A Geometrical Approximation of PCA for Hyperspectral Data Dimensionality Reduction

A. Machidon, R. Coliban, M. Ivanovici, F. Del Frate

prof. Mihai Ivanovici
Universitatea Transilvania din Brașov, Romania

12-16 November 2018
ESA Earth Observation Φ-Week
Frascati, Italy

Paradigm shift

- Principal Component Analysis (PCA) - widely-used statistical tool for multivariate signal analysis
- PCA is model-based, assuming a normal (Gaussian) distribution
- is multi- and hyper-spectral remotely-sensed data really Gaussian?

Context

- PCA directions are given by the majority of the data
- none of the existing PCA approaches (including approximations) can be accelerated by a [fully] parallel implementation
- **we propose a data-driven approach for PCA approximation, which can be implemented in parallel**
- moreover, the PCA directions can be indicated (with a given accuracy) by few points in the data (*outliers*[*])

[*] from the point of view of the Chebyshev inequality for any distribution.

Hypothesis

Observation: the direction given by the furthest points is *relatively close* to the one given by the 1st principal component [†]

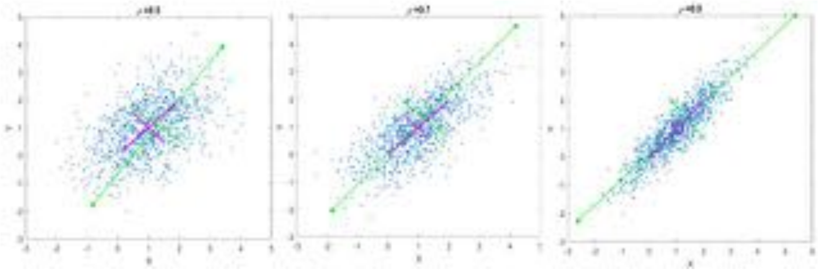


Figure 1: 2D synthetic data (in blue) with standard PCA (in magenta) and approximated PCA (in green) axes for various values of the correlation coefficient ($\rho = 0.5$, 0.7, 0.9).

[†] Machidon, A., Coliban, R., Machidon, O., Ivanovici, M. (2018). Maximum Distance-based PCA Approximation for Hyperspectral Image Analysis and Visualization. In 41st International Conference on Telecommunications and Signal Processing, IEEE, Athens, July 2018.

Approach

1. identify the two elements in a set of n-dimensional vectors $P_1 = \{\mathbf{p_{11}}, \mathbf{p_{12}}, \ldots\} \subset \mathbb{R}^n$ separated by the maximum distance:

$$\{\mathbf{e_{11}}, \mathbf{e_{12}}\} = \arg \max_{\mathbf{p_{1i}}, \mathbf{p_{1j}} \in P1} d(\mathbf{p_{1i}}, \mathbf{p_{1j}}) \tag{1}$$

   The 1st basis vector is $\mathbf{v_1}$ that connects the 2 points:
   $\mathbf{v_1} = \mathbf{e_{11}} - \mathbf{e_{12}}$
2. In order to compute the 2nd basis vector, all the elements in P_1 are projected onto the hyperplane H_1, determined by the normal vector $\mathbf{v_1}$ and containing $\mathbf{m}$ (the midpoint of the segment that connects the two points):

$$H_1 = \{\mathbf{x} \in \mathbb{R}^n | < \mathbf{v_1}, \mathbf{x} > = < \mathbf{v_1}, \mathbf{m} >\} \tag{2}$$

   This results in a set of projections of the original points, $P_2 = \{\mathbf{p_{21}}, \mathbf{p_{22}}, \ldots\}$, computed using the following formula:

$$\mathbf{p_{2i}} = \mathbf{p_{1i}} + (< \mathbf{v_1}, \mathbf{m} > - < \mathbf{v_1}, \mathbf{p_{1i}} >) \cdot \mathbf{v_1} / ||\mathbf{v_1}||^2 \tag{3}$$

The approach illustrated

- After the identification of the two projections separated by the maximum distance $\{\mathbf{e_{21}}, \mathbf{e_{22}}\}$, the second vector, $\mathbf{v_2}$, is computed as the difference between the two values, and so on...

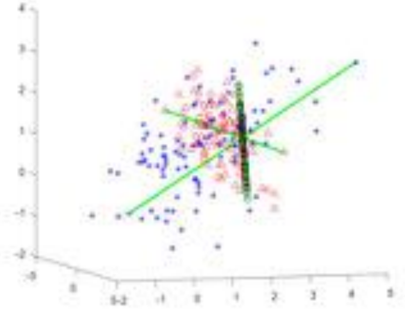


Figure 2: Approximated principal components axes on a 3D correlated cloud of points.

Error metrics

- 100 random sets of 1k 2D points with Gaussian distribution, correlation coefficient from 0.5 to 1, in steps of 0.1.
- Two evaluation metrics:
  1. error angle between the 1st PC and the approximated one
  2. error distance between the mean and the midpoint of the segment given by the furthest points

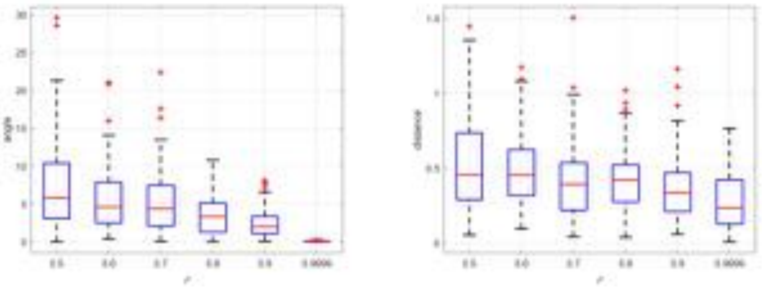


Figure 3: Boxplots of the error angle (left) and error distance (right).

Experimental results - hyper-spectral image visualization

- application: PCA approximation for hyper-spectral image analysis and visualization (Pavia univ. data set).

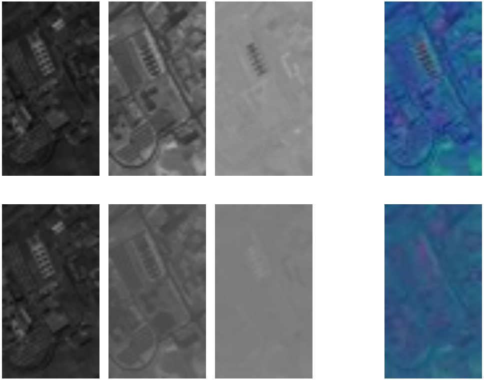


Figure 4: First three computed (top) vs. approximated (bottom) principal components of Pavia image and PCA-based visualization.

Experimental results using INTA-AHS data

- airborne INTA-AHS instrument data set has been acquired during ESA AGRISAR measurement campaign
- test site is the area of Durable Environmental Multidisciplinary Monitoring Information Network (DEMMIN)
  - consolidated test site located in Mecklenburg-Western Pomerania, North-East Germany, covering approx. 25 000 ha
  - fields are very large in this area (in average, 200–250 ha)
  - main crops are wheat, barley, rape, maize, and sugar beet
  - altitudinal range within the test site is around 50 m
- The AHS has 80 spectral channels in the visible, shortwave and thermal IR, pixel size of 5.2 m
- The acquisition taken on the 6 June 2006 has been considered

Input data

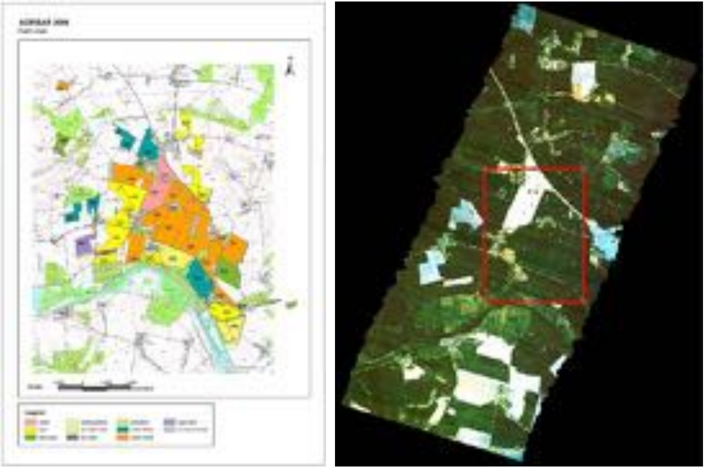


Figure 5: AGRISAR 2006 Ground Truth (left) and AHS crop region (right); Red-Band 6, Green-Band 4, Blue-Band 2.

Experimental setup

- Comparison between classical PCA, approximated (proposed) and non-linear PCA performed by NN[‡]
  - multi-layer perceptron with backpropagation learning
  - topology of NN used: one hidden layer with 20 nodes, 5 input and 6 output nodes
- Only the first 5 PCs were considered (99% of information)
- Learning was performed using the Neumapper sw application provided by EO Laboratory at Univ. of Rome “Tor-Vergata”

[‡] Giorgio A Licciardi and Fabio Del Frate. Pixel unmixing in hyperspectral data by means of neural networks. IEEE Transactions on Geoscience and Remote Sensing, 49(11):4163–4172, 2011.

PCs - computed, approximated and learned


(a) PC1


(b) PC2


(c) PC3


(d) PC4


(e) PC5


(f) Apprx1


(g) Apprx2


(h) Apprx3


(i) Apprx4


(j) Apprx5


(k) Nonln1

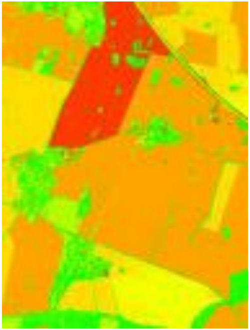

(l) Nonln2


(m) Nonln3

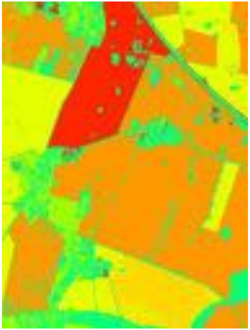

(n) Nonln4


(o) Nonln5

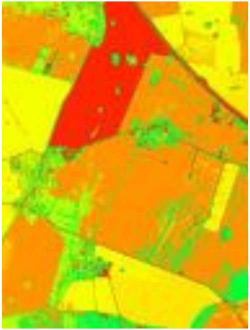
Classification - qualitative comparison



(a) Computed PCA



(b) Approx. PCA



(c) Nonlinear PCA

Figure 7: Classified images based on computed PCA (left), approximated PCA (middle) and non-linear PCA (rigt) of the DEMMIN test site of AHS image.

Classification - quantitative comparison

Class	Description	PCA	Approx.PCA	Nonlin. PCA
1	Rape	100%	100%	100%
2	Set aside: rape	50%	50%	50%
3	Maize	100%	100%	100%
4	Winter wheat	**100%**	90.5%	95%
5	Cutting pasture	66.7%	66.7%	66.7%
6	Grassland	66.7%	**100%**	66.7%
7	Winter barley	-	-	-
8	Urban	75%	75%	75%

Table 1: Classification accuracy (true positives) for 40 randomly-chosen pixels.

Conclusions and future work

- Comparable results in terms of classification for the classical, approximated and learned PCA, however a more thorough analysis is required
- Comparison showed questionable results locally
- Improve the protocol of validation by fully automatic check within the ground truth
- Parallel implementation of the PCA approximation on GPUs using CUDA
- Demonstrate the capabilities of the PCA approximation parallel implementation in a satellite Big Data scenario