On the use of the Hotelling’s T2 statistic for the hierarchical clustering of hyperspectral data

Miguel Angel Veganzones, Joana Frontera-Pons, Jocelyn Chanussot, Jean-Philippe Ovarlez

To cite this version:

Miguel Angel Veganzones, Joana Frontera-Pons, Jocelyn Chanussot, Jean-Philippe Ovarlez. On the use of the Hotelling’s T2 statistic for the hierarchical clustering of hyperspectral data. WHISPERS 2013 - 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Jun 2013, Gainesville, FLO, United States. pp.n/c, 10.1109/WHISPERS.2013.8080647. hal-01010382

HAL Id: hal-01010382
https://hal.science/hal-01010382
Submitted on 19 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ON THE USE OF THE HOTELLING’S $T^2$ STATISTIC FOR THE HIERARCHICAL CLUSTERING OF HYPERSONTRAL DATA

M.A. Veganzones$^1$, J. Frontera-Pons$^2$, J. Chanussot$^1$, J.P. Ovarlez$^3$

$^1$GIPSA-lab, Grenoble-INP, Saint Martin d’Hères, France
$^2$ SONDRA Research Alliance, Supélec, France
$^3$French Aerospace Lab, ONERA DEMR/TSI, France

ABSTRACT

In this work we propose a hierarchical clustering methodology for hyperspectral data based on the Hotelling’s $T^2$ statistic. For each hyperspectral sample data, the statistical sample mean is calculated using a window-based neighborhood. Then, the pairwise similarities between any two hyperspectral samples are computed based on the Hotelling’s $T^2$ statistic. This statistic assumes a Gaussian distribution of the data while hyperspectral data have been observed to be long tailed distributed. In order to improve the statistic robustness we use the Fixed Point estimates, and compare them to the classical sample mean estimator. The similarities are then used to hierarchically cluster the hyperspectral data. We give some preliminary qualitative results of the proposed approach over the Indian Pines hyperspectral scene. Results show that the use of the Fixed Point estimator does not significantly affect the clustering results. Further work will be focused on the use of the robust Hotelling statistic.

Index Terms— hyperspectral imaging, hierarchical clustering, Fixed Point estimates

1. INTRODUCTION

Hyperspectral data have been observed not to be multivariate normal but long tailed distributed [1]. In order to take into account these features, the class of elliptical contoured distributions (ECD) is considered to describe clutter statistical behavior. It provides a multivariate location-scatter family of distributions that primarily serves as heavy tailed alternative to the multivariate normal model. A $m$-dimensional random vector $y = [y_1, y_2, \ldots, y_m]^T$ with mean $\mu$ and scatter matrix $\Sigma$ has an elliptical distribution if its probability density function (PDF) has the following form:

$$f_y(y) = |\Sigma|^{-1} h_m((y - \mu)^T \Sigma^{-1}(y - \mu)),$$  \hspace{1cm} (1)

where $T$ denotes the transpose operator and $h_m(.)$ is any function such as (1) defines a PDF in $\mathbb{R}^m$.

The ECD class includes a large number of distributions, notably the Gaussian distribution, multivariate $t$ distribution, $K$-distribution or multivariate Cauchy. Thus it allows for heterogeneity of the background power with the texture. Although non-Gaussian models are assumed for background modelling and test design, the parameters estimation is still performed using classical Gaussian based estimators; as for the covariance matrix, it is generally determined according to the Sample Covariance Matrix (SCM), $\hat{\Sigma}_{SCM} = \frac{1}{N} \sum_{i=1}^{N} y_i y_i^H$. We use here the Fixed Point (FP) estimators [2] as a robust alternative for background statistical characterization.

In [3], authors propose a general approach for high-resolution polarimetric SAR (POLSAR) data classification in heterogeneous clutter, based on a statistical test of equality of covariance matrices. In hyperspectral domain the information provided by the mean is of high relevance. Thus, we propose here a similar approach for hyperspectral data using the Hotelling’s $T^2$ statistic as a statistical test of equality of means [4]. We define a pairwise similarity between two pixels by using the output of the Hotelling’s test for the means of the pixels calculated using a sliding window. As a preliminary test, we segment the hyperspectral scene by means of a hierarchical clustering algorithm [5] whose input is given by the previously calculated pairwise distances. We test the proposed approach in the Indian Pines hyperspectral scene.

2. THE FIXED POINT ESTIMATES

The Fixed Point estimates proposed by Tyler in [2] belong to the wider class of robust $M$-estimates [6]. According to the FP approach, the joint estimation of the scatter matrix, $\hat{\Sigma}$, and the mean vector, $\hat{\mu}$, leads to [2]:

$$\hat{\Sigma}_{FP} = \frac{m}{N} \sum_{n=1}^{N} (y_n - \hat{\mu})(y_n - \hat{\mu})^H$$

$$\hat{\mu}_{FP} = \frac{m}{N} \sum_{n=1}^{N} \frac{(y_n - \hat{\mu})(y_n - \hat{\mu})^H}{\hat{\Sigma}_{FP}}$$  \hspace{1cm} (2)
and
\[
\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} \left( (y_n - \hat{\mu})^T \hat{M}_{FP}^{-1} (y_n - \hat{\mu}) \right)^{1/2}
\]
where \( N \) denotes the number of secondary data and \( y_n \) the vector under observation. For the matrix estimate, existence and uniqueness have been established in [7]. Although the proof for simultaneous scatter and location estimates is still an open question, they have been found to be useful and reliable for the estimation of ECD parameters because of its easy implementation. They are specified by implicit equations and can be easily computed using a recursive algorithm. We refer to [8] for a detailed performance analysis of the FP covariance matrix estimate. The main results of the statistical properties of the \( \hat{M}_{FP} \) are summarized: \( \hat{M}_{FP} \) is a consistent and unbiased estimate of \( M \); its asymptotic distribution is Gaussian and is the same as the asymptotic distribution of a Wishart matrix with \( mN/(m+1) \) degrees of freedom. They provide a robust alternative to the classical estimators and its definition on hyperspectral detection framework was introduced in [9].

3. HOTELLING’S \( T^2 \) HYPOTHESIS TESTS

Let \( X \sim \text{ECD}(\mu_1, \Sigma, h_{m,1}) \) and \( Y \sim \text{ECD}(\mu_2, \Sigma, h_{m,2}) \) be two independent random vectors, elliptically distributed. We intend to decide if they belong to the same class comparing their mean vectors \( \mu_1 \) and \( \mu_2 \). The classification problem can be formulated as a binary hypothesis test. Symbolically, we aim to distinguish between:

\[
\begin{cases}
H_0 : \mu_1 = \mu_2 \\
H_1 : \mu_1 \neq \mu_2
\end{cases}
\]

where the decision on the hypothesis test is made according to some statistical based criteria. We propose a similarity function based on the parameters estimation.

**Hotelling’s \( T^2 \) statistic.** Assuming \( x_1, \ldots, x_{N_1} \) and \( y_1, \ldots, y_{N_2} \) are i.i.d. samples from \( X \sim \mathcal{N}_m(\mu_1, \Sigma) \) and \( Y \sim \mathcal{N}_m(\mu_2, \Sigma) \) respectively. And the mean vector computed as, \( \hat{\mu}_x = \frac{1}{N_1} \sum_{i=1}^{N_1} x_i \) and \( \hat{\mu}_y = \frac{1}{N_2} \sum_{i=1}^{N_2} y_i \) and the covariance matrix estimated according to:

\[
W = \frac{\sum_{i=1}^{N_1} (x_i - \hat{\mu}_x)(x_i - \hat{\mu}_x)^T + \sum_{i=1}^{N_2} (y_i - \hat{\mu}_y)(y_i - \hat{\mu}_y)^T}{N_1 + N_2 - 2}
\]

which correspond to the unbiased pool estimate. Then the Hotelling \( T^2 \) statistic is defined as:

\[
t^2 = \frac{N_1 N_2}{N_1 + N_2} (\hat{\mu}_x - \hat{\mu}_y)^T W^{-1} (\hat{\mu}_x - \hat{\mu}_y). \tag{4}
\]

This expression is easily derived from the likelihood function of a multivariate normal model when \( H_0 \) is assumed, see e.g. [4]. The distribution of the test under the null hypothesis can be related to the \( F \)-distribution according to

\[
\frac{N_1 + N_2 - m - 1}{(N_1 + N_2 - 2)m} t^2 \sim F(m, N_1 + N_2 - 1 - m).
\]

A different situation arises when samples do not fit on a multivariate gaussian distribution. We can improve the robustness of the test replacing the mean and matrix estimates for some robust alternative (e.g. the FP) if its asymptotic distribution is gaussian. The equality of the covariances matrices is a simplification that will certainly modify the test, as the resulting matrix estimate will not correspond to a Wishart distribution \( (1/N_1 \Sigma_1 + 1/N_2 \Sigma_2) \). The results found on statistics literature for robust Hotelling and for both matrix and mean comparison will be further investigated.

4. STATISTICAL HIERARCHICAL CLUSTERING

A hierarchical clustering algorithm returns a hierarchy of clusters built by merging smaller components into bigger clusters (agglomerative clustering) or by splitting the whole image into smaller regions (divisive clustering). In order to do that, a pairwise similarity function is necessary, i.e. the angular distance, to compare each pair of components. Then, a linkage function indicates which of any possible pair of components is merged to (or split from) a bigger cluster.

A common linkage is the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method which consists in averaging all distances between the samples in each node, that is, the mean distance between elements of each cluster. Given two nodes \( A \) and \( B \), the UPGMA linkage distance between the nodes, denoted \( l(A, B) \) is given by:

\[
l(A, B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \tag{5}
\]

where \(|\cdot|\) denotes the cardinality of a set and \( d(x, y) \) is any pairwise distance. In this work we propose to use the Hotelling’s \( T^2 \) statistic (4) as the pairwise distance. The result is a hierarchy of clusters that can be represented as a dendrogram. A common criterion to stop merging (splitting) is to a priori set the number \( c \) of clusters one is looking for.

5. EXPERIMENTS AND RESULTS

5.1. Indian Pines scene

The Indian Pines scene\(^1\) was gathered by airborne AVIRIS sensor over North-western Indiana and consists of 145 \times 145 pixels and 224 spectral reflectance bands in the wavelength range 0.4-2.5 \( \mu m \). We have reduced the number of bands to 200 by removing bands covering the region of water absorption: [104-108], [150-163], 220. Since the scene is taken in June some of the crops present, corn, soybeans, are in early stages of growth with less than 5% coverage. The ground

\(^1\)https://engineering.purdue.edu/~biehl/MultiSpec/
We have presented an approach to statistically characterize the hyperspectral data and then use the Hotelling’s $T^2$ statistic to define a pairwise similarity function. The definition of a pairwise similarity allows to use a hierarchical clustering algorithm to segment the hyperspectral image. We have shown that even when the Hotelling’s statistic Gaussian assumption differs from the actual ECD, the clustering results are meaningful. The use of a robust FP estimator shows a better performance than the use of the classic estimator. The use of a robust Hotelling’s estimator capable of dealing with non-Gaussian assumption will be further investigated.

6. CONCLUSIONS

Acknowlegements

M.A. Veganzones was supported by the Delegation Generale de l’Armement (DGA) under contract PDE2012600079.

7. REFERENCES


Fig. 2. Results of the hierarchical clustering over the entire dataset using the Hotelling $T^2$ test with the classic estimator (left column) and the Fixed Point estimator (right column). Each row identifies with a different value of the number of clusters, $c$. 
Fig. 3. Results of the hierarchical clustering over the reduced dataset using the Hotelling $T^2$ test with the classic estimator (left column) and the Fixed Point estimator (right column). Each row identifies with a different value of the number of clusters, $c$. 