EMPIRICALLY COMPARING TWO DIMENSIONALITY REDUCTION TECHNIQUES – PCA AND FFT: A SETTLEMENT DETECTION CASE STUDY IN THE GAUTENG PROVINCE OF SOUTH AFRICA

T.L. Grobler†, *W. Kleynhans*⋆ *and B.P. Salmon*‡

†Dept of Mathematical Sciences, Computer Science Division, Stellenbosch University,
Private Bag X1, 7602 Matieland, South Africa
⋆Department of Electrical, Electronic and Computer Engineering University of Pretoria,
Pretoria 0002, South Africa
‡School of Engineering, University of Tasmania, Hobart, TAS 7001, Australia

ABSTRACT

In this paper we present a class label agnostic dimensionality reduction comparison framework. We illustrate the usefulness of this framework at the hand of a case study. For our case study, we consider two prominent land cover classes in the Gauteng province, namely natural vegetation and settlement using an 8 year MODIS dataset. We use the framework to compare two feature extraction techniques, namely PCA and FFT. For the case study we considered in this paper, the PCA technique produced a reduced feature space which was 15% more separable than the feature space produced by the FFT method.

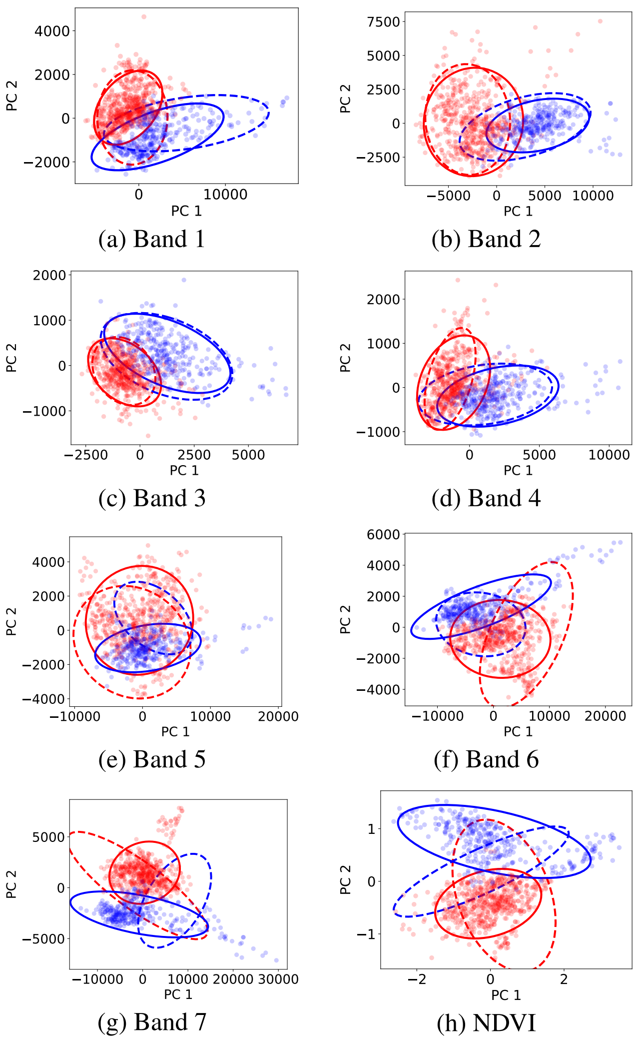
Index Terms— Principal Component Analhysis (PCA), harmonic analysis, hypertemporal remote sensing.

1. INTRODUCTION

Many dimensionality reduction techniques have been proposed in the remote sensing literature [1]. However, there are few empirical comparison studies between the different methods in the literature (especially in the case of hypertemporal remote sensing data). The most straightforward approach one could follow to perform such an empirical study would be to first create x intermediate datasets by applying x reduction techniques on an original dataset. One would then use the x intermediate datasets to train x supervised classifiers (each of the aforementioned classifiers should use the same underlying classification strategy). The last step of this approach would then involve comparing the accuracy of the x classifiers with one another. One drawback of such an approach is that this approach can only be used if your original dataset is labelled. In this paper we present a basic comparison framework that can be used to compare the efficacy of different reduction techniques whether the original dataset is labelled or not. To illustrate how this framework functions, we employ it and a case study to compare two hypertemporal dimensionality reduction techniques, namely PCA (Principal Component Analysis) and the Fast Fourier Transform (FFT). For our case study we consider a hypertemporal Moderate Resolution Imaging Spectroradiometer (MODIS) dataset containing both vegetation and settlement time-series. Settlement expansion is one of the most pervasive forms of land cover change in southern Africa. It is therefore of the utmost importance to be able to discern between these two land cover types [1]. PCA has been applied extensively to remote sensing data [2] and in particular to remote sensing time-series [3]. PCA remains a popular dimensionality reduction technique. It was recently used to analyze the spatio-temporal variability of the Pantanal vegetation cover (which is the largest tropical wetland in the world) [4]. Harmonic features (i.e. the FFT) have also been used extensively for land cover classification and change detection [5]. It was recently used in a tree/grass fractional cover case study conducted in the Kruger national park South Africa [6]. We start the paper by discussing the dataset we considered and then we present the feature extraction comparison framework. Lastly we present our results and conclusions.

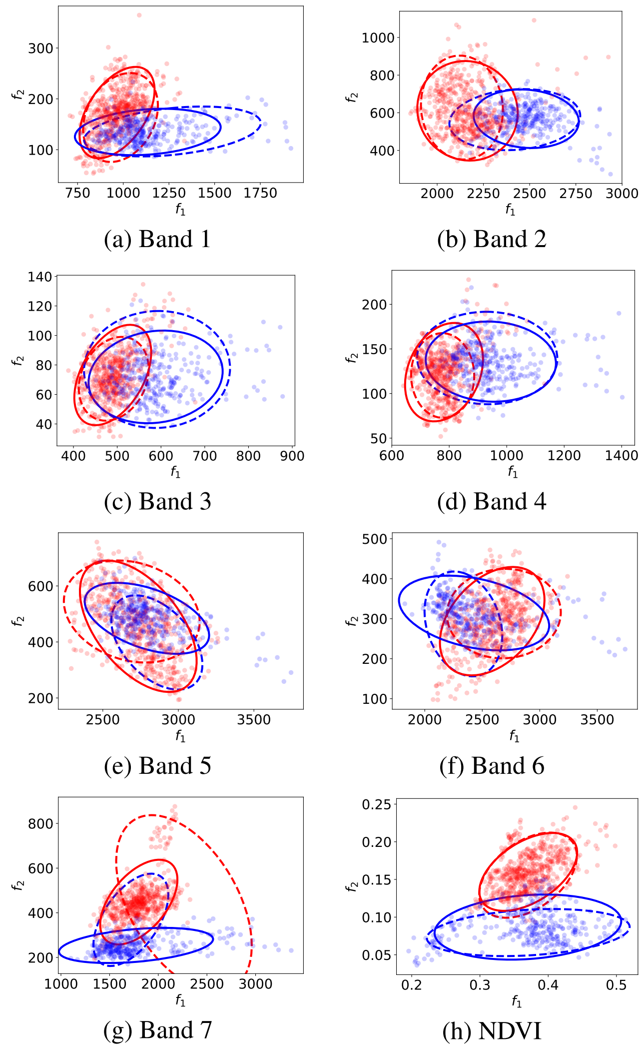
2. DATA DESCRIPTION

The hypertemporal dataset that we used contains MODIS MCD43A4 BRDF (Bidirectional Reflectance Distribution Function) corrected 500 m land surface data (corresponding to a total area of approximately 230 km^2 of the Gauteng province of South Africa). The temporal cadence of the data is 45 observations a year (one every 8 days) In this paper we consider two classes of land cover, namely vegetation and settlement. The settlements class contains pixels (333 pixels) consisting of about 50% buildings, and 50% vegetation, whereas the vegetation class contains pixels (592 pixels) which contain more than 90% vegetation. Each pixel consist of eight time-series that contain 368 samples. The eight time-series can be associated with the first seven MODIS bands and the Normalized Difference Vegetation Index (NDVI). We selected MODIS pixels that according to Systme Probatoire dObservation de la Terre (SPOT) images had the appropriate percentage land cover type in a MODIS pixel and did not change from 2000 to 2008 [1].



(a) Band 1 (b) Band 2 (c) Band 3 (d) Band 4 (e) Band 5 (f) Band 6 (g) Band 7 (h) NDVI

Fig. 1. The reduced feature space obtained after running PCA on the Gauteng dataset (i.e. we plot the feature matrix $\mathbf{Y}_2$ for each MODIS band). The first principal component is associated with the x-axis, while the second principal component is associated with the y-axis. We plotted the features using their corresponding labels (red for vegetation and blue for settlements). The solid ellipses represent the 95% confidence interval of the Gaussian density which was fitted on the feature space of each land cover class. The dashed ellipses represent the 95% confidence interval of the obtained GMM mixture components that were fitted to the dataset. The raw MCD43A4 data are digital numbers (16-bit unsigned integer values – did not convert to reflectance ratios).



(a) Band 1 (b) Band 2 (c) Band 3 (d) Band 4 (e) Band 5 (f) Band 6 (g) Band 7 (h) NDVI

Fig. 2. The reduced feature space obtained after running the FFT on the Gauteng dataset (i.e. we plot the feature matrix $\mathbf{Y}_2$ for each MODIS band). The mean of each pixel is associated with the x-axis, while the seasonal amplitude of each pixel is associated with the y-axis. We plotted the features using their corresponding labels (red for vegetation and blue for settlements). The ellipses represent the 95% confidence interval of the Gaussian density which was fitted on the feature space of each land cover class. The dashed ellipses represent the 95% confidence interval of the GMM mixture components that were fitted to the dataset.

3. METHODOLOGY

We describe the proposed comparison framework in Section 3.1. We then discuss the two dimensionality reduction techniques we considered in Section 3.2 and Section 3.3. We end the section by discussing the distance metric we used to quantify the separability between the two land cover classes (after feature extraction).

3.1. Comparison Framework

The proposed comparison framework:

Feature Reduction Apply the different feature reduction approaches to either a labelled or an unlabelled hypertemporal dataset (repeat for each spectral band).

- In this paper, we used the SCIKIT-LEARN library to run PCA on each MODIS band [7]. We then stored the two largest principal components for each band in a feature matrix (Section 3.2).
- In this paper, we used the NUMPY library to extract the mean and seasonal component out of each MODIS pixel for each band and stored the result in a feature matrix (Section 3.3) [8].

Density Estimation If the dataset is labelled then we estimate the densities associated with the reduced feature space for each feature reduction method and land cover type and if it is unlabbeled we fit a mixture model to the reduced feature space instead.

- In the case of a labelled dataset we assumed that the feature space associated with each land cover type (and band) could be represented by a Gaussian density. In the case of an unlabelled dataset we assumed that the reduced feature space could be represented by a Gaussian Mixture Model (GMM) (we assumed two components, i.e. semi-supervised). We only considered two features (ensure fairness). The aforementioned densities were estimated using SCIKIT-LEARN.

Hellinger Distance Compute the Hellinger distance between the estimated densities (or the different mixture components) associated with the different land cover classes (or clusters) for each reduction technique (Section 3.4). A Hellinger distance close to 0 implies that the densities are inseparable, while a value close to 1 implies the opposite.

3.2. Principal Component Analysis (PCA)

The two largest principal components associated with the time-series of each MODIS pixel were extracted ($\sim$ 66% of variance captured) and stored in a feature matrix for each band. The above is realized via:

- Let $\mathbf{X}_b$ denote an $n \times t$ centered matrix (i.e. column means have been subtracted and are now equal to zero) containing MODIS band b hypertemporal data (to reduce clutter we will omit the b subscript in the rest of the paper). Moreover, assume that n denotes a MODIS pixel index and that t denotes a time-step index.
- Form the $t \times t$ covariance matrix $\mathbf{C}$ as follows:
$$\mathbf{C} = \frac{\mathbf{X}^T\mathbf{X}}{n-1}. \tag{1}$$
- Diagnonalizing $\mathbf{C}$ results in:
$$\mathbf{V}\mathbf{L}\mathbf{V}^T, \tag{2}$$
where $\mathbf{L}$ is a $t \times t$ diagonal matrix containing the eigenvalues of $\mathbf{C}$, in descending order, on its diagonal and $\mathbf{V}$ is a $t \times t$ matrix containing the eigenvectors associated with the eigenvalues found in $\mathbf{L}$ (each column of $\mathbf{V}$ contains an eigenvector).
  - The eigenvectors contained in $\mathbf{V}$ are also known as the principal axes or directions of the data.
  - The principal components are computed by projecting the data onto the principal axes.
- Compute the principal components via $\mathbf{Y} = \mathbf{X}\mathbf{V}$ (the dimensions of $\mathbf{Y}$ is the same as $\mathbf{X}$).
  - The jth principal component is located in the jth column of $\mathbf{Y}$.
- Perform dimensionality reduction using $\mathbf{Y}_2 = \mathbf{X}\mathbf{V}_2$, where $\mathbf{V}_2$ denotes the matrix obtained by keeping only the first two columns of $\mathbf{V}$. $\mathbf{Y}_2$ denotes the reduced $n \times 2$ feature matrix containing the first two principal components of $\mathbf{X}$.

3.3. Harmonic Features

The two largest harmonic components associated with the time-series of each MODIS pixel were extracted for each band via the Fast Fourier Transform and stored in the feature matrix $\mathbf{Y}_2$ (the dimension of $\mathbf{Y}_2$ is $n \times 2$). The above is realized via:

$$[\mathbf{Y}_2]_{ij} = \begin{cases} |\mathcal{F}[x_i(t)][0]| & \text{if } j = 1 \\ 2|\mathcal{F}[x_i(t)][f_s]| & \text{if } j = 2 \end{cases}. \tag{3}$$

In the above equation $x_i(t)$ denotes the time series of MODIS pixel i in band b, $\mathcal{F}\{\}$ denotes the Fourier Transform and $f_s = \frac{1}{45}$ Hz [1].

3.4. Hellinger Distance

The Hellinger distance (HD) between $P \sim \mathcal{N}(\mathbf{u}_1, \mathbf{\Sigma}_1)$ and $Q \sim \mathcal{N}(\mathbf{u}_2, \mathbf{\Sigma}_2)$ is defined as:

$$H(P,Q) = \sqrt{1 - \frac{|\mathbf{\Sigma}_1|^{\frac{1}{4}}|\mathbf{\Sigma}_2|^{\frac{1}{4}}}{|\mathbf{M}|^{\frac{1}{2}}}\exp\{-\frac{1}{8}\mathbf{u}^T\mathbf{M}^{-1}\mathbf{u}\}}, \tag{4}$$

$$\mathbf{u} = (\mathbf{u}_1 - \mathbf{u_2}), \quad \mathbf{M} = \frac{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2}{2}. \tag{5}$$

4. RESULTS

We plot the reduced feature space obtained using the framework described in Section 3 on the dataset (a labelled and an unlabelled version of this dataset) described in Section 2 in Figs. 1 and 2. The graphs in Figs. 1 and 2 suggest that PCA outperforms the FFT method, i.e. the two classes are more separable in the reduced feature space associated with PCA than the feature space generated by the FFT technique (wheter we incorporated the labels of the aformentioned dataset into our analysis or not). This observation is confirmed by Fig. 3, which indicates that PCA produces a feature space which is about 15% more separable than the feature space produced by the FFT method (not in an absolute sense – i.e. good indicator of separability). Figs. 1–3 show us that:

Labelled data The two land cover classes are the most separable in MODIS band 7 and NDVI. PCA and FFT perform similarly for these two bands. The two land cover classes are the least separable in MODIS band 5 and 6. PCA performs significantly better than the FFT method in these two bands. The land cover classes are somewhat separable in MODIS bands 1, 2, 3, and 4. PCA performs slightly better than the FFT approach in these bands.

Unlabelled data If the data is not separable enough, or cannot be accurately represented by the chosen mixture model then the results produced by our framework has no physical meaning. This is the case for band 5, 6, 7 and NDVI. The only exception being: the FFT technique applied to NDVI. In contrast, if the aforementioned conditions are met then the framework performs well and produces results similar to the case when the framework is applied to labelled data. This is the case for band 1, 2, 3 and 4. The only major difference being: the FFT technique performs a little bit better than PCA in band 1.

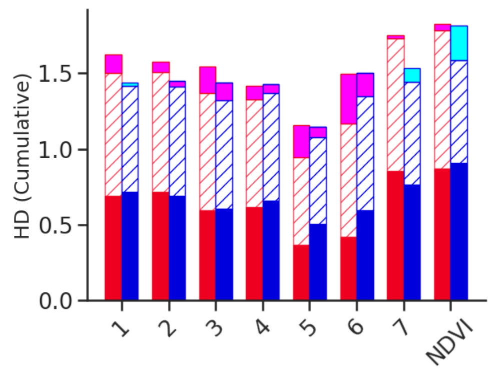


Fig. 3. The Hellinger distance between the densities in Figs. 1 and 2 for the two feature reduction methods we investigated, namely PCA (filled lowest segment) and FFT (second lowest hatched segment). Moreover, the results we obtained by applying our framework to labelled and unlabelled MODIS data are depicted in red and blue, respectively. We also plot the difference between the two reduction methods in magenta (PCA>FFT) and cyan (FFT>PCA). PCA performs on average 1.15 times better than the FFT method for the Gauteng dataset (i.e. $\mathrm{HD}_{\mathrm{PCA}} \approx 1.15 \times \mathrm{HD}_{\mathrm{FFT}}$).

5. CONCLUSION

We presented a label agnostic feature extraction comparison framework in this paper. We demonstrated its usefulness by employing it and a case study to compare two feature extraction methods, namely FFT and PCA (we also found that the PCA approach outperformed the FFT approach).

6. REFERENCES

[1] T.L. Grobler, *Sequential and non-sequential hypertemporal classification and change detection of MODIS time-series*, Ph.D. thesis, University of Pretoria, 2012.

[2] G.F. Byrne, P.F. Crapper, and K.K. Mayo, "Monitoring land-cover change by principal component analysis of multitemporal Landsat data," *Remote sensing of Environment*, vol. 10, no. 3, pp. 175–184, 1980.

[3] M. Hall-Beyer, "Comparison of single-year and multi-year NDVI time series principal components in cold temperate biomes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 11, pp. 2568–2574, 2003.

[4] T.I.R de Almeida, N.C. Penatti, L. G. Ferreira, A.E. Arantes, and C.H. do Amaral, "Principal component analysis applied to a time series of MODIS images: the spatio-temporal variability of the Pantanal wetland, Brazil," *Wetlands ecology and management*, vol. 23, no. 4, pp. 737–748, 2015.

[5] M.E. Jakubauskas, D.R. Legates, and J.H. Kastens, "Crop identification using harmonic analysis of time-series AVHRR NDVI data," *Computers and electronics in agriculture*, vol. 37, no. 1-3, pp. 127–139, 2002.

[6] SS. Ibrahim, H. Balzter, K. Tansey, N. Tsutsumida, and R. Mathieu, "Estimating fractional cover of plant functional types in African savannah from harmonic analysis of MODIS time-series data," *International Journal of Remote Sensing*, vol. 39, no. 9, pp. 2718–2745, 2018.

[7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] T.E. Oliphant, *A guide to NumPy*, vol. 1, Trelgol Publishing USA, 2006.