Tests for the elliptical symmetry of hyperspectral imaging data

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1117/12.861055">http://dx.doi.org/10.1117/12.861055</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>SPIE</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Thu Jan 24 12:09:11 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/60938">http://hdl.handle.net/1721.1/60938</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Tests for the Elliptical Symmetry of Hyperspectral Imaging Data

Sidi Niu\textsuperscript{a}, Vinay K. Ingle\textsuperscript{a}, Dimitris G. Manolakis\textsuperscript{b} and Thomas W. Cooley\textsuperscript{c}

\textsuperscript{a}Northeastern University, 360 Huntington Avenue, Boston, MA 02115; \\
\textsuperscript{b}MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420; \\
\textsuperscript{c}Air Force Research Laboratory, 3550 Aberdeen Ave SE, Kirtland AFB, MN 87117.

ABSTRACT

Accurate statistical models for hyperspectral imaging (HSI) data distribution are useful for many applications. A family of elliptically contoured distribution (ECD) has been investigated to model the unimodal ground cover classes. In this paper we propose to test the elliptical symmetry of real unimodal HSI clutters which will answer the question whether the family of ECD will provide an appropriate model for HSI data. We emphasize that the elliptical symmetry is an inherent feature shared by all ECDs. It is a prerequisite that real HSI clutters must pass these elliptical symmetry tests, so that the family of ECD can be qualified to model these data accurately.

Keywords: hyperspectral imaging, modeling, elliptical symmetry testing

1. INTRODUCTION

Accurate statistical models for HSI data distribution are useful for many applications. These models provide the foundation for development and evaluation of reliable algorithms for detection, classification, clustering, and estimation. A typical hyperspectral image usually contains several different ground cover classes and consequently can be modeled by a finite mixture statistical model. Each component density function of the mixture model is supposed to characterize a single unimodal ground clutter. A family of ECD has been investigated to model the unimodal ground cover classes. Regardless of the mean vector, an ECD can be determined by the scale/covariance matrix and its generating function which are assumed to be independent from each other. Since the covariance matrix can be estimated from the sample covariance matrix (sometimes regularized sample covariance matrix), the only flexibility left is finding an appropriate generating function to characterize real HSI data behaviors. Previous work has focused on fitting a specific ECD in the Mahalanobis distance sense to capture the heavy tail behavior of real HSI data. In this paper we stress on the other issue that testing the elliptical symmetry of real unimodal HSI clutters will answer the question whether the family of ECD will inherently provide an appropriate model for HSI data. Several graphical testing schemes are proposed. Instead of testing elliptical symmetry directly, HSI data are first prewhitened and then checked for their spherical symmetry. A $t$-plot and a $\beta$-plot correlates the spherical symmetry to graphical linearity. Pairwise scatter plots and angular goodness-of-fit test evaluate the symmetry behavior in polar coordinates. Numerical measures are also calculated in each case as auxiliaries to graphical plots. We emphasize that the elliptical symmetry is an inherent feature shared by all ECDs. It is a prerequisite that real HSI clutters must pass these elliptical symmetry tests, so that the family of ECD can qualify to model these data accurately.

The organization of this paper is as follows. In Section 2, we first discuss the mathematical background of statistical modeling as well as definition and properties of ECDs. In Section 3, the proposed graphical elliptical symmetry tests derived from the unique properties of ECDs are discussed and illustrated by synthetic data. These tests are then applied to real clustered HSI data in Section 4. The main results and conclusions are summarized in Section 5.

Further author information: (Send correspondence to Sidi Niu)
Sidi Niu: E-mail: niusidi@gmail.com, Telephone: 1 617 710 0991
2. MATHEMATICAL BACKGROUND

2.1 Statistical modeling

A typical HSI image of remote sensing, for example a down-looking airborne image, usually includes large area and thus contains different ground covers, called clutters. Hence the whole image has unimodal distribution which is hard to represent. We assume that the whole image can be well classified into finite number of distributional unimodal clutters. Each clutter can thus be modeled separately. The multimodal probability density function (pdf) representing the whole image can be obtained by the finite mixture model

\[ f(x; \Psi) = \sum_{i=1}^{K} \pi_i f_p(x; \Theta_i) \]  

where \( \Psi \) is the set including all parameters, \( \Psi = \{\pi_1, \ldots, \pi_K, \Theta_1, \ldots, \Theta_K\} \). The mixture weights \( \pi_i \), or priors, satisfy the constraints, \( 0 \leq \pi_i \leq 1 \) and \( \sum_{i=1}^{K} \pi_i = 1 \), and \( f_p(x; \Theta_i) \) is a single unimodal pdf characterized by parameter \( \Theta_i \) with a dimensionality of \( p \). The finite mixture model describes the whole hyperspectral image as a mixture of pixels from \( K \) different spectrally homogeneous distributions, with the probability of a pixel belonging to one of the distributions equals to its weight \( \pi_i \). In this paper, the constitutional pdfs \( f_p(x; \Theta_i) \) are assumed to be ECDs, that is \( f_p(x; \Theta_i) = \text{ECD}_p(x; \mu, \Sigma, q) \). The graphical symmetry tests proposed in Section 3 will attempt to verify this ECD assumption. We first discuss the definition as well as some key properties of ECDs.

2.2 Elliptically Contoured Distribution

A random vector \( x = [x_1, \ldots, x_p]^T \), with dimensionality \( p \), is said to have distribution belonging to the family of elliptically contoured distribution, denoted by \( x \sim \text{ECD}_p(x; \mu, \Sigma, q) \) if its pdf, when exists, can be expressed as a function of the quadratic form \( (x - \mu)^T \Sigma^{-1}(x - \mu) \) and is given by

\[ f_x(x) = (2\pi)^{-\frac{p}{2}} (\det \Sigma)^{-\frac{1}{2}} q((x - \mu)^T \Sigma^{-1}(x - \mu)), \]  

where \( \mu = [\mu_1, \ldots, \mu_p]^T \) is the location parameter, \( \Sigma \) is the scale matrix (symmetric and positive definite), and \( q(\cdot) \) is called generating function which categorizes subset distributions within the family. Since the location parameter can be set to be zero by changing the origin, an ECD can be fully characterized by its scale matrix \( \Sigma \) and generating function \( q(\cdot) \) without loss of generality.

The ECD family is a generalization of the widely-used normal or Gaussian distribution by introducing extra degrees of freedom. In fact, a representation theorem exists for all ECDs which is summarized below.

**Theorem 1.** If a random vector is an ECD random vector, then there exists a nonnegative random variable \( s \) such that the pdf of the random vector conditioned on \( s \) is a multivariate normal pdf.

Given the representation theorem, any ECD can be generated by modulation of normal distribution. Let \( z = [z_1, z_2, \ldots, z_p]^T \) denote a real, zero mean, normal random vector with covariance matrix \( C \). Let \( s \) denote a nonnegative random variable with pdf \( f_s(s) \), called characteristic pdf. Consider the product defined by \( x = s z \). For the problem of background clutter modeling and simulation, it is desirable to independently control the non-Gaussian clutter envelop pdf and its correlation properties. Therefore, \( z \) and \( s \) are assumed to be statistically independent. The generating function \( q(\cdot) \) is solely determined by the characteristic pdf \( f_s(s) \), while the scale matrix \( \Sigma \) is related to the covariance matrix \( C \) by a positive scaler \( \alpha > 0 \), that is \( \Sigma = \alpha C \). Hence, the scale matrix \( \Sigma \) and the generating function \( q(\cdot) \) characterizes a particular ECD independently. We emphasize that the correlation property given by the scale matrix \( \Sigma \) does not categorize which subclass of an ECD belongs to, it is the generating function \( q(\cdot) \) which discriminates different subclasses within the ECD family.

Since the same scale matrix \( \Sigma \) can be shared by any ECD, in order to unveil the symmetry property, without generality, a whitening process is applied to an ECD. Correspondingly, a spherical distribution is obtained with the same generating function but with zero mean and identity matrix as its covariance matrix. In this case, each marginal distribution is an identical independent distribution (i.i.d.). Important insights can be gained if generalized spherical coordinate representation is applied.
Theorem 2. A random vector \( \mathbf{z} = [z_1, z_2, \ldots, z_p]^T \), which has zero-mean and identity covariance matrix, is ECD if and only if there exist random variables \( r \in (0, \infty), \theta_k \in (0, \pi), 1 \leq k \leq p-2 \) and \( \theta_{p-1} \in (-\pi, \pi) \) such that when the components of \( \mathbf{z} \) are expressed in the generalized spherical coordinates:\(^4\)

\[
\begin{align*}
  r &= \sqrt{z_1^2 + z_2^2 + \ldots + z_p^2}, \\
  \cos \theta_1 &= z_1/r, \quad 0 < \theta_1 < \pi, \\
  \cos \theta_k &= \frac{z_k}{z_{k-1}} \tan \theta_{k-1}, \quad 0 < \theta_k < \pi, \quad 2 \leq k \leq p-2, \\
  \tan \theta_{p-1} &= z_p/z_{p-1}, \quad -\pi < \theta_{p-1} < \pi.
\end{align*}
\]

then the random variables \( r \) and \( \theta_k, 1 \leq k \leq p-1 \) are mutually statistically independent with pdfs of the form:

\[
\begin{align*}
  f(r) &= \frac{r^{p-1}}{2^{(p/2)-1}\Gamma(p/2)} q(r^2), \\
  f_k(\theta_k) &= \frac{\Gamma((p-k+1)/2)}{\sqrt{\pi}\Gamma((p-k)/2)} \sin^{p-1-k}(\theta_k), \quad 0 < \theta_k < \pi, \quad 1 \leq k \leq p-2, \\
  f_{p-1}(\theta_{p-1}) &= (2\pi)^{-1}, \quad -\pi < \theta_{p-1} < \pi,
\end{align*}
\]

where \( q(\cdot) \) is the same generating function as in Eq. (2) and \( \Gamma(v) \) is the Euler’s Gamma function.

In the spherical coordinates representation, the random variable \( r \), say the envelop, contains all the distinctive information characterizing individual ECD member through the generating function \( q(\cdot) \), while distributions in the orthogonal subspace spanned by \( \theta_1, \ldots, \theta_{p-1} \) are totally fixed and shared by the whole ECD family. Consequently, fitting a specified ECD to a particular HSI clutter degrades to a one dimensional problem in the Mahalanobis square distance \( r^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \).\(^5\) We use the same notation \( r^2 \) because this distance equals the Euclidean square distance in spherical coordinates \( r^2 \). To answer the question whether the ECD family will inherently provide a satisfying model for HSI clutters, we need to compare the empirical distributions with the theoretical ones in the angular subspace spanned by \( \theta_1, \ldots, \theta_{p-1} \).

3. ELLIPTICAL SYMMETRY TESTS

In this section, we introduce several graphical symmetry testing techniques, all of which are based on whitened spherical distribution mentioned previously.

3.1 Scatter Plot

The primary way to verify a ECD distribution is to use a scatter plot\(^5\) which arises from the spherical coordinates representation. From Theorem 2, whitened ECDs in spherical coordinates share the same angles distributions and are only identified by unique envelops. We use \( p \times p \) matrix of plots in which off-diagonal scatter plots investigate the independence of pairwise coordinates and diagonal histogram plots illustrate the marginal distributions. Figure 1a is an illustrative scatter plot for synthetic data generated from \( \mathbf{z} \sim t_5(\mathbf{0}, \mathbf{I}) \), where \( t_5(\mathbf{0}, \mathbf{I}) \) denotes a multivariate \( t \) distribution of dimensionality 5 with zero mean and identity matrix as its covariance matrix, which is known as a member of ECD family. We are most interested here in the off-diagonal scatter plots. A scatter plot in which the points are distributed evenly indicates that the two correspondent random variables are uncorrelated which is a looser condition for independence that is hard to verify.

3.2 Angular Goodness-of-fit Plot

In Theorem 2, the pdfs of the angle random variables are explicitly derived which remain unchanged as long as spherically contoured condition is met. As a supplemental test of diagonal histograms in scatter plot, quantile-quantile (Q-Q) plots of theoretical cumulative distribution function (cdf) and empirical cdf are developed. Such a procedure is valid in exploiting quadratic statistics\(^6\) and provides a measure of the difference between the theoretical cdf and the empirical cdf. As shown in Figure 1b, the empirical cdf plots of synthetic data drawn from symmetric distribution \( \mathbf{z} \sim t_5(\mathbf{0}, \mathbf{I}) \) overlap the theoretical ones perfectly. One thing to mention here is that all synthetic data from a particular member of ECD family will theoretically exhibit the exact same results. Three goodness-of-fit metrics\(^7\) are computed as numerical measurements supplemental to the graphical plots.
3.3 \( t \)-plot

The scatter plots and angular goodness-of-fit plots examine symmetry behavior in the whole angular subspace. There are also other techniques, for example some robust statistics, which test whether the data is symmetric in certain direction. We first begin with the \( t \)-statistic which tests the spherical symmetry in the \( t \) direction\(^8\) . Let \( z = [z_1, \ldots, z_p]^T \) be the same whitened random vector. It is well known\(^9\) that the statistic

\[
y_k = t(z_k) = \frac{\sqrt{p} \bar{z}_k}{s_k}
\]

where \( z_k = [z_{k1}, \ldots, z_{kp}]^T \), \( \bar{z}_k = (1/p) \sum_{i=1}^p z_{ki} \), and \( s_k^2 = (1/p) \sum_{i=1}^p (z_{ki} - \bar{z}_k)^2 \). Obviously, \( y_1, \ldots, y_N \) are i.i.d. according to \( t_{p-1} \), where \( N \) is the total number of samples. We plot ascending ordered \( y(k) \) against the \( \frac{2k-1}{2N} \)-quantile of \( t_{p-1} \), \( k = 1, \ldots, N \). If the data distribution is spherically symmetric, then the plot will cling to the 45° line through the origin. Both outliers and presence of systematic differences between \( y(k) \) and the \( \frac{2k-1}{2N} \)-quantile of \( t_{p-1} \) can be detected by visually comparing the plots to the 45° line. Deviation from linearity in the \( t \)-plot indicates deviation of the underlying distribution from elliptical distributions. As a numerical measurement of symmetric behavior auxiliary to the graphical plot, we also calculate the correlation coefficient between ordered empirical \( y(k) \) and theoretical ordered univariate \( t \) samples. The value of 1 indicates a perfectly linear plot. An illustrating example of the \( t \)-plot can be found in Figure 2a where the symmetric synthetic data from \( t_5(0, I) \) is tested. From the plot, we can see that most of the points fall on the reference line except for few outlier ones.

3.4 \( \beta \)-plot

Similar to the \( t \)-plot, the \( \beta \)-plot employs the calculation of robust \( \beta \)-statistics, which tests the symmetric behavior in the \( \beta \) direction. Let \( z = [z_1, \ldots, z_p]^T \) be a random vector with zero mean and identity covariance matrix, then the \( \beta \)-statistics is computed by

\[
b_{alpha}(z) = \frac{\sum_{i=1}^a z_i^2}{\sum_{i=1}^p z_i^2}, \quad a = 1, 2, \ldots, p
\]

It can be shown that \( b_{alpha}(z) \) is distributed according to beta distribution with parameters \( a/2 \) and \( (p - a)/2 \), denoted by Beta\((a/2, (p - a)/2)\). Generally, a good choice for \( a \) is \( a = \lceil p/2 \rceil \), where \( \lceil a \rceil \) denotes the largest integer
Theoretical Quantiles
Empirical Quantiles

Figure 2: Testing results for synthetic data drawn from $z \sim t_5(0, I)$. (a) $t$-plot employs the evaluation of robust $t$-statistics. Symmetry behavior is tested by the degree of alignment between empirical line and reference $45^\circ$ line. (b) $\beta$-plot employs the evaluation of robust $\beta$-statistics. Symmetry behavior is tested by the degree of alignment between empirical line and reference $45^\circ$ line.

which is no more than $a$. Similar to the $t$-plot, we can plot a $\beta$-probability Q-Q plot by assigning the ordered empirical $\beta$-statistics $b_{a,p}(z_k)$, $k = 1, \ldots, N$, where $N$ is the total number of samples, to the correspondent $k$th quantile of theoretical distribution Beta($a/2, (p-a)/2$). We also compute the correlation coefficient as an auxiliary numerical measurement. An illustrating example of the $\beta$ plot is shown in Figure 2b. In this figure, symmetric synthetic data from $t_5(0, I)$ is tested and as seen in the plot, the points fall on the reference line perfectly.

4. EXPERIMENTAL RESULTS

In this section, we now apply the derived graphical symmetry tests to real HSI data. The data set used for this project, shown in Figure 3a, is from an Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data measurement over Fort A. P. Hill, Virginia, on 9 November 1999 and 26 September 2001 on a Twin Otter aircraft flying at an altitude of approximately 12,000 ft. Fort A. P. Hill, located in eastern Virginia, is heavily forested with natural areas of deciduous, coniferous, and mixed deciduous-coniferous forests. There are also numerous large plantations of loblolly pine that are intended for harvest as pulp wood. The trees in the loblolly pine plantations were set in regular rows at close spacing to maximize yield and planted at the same time. The result is a forest land cover that is spectrally and spatially uniform. By contrast, the natural forests show much greater spectral and spatial variability. The forest land covers have been mapped by Fort personnel for resource management and conservation purposes and are contained in a Geographic Information System (GIS) database.

Because the clustering by human experts are raw and without any statistical assumption, we also employ automatic clustering method using iterative Expectation-Maximization (EM) to cluster the data pixels under normal mixture model assumption. The unimodal data sets we tested here are comprised only by the pixels labeled as the same group by automatic method within the human predefined ground cover regions. We will show the test results for Data set 2 Coniferous forest, Data set 3 Deciduous forest, and Data set 6 South panel field in this paper.

Figure 4 and 7 show the test results for Data set 2 Coniferous forest. The scatter plot in Figure 4a shows the points are distributed evenly so that the uncorrelated condition is met. Only slight deviation can be found in Figure 4b which means the empirical cdfs are quite similar to the prescribed theoretical ones. Both $t$-plot shown in Figure 7a and $\beta$-plot shown in Figure 7b exhibit good linearity. The test results verify that the Coniferous forest data exhibit strong symmetry behavior and thus can be modeled by ECD accurately. Figure 5 and 8 show
Figure 3: (a) Data set used in this project. The area is heavily forested with natural stands of deciduous, coniferous, and mixed deciduous-coniferous forests. There are also numerous large plantations of loblolly pine that are intended for harvest as pulp wood. The whole image has been mapped for different forest land covers by human experts. (b) Automatic generated clutters using expectation-maximization method under normal mixture assumption.

the test results for Data set 6 South panel field. The test result is similar to Coniferous forest so that South panel field data is also distributed symmetrically.

Figure 6 and 9 show the test results for Data set 3 Deciduous forest. The scatter plot in Figure 6a shows some particular shape of the points distribution which means there are some correlation between the corresponding random variables. In Figure 6b, some explicit deviation can be seen in the angular Q-Q plots but in the overall the empirical cdfs are similar to the prescribed theoretical ones. Both in Figure 9a t-plot and in Figure 9b, we can see some points especially outlier ones deviate explicitly from the reference lines. If we look back at Figure 3b, the area of deciduous forest is more complex and contains different ground covers. In such case, pixels in this area would probably include different kinds of trees which may explain the degradation in symmetry behavior.

5. CONCLUSIONS

In summary, we developed three graphical tests to examine the symmetry behavior of unimodal data sets. According to the experimental results, most of the resulting clutters exhibit strong symmetric behavior which demonstrates that ECD based statistical model can fit HSI data quite accurately. The physical qualification of ECD modeling HSI data is discussed by Sangston.11 In order to model heavy tail behavior of HSI data, other members of ECD family can be employed rather than normal distribution.5,12 The symmetry performance of the data depends on the quality of classification. In this work, the performance is improved by applying iterative clustering methods.

REFERENCES

Figure 4: Testing results for coniferous forest. (a) scatter plot shows the points are distributed evenly so that the uncorrelated condition is met; (b) only slight deviation is visible in the angular Q-Q plots which means the empirical cdfs are quite similar to the prescribed theoretical ones.

Figure 5: Testing results for south panel field. (a) scatter plot shows the points are distributed evenly so that the uncorrelated condition is met; (b) only slight deviation is visible in the angular Q-Q plots which means the empirical cdfs are quite similar to the prescribed theoretical ones.
Figure 6: Testing results for deciduous forest. (a) scatter plot shows some particular shape of the points distribution which means there are some correlation between the corresponding random variables; (b) some explicit deviation can be seen in the angular Q-Q plots but in the overall the empirical cdfs are similar to the prescribed theoretical ones.

Figure 7: Testing results for coniferous forest. (a) $t$-plot exhibits only slight deviation from perfect linearity; (b) $\beta$-plot is also linear except for few outlier points.
Figure 8: Testing results for south panel field. (a) $t$-plot exhibits only slight deviation from perfect linearity; (b) $\beta$-plot is also linear except for few outlier points.

Figure 9: Testing results for deciduous forest. (a) $t$-plot shows some explicit deviation from linearity; (b) $\beta$-plot shows many points especially outlier ones deviate from the reference line.


