2019

1 Problem Description

The problem that I am looking at is to do classification between 16 different classes of terrain based on the radiance spectra between 400 nanometres and 2500 nanometres. The classes of terrain vary from biological, for instance crops like corn and oats, to human-influenced terrain like buildings and mixtures of the two. For the data preprocessing I want to look at data filtering, outlier detection and feature selection. Specifically I want to apply a PCA-based outlier detection method, a moving average method for data filtering and whether or not derivatives or specific bands of the radiance spectras can be good features for classication. For the classification part I want to demonstrate the use of a linear classification method, like PCA-LDA or PLS-LDA, and a nonlinear classification method, either SVM or NN.

2 The Dataset

The dataset that I have chosen to analyze is the "Indian Pines Test Site 3" dataset provided by Purdue University in Indiana, United States. The dataset is from a flight line of an airborne AVIRIS hyperspectral imager that was flown over an area where two thirds is agriculture and one third is forest. The data acquisition was done during the month of June and hence some crops, like corn and soybeans, are present. The dataset contains spectral radiance measurements of 147 scan lines of 147 samples of 220 channels between 400 nm and 2500 nm concatenated into a $147 \times 147 \times 220$ hypercube and the ground truth class labels for each sample concatenated into a 147×147 array. Additionally calibration information to compute the measured radiance and mapping spectra indices to wavelengths is provided. However reference irradiance spectras from the atmosphere have not been provided with the dataset (if it ever was collected). This means that I am not able to correct for atmospheric and solar conditions, and that the inherent properties of the objects on the ground will be less apparent in the data. The provided ground truth classes and their respective number of samples can be seen in table 1 and the RGB channels of the hypercube can be seen in fig. 1.

Index	Class label	Number of samples
0	Background	-
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowd	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93

Table 1: Dataset ground truth classes and their respective number of samples.

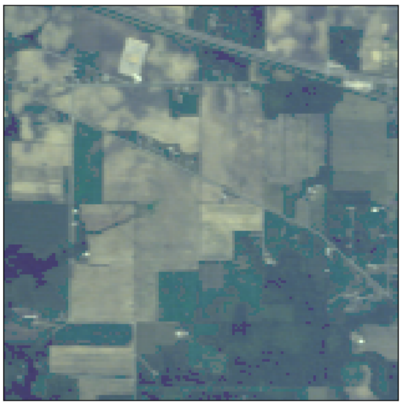

Figure 1: RGB channels of the hypercube.

3 Preprocessing and Data Inspection

The spectras provided in the dataset are offset and scaled versions of the measured spectral radiance. In order for us to retrieve the measured spectral radiance spectras the scaling and offset factors are provided. The spectral radiance spectras for scan line m and sample n are then obtained from the data X_{raw}, offset o and scale s as follows:

$$E(m,n) = \frac{X_{\text{raw}} - o}{s} \tag{1}$$

From the spectral radiance the measured radiance can be found from the following equation:

$$E(m,n,\lambda) = E(m,n) \cdot \lambda \tag{2}$$

The obtained spectral radiance and radiance spectras for the first scan line can be seen in fig. 2 and fig. 3.

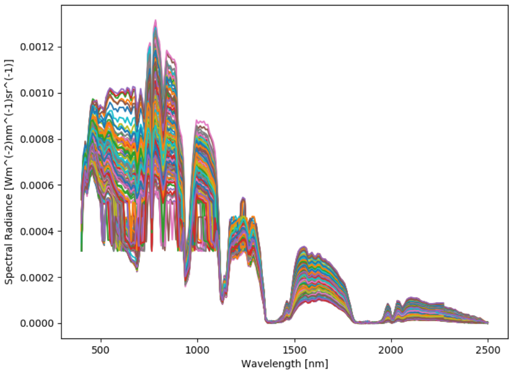

Figure 2: Spectral radiance spectra of the first scan line.

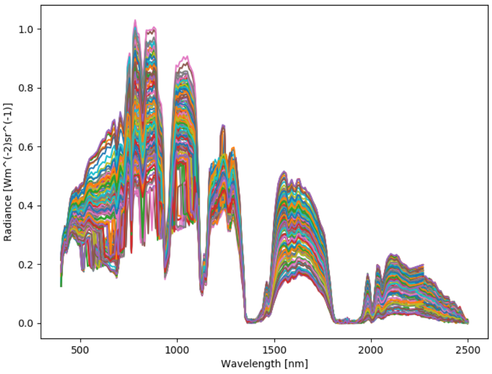

Figure 3: Radiance spectra of the first scan line.

From the obtained radiance spectra in fig. 3 one can see that there are five quite distinctive wave bands were the measured radiance is quite low. The wave band around 800 nanometres is the absorption band of oxygen, while the other four are the absorption bands of water. Using this domain knowledge to my advantage, the number of features might be reduced by ignoring these bands. Additionally one can see that the measurements contains some noise and that some of the sensor readings are a bit off. The noise is not a problem for the application of PCA or PLS, but differentation of the spectras enhance the noise. Therefore the noise is a problem when computing additional feature in the form of derivatives.

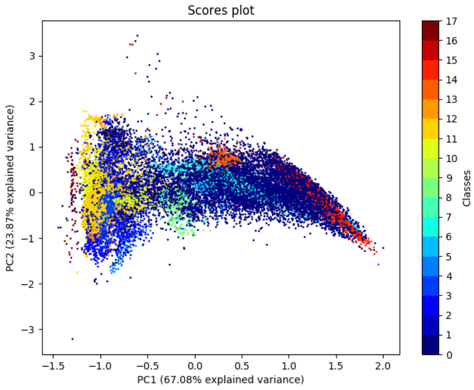

Figure 4: PCA scores of the first two principal components.

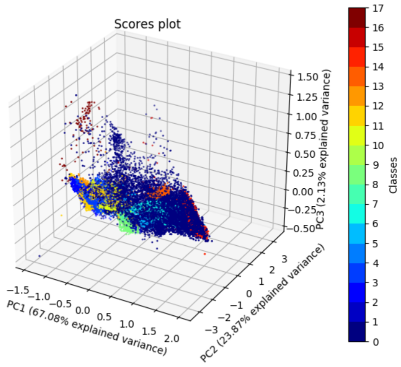

Figure 5: PCA scores of the first three principal components.

The score plots in fig. 4 and fig. 5 were constructed by performing PCA on all the radiance spectras as a mean of inspecting the data. About 90 % of variance in the data is explained in the first two principal components. From the score plot one can see tendencies of class clusters, but that in general the classes are intertwined. Additionally there are tendencies of some outliers.

4 Further Work

Based on the score plots, classification on the principal components seems to be challenging for this dataset due to the overlapping class clusters. The plan forward is to remove outliers by filtering out observations based the distance measure of the PC-scores, and see if this has any effect on the principal components. Additionally I want to construct more features by filtering the radiance spectras and computing their 1st and 2nd derivative. From bio-optics and chemometrics I know that peaks and curvatures are good features for classification. Hopefully these would provide more good features for discriminating between the classes. If I manage to find the time I might do some statistical analysis on the readings from each individual sensors (i.e. the different samples of each scan line) to correct for differences in the sensors. These are the main elements of the data preprocessing procedure that I want to do before performing discriminant analysis on the data. For the discriminant anaylysis I first want to analyze the problem as a all-vs-all discrimination problem and then as a one-vs-rest if there is time. I want to use k-fold cross validation for model selection and then just a regular test set for model assessment. I will start by using the linear PCA-LDA model before utilizing the non-linear SVM model.