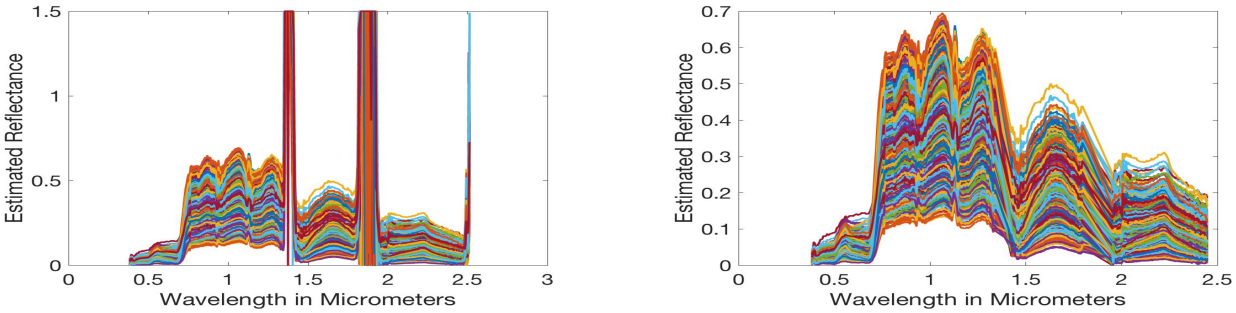
Comparing Dimensionality Reduction Techniques

Hudanyun Sheng
Dept. Electrical and Computer Engineering
University of Florida
Gainesville, FL
hdysheng@ufl.edu

Dylan Stewart
Dept. Electrical and Computer Engineering
University of Florida
Gainesville, FL
d.stewart@ufl.edu

I. Methodology

A. Removing Bad Bands



(a) Before bad band removal

(b) After bad band removal

Fig. 1: Before and after bad band removal

B. SVM Training

Given that our objective is comparing dimensionality reduction methods and not optimizing an SVM classifier, we use the built in MATLAB SVM functions and do not modify any of the classification parameters manually. Each method is trained and tested separately and have no influence on the model parameters of any other method.

C. Principal Component Analysis

PCA projects the data into a subspace that maximizes variance. Considering the notion of using PCA in dimensionality reduction and evaluation by classification accuracy there are two main quantitative values of importance that influence our parameter selection experiments. Mainly, how well can we classify the data using $D < B$ features or bands? More subtly, how much variance are we keeping by selecting this number of principal components? Is maximizing variance maximizing separability?

For our first experiment of comparing the number of bands, we vary the number of bands to keep D from the maximum number of bands D_{max} to 1 and plot the classification accuracy and variance kept. A supplemental plot of the lowest varying set of eigenvectors is also constructed. Lastly we show a plot of our data in 3 dimensions using the top 3 components and compare to the lowest 3 components that are greater than 0.

D. Maximum Noise Fraction

Similarly to PCA, MNF is a linear projection into a subspace. However, in this case that subspace maximizes the SNR. We are interested in estimating the noise correctly. In theory, if we estimate the noise well and the classes are separable and compact, classification rate should be high. Given the parameters for MNF we choose varying amounts of noise to be added to the spectra of the classes for our MNF estimation. We also choose a varying amount of components given that we assume some bands are primarily noise.

Given that this method relies on two main inputs (noise and number of components), we ran a single experiment that varies the variance of the noise added given 0 mean and the number of components from D_{max} to 1. We aim to view the effects of each parameter individually and their possible influence on each other.

The authors would like to thank Darth Gader for the base code of these experiments.

E. Hierarchical Dimensionality Reduction

Unlike the previous methods, HDR is not a linear projection. HDR merges similar bands and therefore we must only consider modifying the number of bands desired and how we compute similarity. In the code we were given, our HDR is performed by merging the two most similar bands by their Jensen-Shannon Divergence (JSD) [4]. Given that the author of this paper has yet to publish his methods on comparing high-dimensional distributions, we decided to modify HDR with JSD by using a parzen window estimator [5].

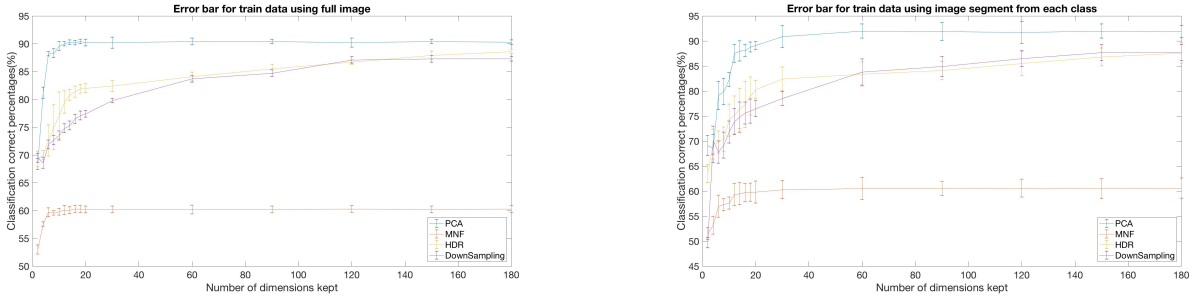
For this method we will run experiments with varying numbers of bands and compare the performance of the JSD with and without parzen window estimation.

F. Spectral Downsampling

Given that this method is simply throwing out bands and averaging them we conduct an experiment with varying number of bands for averaging a group of bands and compute the classification accuracy.

II. Results

We first looked at each method over the training data.



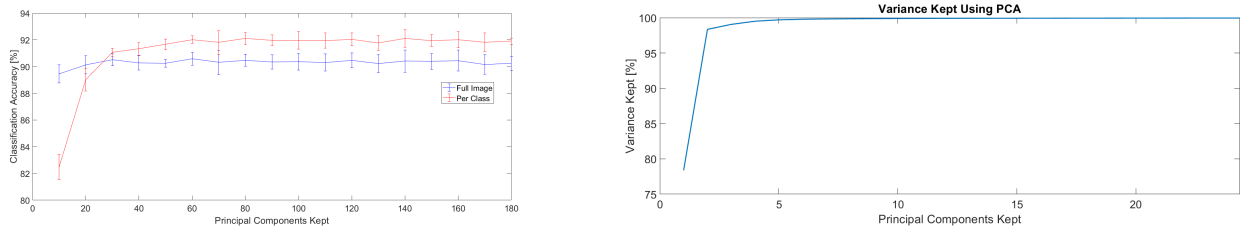
(a) Full Image Training

(b) Class Specific Training

Fig. 2: 10 fold cross-validation means and 2 standard deviation error bars

A. Principal Component Analysis

Our first PCA experiment compares PCA given a specific number of bands. In order to understand how much variance we are keeping we look at the variance from samples of the full image in relation to the number of kept principal components.



(a) PCA Training

(b) Variance over 1 run

Fig. 3: 10 fold cross-validation means and 2 standard deviation error bars for training and % variance kept for one run

We also look at the projection of the data in 3 dimensions.

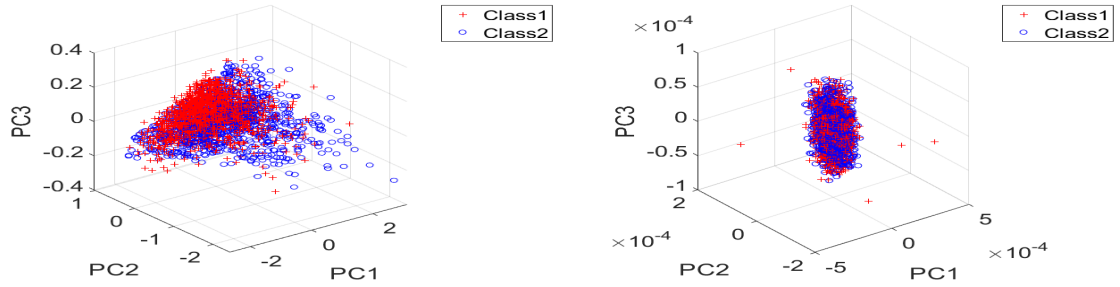
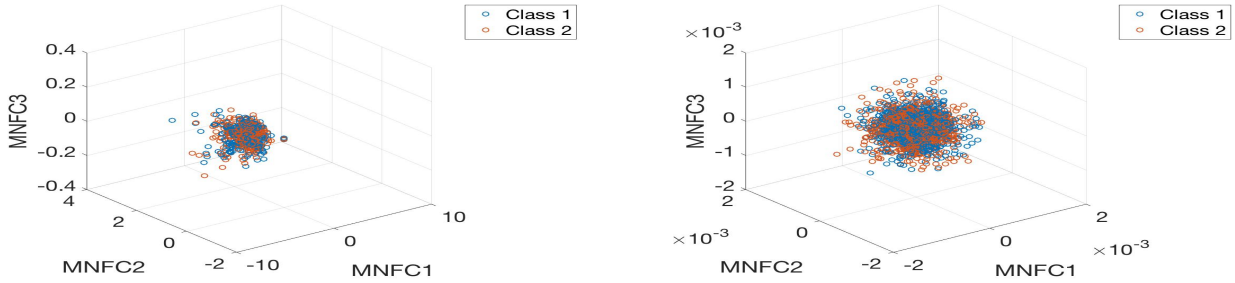


Fig. 4: L:R Top 3 and bottom 3 PCA components

B. Maximum Noise Fraction

Similarly to PCA we looked at the first and last 3 components of MNF to view possible separation.

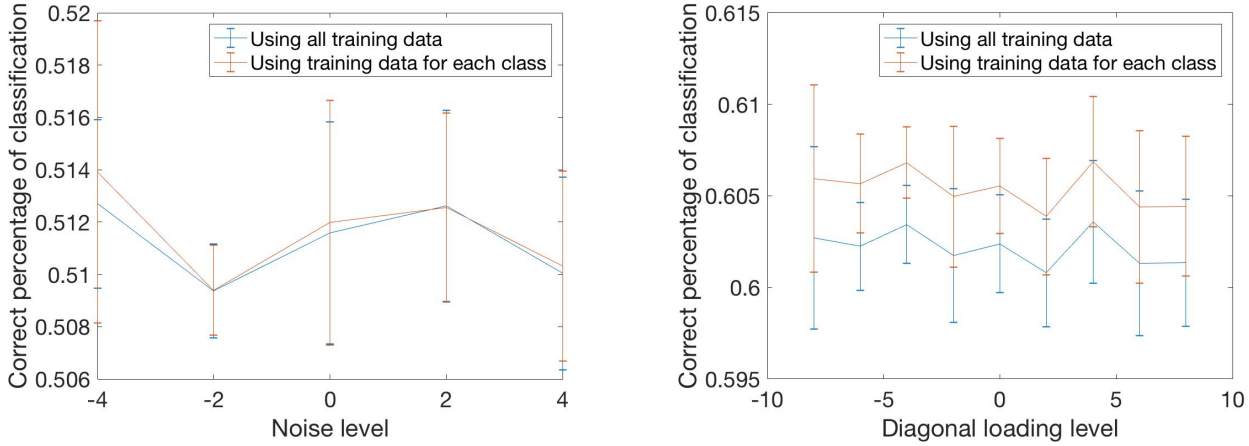


(a) Top 3 MNF Components

(b) Last 3 MNF Components

Fig. 5: Top 3 and bottom 3 MNF components

If we add noise with different covariances or different diagonal loading numbers, the results are shown below:



(a) Different performances with different covariance matrix of the noise added

(b) Different performances with different diagonal loading factor

Fig. 6: The effect of different covariance matrices or diagonal loading factors.

C. Hierarchical Clustering

For Hierarchical Clustering we ran experiments with and without a parzen window estimator with window size 4.

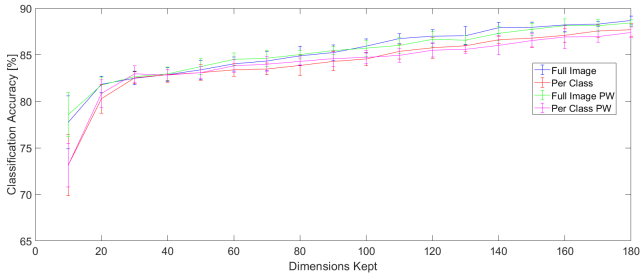
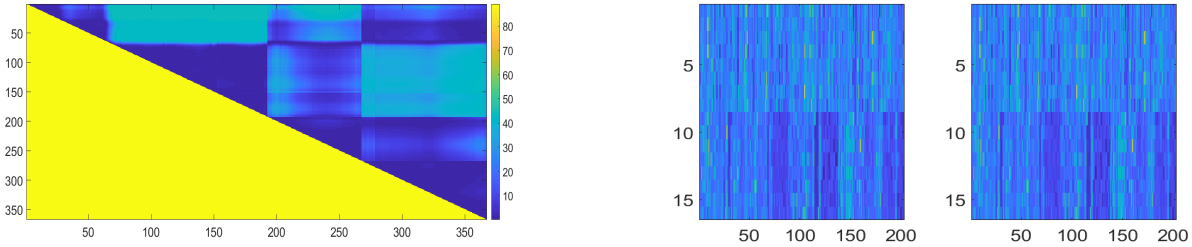


Fig. 7: 10 fold cross-validation with means and 2 standard deviation in error bars

We also wanted to look at which bands have the smallest KL-divergence to ensure that the method is picking similar bands.



(a) Similarity between bands using parzen windowed smoothed pdf and KL-Divergence

(b) Most similar bands using parzen windowed smoothed pdf and KL-Divergence

Fig. 8: Verifying parzen window and KL-divergence

Classification results for the full image and per class are shown in tables I and II and tables III and IV respectively.

No. experiment	PCA	MNF	HDR	Down sampling
1	90	60	81	78
2	90	60	83	77
3	90	60	82	77
4	90	60	82	77
5	90	60	82	78
6	90	60	82	77
7	90	60	81	77
8	90	60	82	78
9	90	60	82	78
10	90	93	86	87
mean	90.2	60.2	82.0	77.4
std.dev	0.3	0.3	0.4	0.3

TABLE I: Table of training correct percentages when using whole data set to reduce dimensionality

No. experiment	PCA	MNF	HDR	Down sampling
1	91	61	82	76
2	90	60	82	77
3	90	60	81	77
4	90	59	82	77
5	90	61	81	77
6	89	60	81	77
7	91	59	83	78
8	90	61	83	76
9	90	58	81	77
10	90	61	82	77
mean	90.1	60.1	81.7	77.0
std.dev	0.1	1.1	0.6	0.7

TABLE II: Table of test correct percentages when using whole data set to reduce dimensionality

No. experiment	PCA	MNF	HDR	Down sampling
1	89	59	80	77
2	89	60	81	76
3	89	60	81	76
4	89	60	81	77
5	89	60	80	77
6	90	60	81	76
7	89	60	78	76
8	89	60	79	77
9	89	60	80	77
10	89	59	81	77
mean	89.1	59.8	80.3	76.5
std.dev	0.3	0.3	1.2	0.4

TABLE III: Table of training correct percentages when reduce dimensionality for each class

No. experiment	PCA	MNF	HDR	Down sampling
1	89	61	79	76
2	89	60	82	76
3	89	60	80	77
4	89	59	81	76
5	89	60	79	76
6	89	60	81	76
7	89	59	79	78
8	90	60	79	75
9	89	57	79	76
10	88	61	81	76
mean	89.0	59.6	80.0	76.3
std.dev	0.4	1.1	0.9	0.8

TABLE IV: Table of test correct percentages when reduce dimensionality for each class

D. Pavia MNF

For the Pavia data set, two versions of original data set are presented: the Pavia and the squeezed Pavia. The visualization of the mean over all bands for both of the fake images are shown below:

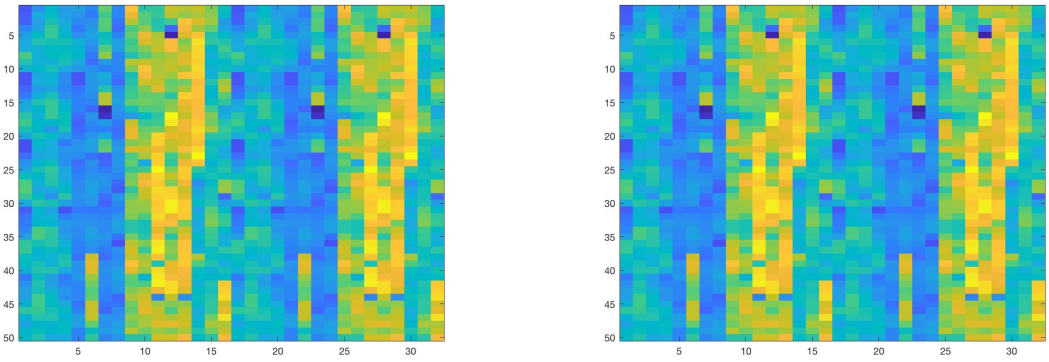


Fig. 9: Visualization of mean value of the original data set over all bands. Left: fake Pavia image. Right: squeezed fake Pavia image.

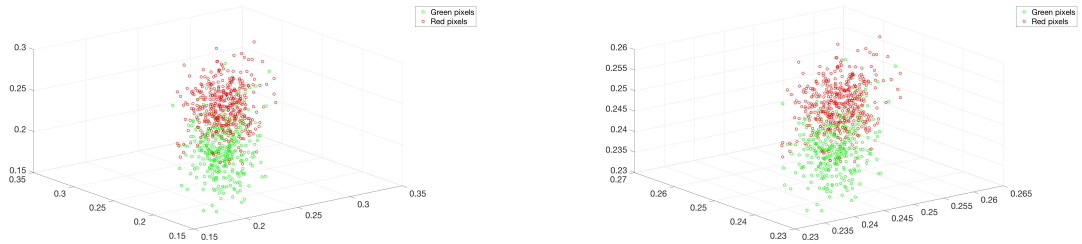


Fig. 10: Scatter plots of the original data set before dimensionality reduction. Left: fake Pavia image. Right: squeezed fake Pavia image.

The noise levels are changed to be $10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}$, the mean over all bands with noise added for the fake Pavia image are shown below in figure 11; the corresponding mean over all bands of the corresponding dimensionality reduction results are shown in figure 12; and the corresponding scatter plot of first 3 components after dimensionality using MNF is shown in figure 13.

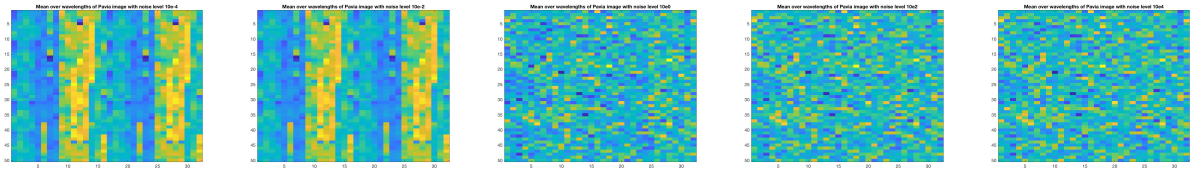


Fig. 11: Mean over all bands after adding noise for fake Pavia image

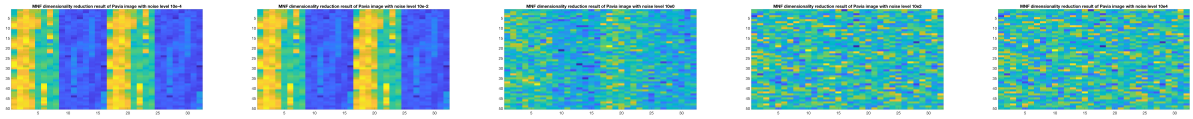


Fig. 12: Mean over all bands after MNF dimensionality reduction with noise added to the fake Pavia image

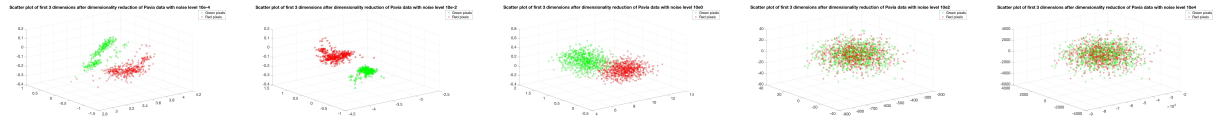


Fig. 13: Scatter plot of first 3 components after dimensionality using MNF with noise added to the fake Pavia image

The mean over all bands with noise added for the squeezed fake Pavia image are shown below in figure 14; the corresponding mean over all bands of the corresponding dimensionality reduction results are shown in figure 15; and the corresponding scatter plot of first 3 components after dimensionality using MNF is shown in figure 16.

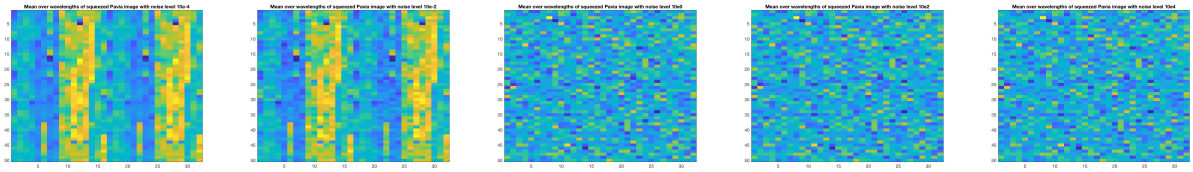


Fig. 14: Mean over all bands after adding noise for squeezed fake Pavia image

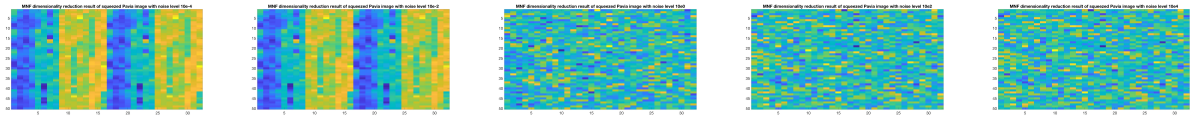


Fig. 15: Mean over all bands after MNF dimensionality reduction with noise added to the squeezed fake Pavia image

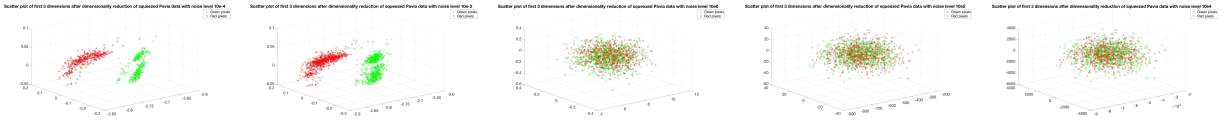


Fig. 16: Scatter plot of first 3 components after dimensionality using MNF with noise added to the squeezed fake Pavia image

III. Discussion

Before discussing each individual experiment it is important to note a few main factors that effect all testing results. All parameters for our testing results were selected directly from our cross-validated training results. It is also important to recall our removal of bad bands. Each dimensionality reduction method depends upon the data given to it, if there are any errors in our bad band removal step as described in section A of the methodology, these will be passed through to the rest of the experiments. Lastly, due to spectral down sampling needing to be passed a number at least or lower than half of the amount of features originally used, we did not experiment with dimensionality greater than half the original dimensionality after removing bad bands.

Within our investigation into PCA, we noticed that training on a per class basis was more successful than the entire image for higher numbers of principal components as seen in figure 3a. In our view, this occurs because of what PCA aims to do in the first place - project the data in directions of decreasing variance. When a single transform is used (full image plot), the data is projected into directions that represent the most variance for the entire image and not on a class by class basis. It is fairly intuitive that by building a transformation that maximizes variance specifically for each class, performance should be improved. Each class should have a unique transformation that helps to separate it from others. Another interesting plot shown in figure 3b demonstrates that although we have many many features, variance can be maximized in around 10-20 principal components. This does not mean that only 10-20 of our features are useful or contain variance. This merely states that based on a single experiment from the training data, the data can be projected into a new space where 10-20 principal components contain almost all of the variance. At around 3 principal components, 97% of the variance is preserved. Does this mean that those 3 components that contain most of the variance can also separate the data? Not necessarily. As seen in figure 4, although

the top 3 principal components contain $> 95\%$ of the variance, these 3 components do not separate the classes. We included the plot of the bottom 3 components to show that neither maximum nor minimum variance in the PCA space can separate these classes.

Before a deeper investigation into Hierarchical Clustering, we verified the computation of KL-divergence as seen in figure 8b. In figure 7, there are slight differences in classification accuracy, however it is important to remember this is for 10 experiments. Random initializations of the training data can cause slight variations in the features chosen during HDR and consequently the classification accuracy. Overall HDR performs relatively robustly with small changes in classification accuracy when removing small numbers of features. The largest change in performance happens between 10 and 20 features and this drop can be seen in other methods. As a result of this trend over all other methods we consider this to be inherent to the data chosen and the difficulty of separating the two classes.

Lastly, we investigated down sampling by simply averaging bands. This is by far the most intuitive method and it shows some interesting results when compared with the other methods on the training data in a cross-validation scheme. In both full and per class training methods, down sampling has comparable performance with HDR until around 60 features. This leads to an interesting question that can not be answered given the time allotted for this investigation: could a combined down sampling and HDR method be used to quickly reduce dimensionality of data for a classification approach? As can be easily understood, HDR relies on a square similarity matrix. Instead of computing this large matrix from the beginning, maybe some bands can be initially averaged and then a much smaller similarity matrix can be constructed to enhance the speed of computation? This is outside of the scope of our problem, but an interesting question nevertheless.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[2] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, pp. 65–74, Jan 1988.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.

[4] B. Fuglede and F. Topsoe, "Jensen-shannon divergence and hilbert space embedding," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, pp. 31–, June 2004.

[5] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, pp. 51–83, Jan 1978.