Stochastic Contrast Measures for SAR Data: A Survey

Alejandro C. Frery*

Abstract: “Contrast” is an generic denomination for “difference”. Measures of contrast are a powerful tool in image processing and analysis, e.g., in denoising, edge detection, segmentation, classification, parameter estimation, change detection, and feature selection. We present a survey on techniques that aim at measuring the contrast between (i) samples of SAR imagery, and (ii) samples and models, with emphasis on those that employ the statistical properties of the data.

Key words: Contrast; Divergence; Entropy; Geodesic distance; Statistics; Synthetic Aperture Radar (SAR)

CLC index: TN957.7

Document code: A

Article number: 2095-283X(2019)06-0758-24

DOI: 10.12000/JR19108


1 Introduction

Synthetic Aperture Radar (SAR) has been widely used as an important system for information extraction in remote sensing applications. Such microwave active sensors have as main advantages the following features: (i) their operation does not depend on sunlight, either of weather conditions and (ii) they are capable of providing high spatial image resolution.

In recent years, the interest in understanding such type of imagery has increased. However, since the acquired images stem from a coherent illumination process, they are affected by a signal-dependent granular noise called “speckle”[1]. Such noise has a multiplicative nature, and its intensity does not follow the Gaussian law. Moreover, the availability of fully polarimetric images requires processing data in the form of complex matrices. Thus, analyzing SAR images requires tailored image processing based on the statistical properties of speckled data.

There is a vast literature on techniques for SAR image processing and analysis that employ a diversity of approaches for the related problems these images pose. Among them, advances in modeling with Statistical Information Theory and Geometry have led to several exciting solutions for the problems of denoising, edge detection, segmentation, classification, parameter estimation, change detection, and feature selection. Moreover, these approaches provide unique insights on the properties of the data and the information they convey.

This paper presents a survey of those results.

This work is organized in nine sections following this Introduction.

Section 2, where we show that it is possible to assess the quality of despeckling filters by measuring the distance of the denoised image to the ideal model.

Once we see that the approach provides interesting results, we move to Section 3, where we review the main inference techniques, namely analogy (Sec. 3.1), maximum likelihood (Sec. 3.2), and Nonparametric and Robust estimation (Sec. 3.3).

Section 4 discusses the main approaches for computing differences between distributions: the logarithm of the ratio of likelihoods (Sec. 4.1), tests based on divergences (Sec. 4.2), tests that rely on the quadratic difference of entropies (Sec. 4.3), and tests that employ geodesic distances between distributions (Sec. 4.4).

Section 5 reviews the main models that describe SAR data in intensity format, and provides details about the Gamma (Sec. 5.1) and G0 (Sec. 5.2) distributions.

Section 6 presents applications of the con-
cepts discussed in Section 4 to the intensity models of the previous one: edge detection by maximum likelihood (Sec. 6.2) and by maximum geodesic distance between samples (Sec. 6.3); parameter estimation by distance minimization (Sec. 6.4); and speckle reduction (Sec. 6.5).

Section 7 reviews the main distribution for fully polarimetric data: the Wishart law. We present the reduced log-likelihood of a sample from such a model and estimation by maximum likelihood.

In Section 8, we study applications of measures of difference to polarimetric data: we propose a nonlocal filter whose weights are proportional to the closeness between samples (Sec. 8.1); we classify segments by minimizing the distance between samples and prototypes (Sec. 8.2); we go back to the problem of edge detection in this kind of data (Sec. 8.3); we show a technique for polarimetric decomposition, which corrects the effects of the orientation angle (Sec. 8.4), and we conclude showing how to detect changes in multitemporal images with a test based on likelihoods and by comparing entropies (Sec. 8.5).

Section 9 presents the idea of projecting polarimetric data onto the surface of a 16-dimensional sphere, measuring distances and, with this, performing decompositions and clustering.

Fig. 1 shows the mind map that led the writing of this survey. The arrows denote the model and/or technique used to tackle each problem.

The paper concludes with Section 10, where we suggest research avenues and further readings. We also give our assessment of the value of these techniques in perspective in an era dominated by the Deep Learning approach.

2 Filter Assessment

The multiplicative model is widely accepted as a good descriptor for SAR data. It assumes that the observed intensity in each pixel is the product of two independent random variables: $X$, which describes the unobserved backscatter (the quantity of interest), and $Y$, the speckle.

One can safely assume that the speckle obeys a Gamma law with unitary mean and shape parameter $L$, the number of looks, which depends on the processing level and is related to the signal-to-noise ratio.

Without making any assumption about the distribution of $X$, except that it has positive support, any despeckling filter $\hat{X}$ may be seen as an estimator of $X$ based solely on the observation of $X$, i.e., $\hat{X} = \hat{X}(Z)$. Argenti et al. present a comprehensive survey of such filters, along with measures of their performance.

Image processing platforms suited for SAR, e.g., ENVI and Snap, offer users a plethora of despeckling filters, and all of them require stipu-
lating parameters. The final user seldom chooses measures of quality, as they capture different features, and there is no obvious way to combine them.

Gomez et al. noted that the ideal filter $\hat{X}$ should produce the actual backscatter $X$ as output. The ideal ratio image $\hat{R} = Z/\hat{X} = Z/X = XY/X = Y$ is a collection of independent identically distributed (iid) Gamma deviates with unitary mean and shape parameter $L$. The worse the filter $\hat{X}$ is, the further (less adherent) the ratio image $\hat{R} = Z/\hat{X}$ will be from this model. They then proposed measuring this distance with two components: one, based on first-order statistics (mean and looks preservation), and the other based on second-order statistics (how independent the observations are). This distance can also be used to fine-tune the parameters of a filter in order to obtain optimized results.

The first component is comprised of quadratic differences between (i) local means that should be 1 and (ii) local estimates of the shape of the Gamma distribution that should be equal to the original number of looks.

The second component was initially measured through the homogeneity from Haralik’s co-occurrence matrices. The original proposal used as reference an ensemble of shuffled ratio images. Vitale et al. improved this step by comparing the joint and marginal distributions of the co-occurrence matrices.

Later, Gomez et al. studied the statistical properties of this measure, leading to a test whose null hypothesis is that the filter is ideal. This approach led to a powerful tool without assuming any distribution for $X$, thus for $Z$.

However, we need potent models in order to extract information from SAR images. These models will be reviewed in Sections 5 and 7.

3 Statistical Inference

This Section is based on the book by Frery and Wu, freely available with code and data.

Assume we have data and models for them. We will see ways of using the former to make inferences about the latter.

Without loss of generality, consider the parametric model $(\Omega, \mathcal{A}, \Pr)$ for real independent random variables $X$ and $X = (X_1, X_2, \ldots)$, in which $\Omega$ is the sample space, $\mathcal{A}$ is the smallest $\sigma$-algebra of events, and $\Pr$ is a probability measure defined on $\mathcal{A}$. The distribution of $X$ is indexed by $\theta \in \Theta \subset \mathbb{R}^p$, where $p \geq 1$ is the dimension of the parametric space $\Theta$.

Parametric statistical inferences consist in making assertions about $\theta$ using the information provided by the sample $X$.

3.1 Inference by analogy

This approach requires the knowledge of the expected value of transformations of the random variable $X$:

$$E[\psi(X)] = (E[\psi_1(X)], E[\psi_2(X)], \ldots, E[\psi_p(X)])$$

where each $\psi_j$ is a measurable function $\psi_j : \mathbb{R} \to \mathbb{R}$. Each element of Eq. (1) is given by

$$E[\psi_j(X)] = \int_{\mathbb{R}} \psi_j(x) dF(x)$$

and $F$ is the cumulative distribution function of $X$. If $\psi(X) = X^k$, we say that $E(X^k)$ is the $k$-th order moment of $X$ (if it exists).

The quantity $E(X - EX)^k$ is called “the central moment” of $X$, if it exists. The second central moment $E(X - E(X))^2 = E(X^2) - (E(X))^2$ is called “the variance” of $X$. We denote it $\text{Var}(X)$.

The Law of Large Numbers is the basis for inference by analogy. This law states that (under relatively mild conditions) holds that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \psi_i(X_i) = E[\psi(X)]$$

provided $X, X_1, X_2, \ldots$ are iid.

With this in mind, and assuming one has a large sample, it seems reasonable to equate sample quantities (the left-hand side) and parametric expressions (the right-hand side).

When we have $p$ parameters, i.e., $\theta \in \Theta \subset \mathbb{R}^p$, we need $p$ linearly independent equations to form an estimator of $\theta$. Such estimator has the following form:

$$g_1(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p) - \frac{1}{n} \sum_{i=1}^{n} \psi_1(X_i) = 0$$

$$\vdots$$

$$g_p(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p) - \frac{1}{n} \sum_{i=1}^{n} \psi_p(X_i) = 0$$
More often than not, one has to rely on numerical routines to obtain $\hat{\theta}$ with this approach. In some cases, Eq. (4) has no solution; cf. [10].

### 3.2 Inference by maximum likelihood

Consider again the sample of iid random variables $X = (X_1, X_2, \ldots, X_n)$ each with the same distribution characterized (without lack of generality) by the density $f(X_i; \theta)$. The likelihood function is

$$L(\theta; X) = \prod_{i=1}^{n} f(\theta; X_i)$$

(5)

Notice that $L$ is not a joint density function, as it depends on $\theta$, not on the variables.

The principle of maximum likelihood proposes as estimator for $\theta$ the parameter that maximizes Eq. (5):

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} \{ L(\theta; X) \}$$

(6)

that is, the point in $\Theta$ that makes the observations most plausible. It sometimes coincides with some analogy estimators.

The two most widely used approaches for maximizing Eq. (5) are by optimization and by the solution of a system of equations. They both operate on the reduced log-likelihood rather than on Eq. (5).

Using the fact that Eq. (5) is a product of positive functions, one may consider

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} \left\{ \ln L(\theta; X) \right\}$$

(7)

instead of Eq. (6). If we now discard from $\ln L(\theta; X)$ the terms that do not depend on $\theta$, we obtain the reduced log-likelihood $\ell(\theta; X)$, our function of interest.

A maximum likelihood estimator may be found by maximizing $\ell(\theta; X)$, often with numerical tools, or by solving the system of equations given by

$$\nabla \ell(\hat{\theta}; X) = 0$$

(8)

i.e., the partial derivatives with respect to each $\theta_i$ equated to zero.

Both approaches often require an initial solution, and a general recommendation is towards using an analogy estimator (if available) for that purpose.

Provided there are no doubts about the validity of the model, the maximum likelihood estimators should be preferred over those obtained by analogy.

### 3.3 Other inference techniques

There are two remarkable approaches to statistical inference that differ from analogy and maximum likelihood: nonparametric and robust estimation.

The nonparametric approach makes little or no assumptions about the underlying distribution of the data. Cintra et al. [11] employed such techniques with success for the problem of finding edges in SAR imagery. Nonparametric estimators are resistant to contamination, at the expense of having less efficiency. They should be part of every practitioner’s toolbox. The interested reader is referred to the books listed in Refs. [12–14] for details about this approach.

The reader must always bear in mind the underlying hypotheses used to devise an inference procedure. All the techniques discussed above assume that the sample is comprised of observations from iid random variables. A middle path between parametric (Analogy and maximum likelihood) and nonparametric is the robust approach.

Robust inference assumes there is a model of interest $D(\theta)$, and that we aim at estimating $\theta$ with a sample $Z = (Z_1, Z_2, \ldots, Z_n)$. The sample may or may not be produced by $D(\theta)$ as hypothesized.

A robust estimator for $\theta$ should behave well when the model is correct, although not as good as the maximum likelihood estimator $\hat{\theta}_{ML}$ (by definition). However, differently from $\hat{\theta}_{ML}$, it should provide acceptable estimates when $Z$ originates from a different distribution, i.e., under the presence of contamination.

Contamination may describe several departures from the primary hypothesis as, for instance, a few gross errors when registering the observations, or the presence of data originated from a different law. Contamination happens, for instance, when computing local statistics for designing a filter. The underlying assumption that $Z$ consists of observations from iid random variables breaks when the estimation window is transitioning over two or more different areas. The
works by Palacio et al.\cite{15} and by Negri et al.\cite{16} examine the impact that mild violations of this hypothesis on the behavior of subsequent operations, when the inference procedure is unaware of such departure.

Frey et al.\cite{17} proposed robust estimators for speckle reduction in single-look amplitude data. The authors used an empirical approach and devised estimators based on trimmed statistics, on the median absolute deviation, on the inter-quartile range, on the median, and on the best linear unbiased estimator. They compared the performance of these techniques with respect to estimators based on analogy (moments) and maximum likelihood and concluded that it is advisable to consider robust procedures.

More recently, Chan et al.\cite{18} used a connection between the model defined by (21) and the Pareto distribution to propose robust estimators with low computational cost. The authors show applications to region classification in the presence of extreme values caused by, for instance, corner reflectors.

Refs. \cite{19–21} discuss other robust techniques for inference with small samples under contamination.

The visible gain in using robust estimators in speckle reduction techniques is a reduced blurring of the edges and the preservation of small details in the image. This advantage usually comes at the cost of more computational resources.

4 Measures of Difference

Several relevant problems in signal and image processing and analysis can be solved by formulating them in the form of the following hypothesis test:

Consider the \( N \geq 2 \) samples \( z_1, z_2, \ldots, z_N \) of the form \( z_j = z_{1,j}, z_{2,j}, \ldots, z_{n_j,j} \) for every \( 1 \leq j \leq N \). Is there enough evidence to reject the null hypothesis that they were produced by the same model \( D(\theta) \)?

This problem assumes that the \( N \) samples might be of different sizes \( n_1, n_2, \ldots, n_N \), and that the model \( D(\theta) \) is a probability distribution indexed by the unknown \( q \)-dimensional parameter \( \theta \in \Theta \subset \mathbb{R}^q \). Eventually, there might be no known model, and it will have to be learned from the data and the underlying hypotheses of the problem.

In most of this review, with the exception of Section 4.3, we will be handling two samples \( z_1, z_2 \) of sizes \( n_1 \) and \( n_2 \) of the form \( z_1 = z_{1,1}, z_{2,1}, \ldots, z_{n_1,1} \) and \( z_2 = z_{1,2}, z_{2,2}, \ldots, z_{n_2,2} \).

In the following, we will review some of the relevant approaches in the context of SAR image analysis, which have proven to be relevant at handling this problem.

4.1 Likelihoods ratio

The ratio of likelihoods (LR) statistic is of great importance in inference on parametric models. It is based on comparing two (reduced) likelihoods utilizing a ratio of two maxima: the numerator is the likelihood in the alternative hypothesis \( \mathcal{H}_1: \theta \in \Theta_1 \), and the denominator is the likelihood in the null hypothesis \( \mathcal{H}_0: \theta \in \Theta_0 \). These hypotheses must be a partition of the parametric space: \( \Theta_0 \cap \Theta_1 = \emptyset \) and \( \Theta_0 \cup \Theta_1 = \Theta \). Under the null hypothesis, the denominator of the test statistic should be large compared to the numerator yielding a small value.

As discussed in Ref. \cite{22}, such statistic has, under \( \mathcal{H}_0 \), an asymptotic distribution \( \chi^2_q \), where \( q \) is the difference between the dimensions of the parameter spaces under the alternative and the null hypotheses. It is, thus, possible to obtain \( p \)-values.

4.2 Tests based on divergences

This section is based on the work by Nascimento et al.\cite{23}. Contrast analysis requires quantifying how distinguishable two samples are requiring. Within a statistical framework, this amounts to providing a test statistic for the null hypothesis that the same distribution produced the samples \( D(\theta) \), with \( \theta \in \Theta \subset \mathbb{R}^q \) the parameter space.

Information theoretical tools collectively known as divergence measures, offer entropy-based methods to discriminate stochastic distributions\cite{24} statistically. Divergence measures were submitted to a systematic and comprehensive treatment in Refs. \cite{25–27} and, as a result, the class of \( (h, \phi) \)-divergences was proposed\cite{27}.

Let \( X \) and \( Y \) be random variables defined over the same probability space, equipped with
densities $f_X(x; \theta_1)$ and $f_Y(x; \theta_2)$, respectively, where $\theta_1$ and $\theta_2$ are parameter vectors. Assuming that both densities share a common support $I \subset \mathbb{R}$, the $(h, \phi)$-divergence between $f_X$ and $f_Y$ is defined by

$$
D_h^\phi(X, Y) = h \left( \int_I \phi \left( \frac{f_X(x; \theta_1)}{f_Y(x; \theta_2)} \right) f_Y(x; \theta_2) dx \right) \tag{9}
$$

where $\phi: (0, \infty) \to [0, \infty)$ is a convex function, $h: (0, \infty) \to [0, \infty)$ is a strictly increasing function with $h(0) = 0$, and indeterminate forms are assigned value zero.

By a judicious choice of functions $h$ and $\phi$, some well-known divergence measures arise. Tab. 1 shows a selection of functions $h$ and $\phi$ that lead to well-known distance measures. Specifically, the following measures were examined: (i) the Kullback-Leibler divergence, (ii) the relative Rényi (also known as Chernoff) divergence of order $\beta$, (iii) the Hellinger distance, (iv) the Bhattacharyya distance, (v) the relative Jensen-Shannon divergence, (vi) the arithmetic-geometric divergence, (vii) the triangular distance, and (viii) the harmonic-mean distance.

Often not rigorously a metric, since the triangle inequality does not necessarily hold, divergence measures are mathematically suitable tools in the context of comparing distributions. Additionally, some of the divergence measures lack symmetry property. Although there are numerous methods to address the symmetry problem, a simple solution is to define a new measure $d_h^\phi$ given by

$$
d_h^\phi(X, Y) = \frac{D_h^\phi(X, Y) + D_h^\phi(Y, X)}{2} \tag{10}
$$

regardless whether $D_h^\phi(\cdot, \cdot)$ is symmetric or not. Henceforth, the symmetrized versions of the divergence measures are termed “distances”. By applying the functions of Tab. 1 into Eq. (9), and symmetrizing the resulting divergences, integral formulas for the distance measures are obtained. For simplicity, in the list below, we suppress the explicit dependence on $x$ and the support $I$ in the notation.

(i) The Kullback-Leibler distance:

$$
d_{KL}(X, Y) = \frac{1}{2} \int (f_X - f_Y) \ln \frac{f_X}{f_Y} dx
$$

(ii) The Rényi distance of order $\beta$:

$$
d_\beta^\phi(X, Y) = \frac{1}{\beta - 1} \ln \left\{ \int \frac{f_\beta^\phi(X)}{f_\beta^\phi(Y)} + \frac{f_\beta^\phi(Y)}{f_\beta^\phi(X)} \right\} \tag{11}
$$

(iii) The Hellinger distance:

$$
d_H(X, Y) = 1 - \sqrt{\int f_X f_Y} = 1 - \exp \left\{ -\frac{1}{2} d_R^\phi(X, Y) \right\}
$$

(iv) The Bhattacharyya distance:

$$
d_B(X, Y) = -\ln \int \sqrt{f_X f_Y} = -\ln (1 - d_H(X, Y))
$$

(v) The Jensen-Shannon distance:

$$
d_{JS}(X, Y) = \frac{1}{2} \left[ \int f_X \ln \frac{2f_X}{f_Y + f_X} + \int f_Y \ln \frac{2f_Y}{f_Y + f_X} \right]
$$

(vi) The arithmetic-geometric distance:

<table>
<thead>
<tr>
<th>$(h, \phi)$-divergence</th>
<th>$h(y)$</th>
<th>$\phi(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kullback-Leibler</td>
<td>$y$</td>
<td>$\frac{x \ln(x)}{\beta - 1}$, $0 &lt; \beta &lt; 1$</td>
</tr>
<tr>
<td>Rényi (order $\beta$)</td>
<td>$(\beta - 1)y + 1$, $0 \leq y &lt; \frac{1}{1 - \beta}$</td>
<td>$\frac{x \ln(x)}{\beta - 1}$, $0 &lt; \beta &lt; 1$</td>
</tr>
<tr>
<td>Hellinger</td>
<td>$y/2$, $0 \leq y &lt; 2$</td>
<td>$(\sqrt{x} - 1)^2$</td>
</tr>
<tr>
<td>Bhattacharyya</td>
<td>$-\ln(1 - y)$, $0 \leq y &lt; 1$</td>
<td>$-\sqrt{x} + \frac{x + 1}{2}$</td>
</tr>
<tr>
<td>Jensen-Shannon</td>
<td>$y$</td>
<td>$x \ln(x)$</td>
</tr>
<tr>
<td>Arithmetic-geometric</td>
<td>$y$</td>
<td>$\sqrt{x} + \frac{x + 1}{2}$</td>
</tr>
<tr>
<td>Triangular</td>
<td>$y$, $0 \leq y &lt; 2$</td>
<td>$(x - 1)^2$</td>
</tr>
<tr>
<td>Harmonic-mean</td>
<td>$-\ln \left( -\frac{y}{2} + 1 \right)$, $0 \leq y &lt; 2$</td>
<td>$\frac{(x - 1)^2}{x + 1}$</td>
</tr>
</tbody>
</table>
\[ d_{AC}(X, Y) = \frac{1}{2} \int (f_X + f_Y) \ln \frac{f_Y + f_X}{2 \sqrt{f_Y f_X}} \]

(vii) The triangular distance:
\[ d_T(X, Y) = \int \frac{(f_X - f_Y)^2}{f_X + f_Y} \]

(viii) The harmonic-mean distance:
\[ d_{HM}(X, Y) = -\ln \left(\int \frac{2f_X f_Y}{f_X + f_Y} = -\ln \left(1 - \frac{d_T(X, Y)}{2} \right)\right) \]

We conclude this section stating one of the basilar results that support all the forthcoming techniques.

The distances mentioned above are neither comparable nor semantically rich. Refs. [35–37] are pioneering works that make a connection between any \( h \)-\( \phi \) distance and a test statistic.

Consider the model \( \mathcal{D}(\theta) \), with \( \theta \in \Theta \subset \mathbb{R}^M \) an the unknown parameter that indexes the distribution. Assume the availability of two samples of \( \text{iid} \) observations \( Z = Z_1, Z_2, \ldots, Z_{n_1} \) from \( \mathcal{D}(\theta_1) \), and \( Z = Z_{n_1+1}, Z_{n_1+2}, \ldots, Z_{n_1+n_2} \) from \( \mathcal{D}(\theta_2) \), where \( \theta_1, \theta_2 \in \Theta \). Compute the maximum likelihood estimators of \( \theta_1 \) and \( \theta_2 \) of \( Z_1 \) and \( Z_{2} \). We are interested in verifying if there is enough evidence in \( \hat{\theta}_1, \hat{\theta}_2 \) to reject the null hypothesis \( H_0 : \theta_1 = \theta_2 \).

Under the regularity conditions discussed in Ref. [27, p. 380] the following lemma holds:

Lemma 1: If \( n_1/(n_1 + n_2) \) converges to a constant in \( (0, 1) \) when \( n_1, n_2 \to \infty \), and \( \theta = \theta_2 \), then
\[
S_h^\phi(\hat{\theta}_1, \hat{\theta}_2) = \frac{2n_1 n_2}{n_1 + n_2} \frac{d_h^\phi(\hat{\theta}_1, \hat{\theta}_2)}{h(0) d''(1)} \chi^2_M
\]
where “\( \chi^2_M \)” denotes convergence in distribution and \( \chi^2_M \) represents the chi-square distribution with \( M \) degrees of freedom.

Based on Lemma 1, we obtain a test for the null hypothesis \( \theta_1 = \theta_2 \) in the form of following proposition.

Proposition 1: Let \( n_1 \) and \( n_2 \) be large and \( S_h^\phi(\hat{\theta}_1, \hat{\theta}_2) = s \), then the null hypothesis \( \theta_1 = \theta_2 \) can be rejected at level \( \alpha \) if \( \Pr(\chi^2_M > s) \leq \alpha \).

This result holds for every \( h \)-\( \phi \) divergence, hence its generality. The (asymptotic) \( p \)-value of the \( S_h^\phi \) statistic can be used as a measure of the evidence in favor of \( H_0 \), so it has rich semantic information. As the convergence is towards the same distribution, these values are comparable.

4.3 Tests based on the quadratic difference of entropies

The test statistic presented in Section 4.2 compares a pair of maximum likelihood estimators \( \hat{\theta}_1, \hat{\theta}_2 \). Many applications, as, for instance, change detection in time series, requires using more evidence. In such cases it is possible to compare entropies.

For such situations, one may rely on tests based on entropies, as described in the following results, which are based on Ref. [38].

Let \( f(z; \theta) \) be a probability density function with parameter vector \( \theta \). The \( h \)-\( \phi \) entropy relative to \( \mathcal{Z} \) is defined by
\[
H_h^\phi(\theta) = h\left(\int \phi(f(z; \theta))dz\right)
\]
where either \( \phi : [0, \infty) \to \mathbb{R} \) is concave and \( h : \mathbb{R} \to \mathbb{R} \) is increasing, or \( \phi \) is convex and \( h \) is decreasing. Tab. 2 shows the specification of five of these entropies: Shannon, Rényi, Arimoto, Sharma-Mittal, and restricted Tsallis.

The following result, derived by Pardo et al.[37], paves the way for the proposal of asymptotic statistical inference methods based on entropy.

Lemma 2: Let \( \hat{\theta} \) be the ML estimate of the parameter vector \( \theta \) of dimension \( M \) based on a random sample of size \( n \) under the model \( f(z; \theta) \). Then
\[
\sqrt{N} \left[ H_h^\phi(\hat{\theta}) - H_h^\phi(\theta) \right] \xrightarrow{n \to \infty} \mathcal{N}(0, \sigma^2_h(\theta))
\]
where \( \mathcal{N}(\mu, \sigma^2) \) is the Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \),
\[
\sigma^2_h(\theta) = \delta^T \mathcal{K}(\theta)^{-1} \delta
\]
where \( \mathcal{K}(\theta) = E\left[-(\partial^2 f(z; \theta)/\partial \theta^2)\right] \) is the Fisher information matrix, and \( \delta = [\delta_1, \delta_2, \ldots, \delta_p]^T \) such that \( \delta_i = \partial H_h^\phi(\theta)/\partial \theta_i \) for \( i = 1, 2, \ldots, M \).

We are interested in testing the following hypotheses:
\[
\{ H_0 : H_h^\phi(\theta_1) = H_h^\phi(\theta_2) = \cdots = H_h^\phi(\theta_N) = v \}
\]
\[
\{ H_1 : H_h^\phi(\theta_i) \neq H_h^\phi(\theta_j) \text{ for some } i \text{ and } j \}
\]
In other words, we seek for statistical evidence for assessing whether at least one of the \( N \) samples has different entropy when compared to the remaining ones.
Let \( \hat{\theta}_i \) be the ML estimate for \( \theta_i \) based on a random sample of size \( n_i \) under \( Z_i \), for \( i = 1, 2, \ldots, N \).

From Lemma 2 we have that
\[
\frac{\sqrt{n_i} (H^\phi_h(\hat{\theta}_i) - v)}{\sigma_{H}^2(\hat{\theta}_i)} \xrightarrow{D} \mathcal{N}(0,1)
\]
for \( i = 1, 2, \ldots, N \). Therefore,
\[
\sum_{i=1}^{N} \frac{n_i (H^\phi_h(\hat{\theta}_i) - v)^2}{\sigma_{H}^2(\hat{\theta}_i)} \xrightarrow{D} \chi^2_r
\]
(13)

Since \( r \) is, in practice, unknown, in the following we modify this test statistic in order to take this into account. We obtain:
\[
\sum_{i=1}^{N} \frac{n_i (H^\phi_h(\hat{\theta}_i) - v)^2}{\sigma_{H}^2(\hat{\theta}_i)} = \sum_{i=1}^{N} \frac{n_i (H^\phi_h(\hat{\theta}_i) - v)^2}{\sigma_{H}^2(\hat{\theta}_i)} + \sum_{i=1}^{N} n_i (v - \bar{v})^2
\]
(14)

where
\[
\bar{v} = \left[ \sum_{i=1}^{N} \frac{n_i}{\sigma_{H}^2(\hat{\theta}_i)} \right]^{-1} \sum_{i=1}^{N} \frac{n_i H^\phi_h(\hat{\theta}_i)}{\sigma_{H}^2(\hat{\theta}_i)}
\]

Salicrú et al.\(^{[35]}\) showed that the second summation in the right-hand side of Eq. (14) is chi-square distributed with one degree of freedom. Since the left-hand side of Eq. (14) is chi-square distributed with \( N \) degrees of freedom (cf. Eq. (13)), we conclude that:
\[
\sum_{i=1}^{N} \frac{n_i (H^\phi_h(\hat{\theta}_i) - v)^2}{\sigma_{H}^2(\hat{\theta}_i)} \xrightarrow{D} \chi_{N-1}^2
\]

In particular, consider the following test statistic:
\[
S_{\phi}^h(\theta_1, \theta_2, \ldots, \theta_N) = \sum_{i=1}^{N} \frac{n_i (H^\phi_h(\hat{\theta}_i) - v)^2}{\sigma_{H}^2(\hat{\theta}_i)} \quad (15)
\]

We are now in the position to state the following result.

Proposition 2: Let each sample size \( n_i, i = 1, 2, \ldots, N \), be sufficiently large. If \( S_{\phi}^h(\theta_1, \theta_2, \ldots, \theta_N) = s \), then the null hypothesis \( H_0 \) can be rejected at a level \( \alpha \) if \( \Pr(\chi^2_{N-1} > s) \leq \alpha \).

Again, this Proposition transforms a mere distance into a quantity with concrete meaning: a \( p \)-value.

4.4 Tests based on geodesic distances

Assume we are characterizing the data by the distribution \( D(\theta) \), so each population is completely described by a particular point \( \theta = (\theta_1, \theta_2, \ldots, \theta_q) \in \Theta \subset \mathbb{R}^q \). The information matrix of the model has entries
\[
[g_{ij}(\theta)]_{1 \leq i, j \leq q} = \mathbb{E} \left( \frac{\partial^2 \ln f(Z; \theta)}{\partial \theta_i \partial \theta_j} \right)_{1 \leq i, j \leq q}
\]
(16)

According to Ref. [40], it can be used to measure the variance of the relative difference between two close models as
\[
ds^2 = \sum_{i,j=1}^{q} g_{ij}(\theta)d\theta_id\theta_j
\]

This quadratic differential form is the basis for computing distances between models with
\[
s(\theta_1, \theta_2) = \sqrt{\int_{t_1}^{t_2} \sum_{i,j=1}^{q} g_{ij}(\theta(t)) \frac{d\theta_i(t)}{dt} \frac{d\theta_j(t)}{dt} dt}
\]

where \( \theta(t) \) is a curve joining \( \theta_1 \) and \( \theta_2 \) such that \( \theta(t_1) = \theta_1 \) and \( \theta(t_2) = \theta_2 \). Among all possible curves, we are interested in the shortest one; this leads to the geodesic distance between the models. Finding such a curve is, in general, very difficult, in particular, for distributions with parameters other than \( q = 1 \).

An alternative approach consists in consider-
ing distribution families with a scalar parameter \( \theta \in \Theta \subset \mathbb{R} \), i.e., with \( q = 1 \). With this restriction in mind, we can define a generalized form of distance: that based on \( h-\phi \) entropies as the solution of

\[
\frac{d}{d\theta} \left[ \phi(f(z; \theta)) \right] = \frac{1}{g(\theta)} \int_{0}^{d\theta} \sqrt{g(\theta)} d\theta
\]

where

\[
g(\theta) = h'' \left( \int \phi(f(z; \theta)) dz \right) \\
\cdot \left[ \int \phi'(f(z; \theta)) dz \frac{\partial f(z; \theta)}{\partial \theta} dz \right]^2 \\
+ h' \left( \int \phi(f(z; \theta)) dz \right) \\
\cdot \int \phi''(f(z; \theta)) dz \left[ \frac{\partial f(z; \theta)}{\partial \theta} \right]^2 dz
\]

(17)

In analogy with the results presented for entropies and divergences, Menéndez et al.\cite{41} proved the following asymptotic result: consider, for simplicity, \( h(y) = y \) and \( \phi \) as in previous sections. The test statistic

\[
S^h_\phi(\hat{\theta}_1, \hat{\theta}_2) = \frac{n_1 n_2}{n_1 + n_2} \left( \frac{\hat{\theta}_1 - \hat{\theta}_2}{\phi'(0)} \right)
\]

has, under the null hypothesis \( \theta_1 = \theta_2 \) and for sufficiently large \( n_1 \) and \( n_2 \), \( \chi^2_1 \) asymptotic distribution.

The classical geodesic distance is obtained when \( h \) and \( \phi \) are those leading to the Shannon entropy; cf. Tab. 2. In this case, \( g(\theta) \) becomes \( g_{11}(\theta) \), the only element of Eq. (16).

This alternative approach is more tractable, but requires a further step, namely joining the \( q \) test statistics into a single measure of dissimilarity. After Naranjo-Torres et al.\cite{42} showed the ability of the Shannon geodesic distance at finding edges in intensity SAR data, Frey and Gambini\cite{43} compared three fusion techniques.

5 Intensity Data

The most famous models for intensity SAR data are the Gamma, \( \mathcal{K} \), and \( \mathcal{G}^0 \) distributions. Ref. [44] presents a survey of models for SAR data. The first is adequate for fully developed speckle, while the two others can describe the texture, i.e., variations of the backscatter between pixels with an additional parameter. Their probability density functions are, respectively:

\[
f(z; \mu, L) = \frac{L}{\Gamma(L) \mu^{L-1}} z^{L-1} \exp\{-Lz/\mu\}
\]

(19)

\[
f(z; \alpha, \lambda, L) = \frac{2L\lambda^{\alpha}}{\Gamma(\alpha) \Gamma(L)} (Lz)^{\alpha-1} \\
\cdot K_{\alpha-1}(2\sqrt{Lz})
\]

(20)

\[
f(z; \alpha, \gamma, L) = \frac{L^\gamma \Gamma(L - \alpha)}{\gamma \Gamma(L) \Gamma(-\alpha)} (\gamma + Lz)^{L-\alpha-1}
\]

(21)

where, in Eq. (19) \( \mu > 0 \) is the mean; in Eq. (20) \( \lambda, \alpha > 0 \) are the scale and the texture, and \( K_{\nu} \) is a modified Bessel function of order \( \nu \); and in Eq. (21) \( \gamma > 0 \) is the scale and \( \alpha < 0 \) is the texture. In all these expressions, \( L \geq 1 \) is the number of looks.

The \( \mathcal{G}^0 \) law is flexible (it can describe observations with a wide range of texture, and it can approximate both the Gamma and \( \mathcal{K} \) distributions) and tractable (its density does not involve Bessel functions, while the \( \mathcal{K} \) does). Parameter estimation under the \( \mathcal{G}^0 \) model may be difficult, however.

Maximum likelihood estimators require iterative methods that may not converge, and estimators by analogy may not have feasible solutions. Moreover, contamination is a significant issue, mostly when dealing with small samples and/or in the presence of strong backscatter as the one produced by a double bounce.

The models given by Eq. (19) and Eq. (21) are baseline to this work. The next subsections present properties that will be useful in the sequence.

5.1 The Gamma model

Denote by \( Z \sim \Gamma(\mu, L) \) a random variable that follows the Gamma distribution characterized by density Eq. (19). The \( k \)-th order moment of \( Z \) is

\[
E(Z^k) = \frac{\Gamma(L+k)}{\Gamma(L)}
\]

(22)

The Gamma model given by Eq. (19) includes the Exponential distribution when \( L = 1 \), i.e., in the single-look case. Fig. 2 shows the densities that characterize this distribution with varying means. The linear scale (Fig. 2(a)) shows the well-known exponential decay, while in the semilogarithmic scale (Fig. 2(b)) they appear as straight lines with a negative slope.
Fig. 3 shows the effect of multilooking the data on their distribution: it shows three densities with the same mean ($\mu = 1$) and $L = 1, 3, 8$. The larger $L$ becomes, the smaller the probability of extreme events, as can be seen in the semilogarithmic scale (Fig. 3(b)). Also, as $L$ increases, the densities become more symmetric; cf. Fig. 3(a).

Assume one has a sample of $n$ iid $\Gamma(\mu, L)$ deviates $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$. Inference about $(\mu, L)$ can be made by the analogy method using the moments given in Eq. (22). For instance, choosing to work with the first moment and the variance, one has:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Z_i \quad (23)$$

$$\hat{L} = \left( \frac{1}{n} \sum_{i=1}^{n} Z_i \right)^2 - \frac{1}{n} \sum_{i=1}^{n} (Z_i - \hat{\mu})^2 \quad (24)$$

The maximum likelihood estimator of $(\mu, L)$ is the maximum of Eq. (25). This maximum may be found by optimization of Eq. (25), or by solving the system of equations given by $\nabla \ell = 0$.

As an alternative to any analogy method, one may rely on maximum likelihood estimation. From Eq. (19), one has that the reduced loglikelihood of the sample $\mathbf{Z}$ is

$$\ell(\mu, L; \mathbf{Z}) = L \ln L - \ln \Gamma(L) - L \ln \mu$$

$$+ \frac{L}{n} \sum_{i=1}^{n} \ln Z_i + \frac{L}{\mu n} \sum_{i=1}^{n} Z_i \quad (25)$$

This second approach to obtaining the maximum likelihood estimator of $(\mu, L)$ leads to Eq. (23) and to the solution of

\begin{align*}
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} Z_i \\
\hat{L} &= \left( \frac{1}{n} \sum_{i=1}^{n} Z_i \right)^2 - \frac{1}{n} \sum_{i=1}^{n} (Z_i - \hat{\mu})^2
\end{align*}
\[
\ln \hat{L} + 1 - \psi(\hat{L}) - \ln \hat{\mu} + \frac{1}{n} \sum_{i=1}^{n} \ln Z_i + \frac{1}{\mu n} \sum_{i=1}^{n} Z_i = 0
\]

(26)

where \( \psi \) is the digamma function.

Either way, optimization or finding the roots of Eq. (26), one relies on numerical procedures that, more often than not, require a starting point or initial solution. Moment estimators are convenient initial solutions for this.

5.2 The \( G^0 \) model

Frery et al.\cite{45} noticed that the \( K \) distribution failed at describing data from extremely textured areas as, for instance, urban targets. The authors then proposed a different model for the backscatter \( X \): the Reciprocal Gamma distribution.

We say that \( X \sim \Gamma^{-1}(\alpha, \gamma) \), with \( \alpha < 0 \) and \( \gamma > 0 \) follows a Reciprocal Gamma distribution is characterized by the density

\[
f_X(x; \alpha, \gamma) = \frac{\gamma^{-\alpha}}{\Gamma(-\alpha)} x^{\alpha-1} \exp\{-\gamma/x\}
\]

(27)

for \( x > 0 \) and zero otherwise.

Now introducing the Reciprocal Gamma model for the backscatter in the multiplicative model, \( i.e. \) by multiplying the independent random variables \( X \sim \Gamma^{-1}(\alpha, \gamma) \) and \( Y \sim \Gamma(1, L) \), one obtains the \( G^0 \) distribution for the return \( Z = XY \), which is characterized by the density Eq. (21). It is noteworthy that, differently from Eq. (20), this density does not involve Bessel functions.

We denote \( Z \sim G^0(\alpha, \gamma, L) \) the situation of \( Z \) following the distribution characterized by Eq. (21). The \( k \)-order moments of \( Z \) are

\[
E(Z^k) = \left( \frac{\gamma}{L} \right)^k \frac{\Gamma(L+k)\Gamma(-\alpha-k)}{\Gamma(L)\Gamma(-\alpha)}
\]

(28)

provided \(-\alpha > k \), and infinite otherwise. Eq. (28) is useful, among other applications, for finding \( \gamma^* = -\alpha - 1 \), the scale parameter that yields a unitary mean distribution for each \( \alpha \) and any \( L \).

It also allows forming systems of equations for estimating the parameters by analogy.

Fig. 4 shows \( G^0(\alpha, \gamma^*, 1) \) densities along with an Exponential density also with unitary mean. The differences in tail behavior are clearly exhibited in the semilogarithmic scale; cf. Fig. 4(b). Whereas the exponential distribution decreases linearly, the \( G^0 \) law assigns more probability to larger events increasing, thus, the variability of the return.

Fig. 5 shows the effect of varying the number of looks, for the same \( \alpha = 5 \) and \( \gamma = 4 \).

Notice, again, in Fig. 5(b) the effect multilook processing has mostly on the distribution of minimal values. Such effect, along with the reduced probability of huge values, have with multilook processing, yields less contrasted images.

The \( G^0 \) distribution has the same number of parameters as the \( K \) law, but it has been shown to be more apt at modeling return with extreme variability. Moreover, it is also able to describe the same kind of return from textured areas for which the latter was proposed; Mejail et al.\cite{16} showed that, with proper choices of parameters, the \( G^0 \) law is able to approximate with any error any \( K \) distribution. For these reasons, the \( G^0 \)
The \( G^0 \) distribution relates to the well-known Fisher-Snedekor law in the following manner: 
\[
F_{G^0(\alpha, \gamma, L)}(t) = \frac{\gamma \alpha}{\gamma + \alpha t} 
\]
where \( \gamma_{u,v} \) is the cumulative distribution function of a Fisher-Snedekor distribution with \( u \) and \( v \) degrees of freedom, and \( F_{G^0(\alpha, \gamma, L)} \) is the cumulative distribution function of a \( G^0(\alpha, \gamma, L) \) random variable. Notice that \( \gamma \) is readily available in most software platforms for statistical computing. Since such platforms usually also provide implementations of the inverse of cumulative distribution functions, the Inversion Theorem can be used to sample from the \( G^0 \) law.

Assume we have a sample \( Z = (Z_1, Z_2, \ldots, Z_n) \) of iid random variables that follow the \( G^0(\alpha, \gamma, \tilde{L}) \) distribution, i.e., \( L \) has been already estimated by \( \tilde{L} \). The unknown parameter \( \theta = (\alpha, \gamma) \) lies in \( \Theta = \mathbb{R}_+ \times \mathbb{R}_+ \).

The maximum likelihood estimator of \( (\alpha, \gamma) \), is any point that maximizes the reduced log-likelihood:
\[
\ell(\alpha, \gamma; \tilde{L}, Z) = \ln \Gamma(\tilde{L}) - \ln(\Gamma(\alpha) + \Gamma(\gamma)) + \sum_{i=1}^n \ln \frac{Z_i}{\gamma + \tilde{L}Z_i} + \ln \Gamma(\gamma + \tilde{L}) 
\]
provided it lies in \( \Theta \). Maximizing Eq. (30) might be a difficult task, in particular in textureless areas where \( \alpha \to -\infty \) and the \( G^0 \) distribution becomes very close to a Gamma law.

The \( G^0 \) distribution offers several sampling approaches, and Chan et al.\(^{[46]} \) compared them for a variety of situations and platforms regarding accuracy and precision.

6 Applications to Intensity Data

6.1 Estimation of the equivalent number of looks

The three models for intensity data presented in Eqs. (19)—(21) have one parameter in common: \( L \), the number of looks. Such feature is absent in empirical models\(^{[44]} \), and describes the number of independent models used to form the observation \( Z \) at each pixel. It is a direct measure of the signal-to-noise ratio and, thus, of the visual quality of the image.

Although it is usually part of the image information, such nominal value is often larger than the evidence provided by the data. It is, thus, important to estimate this quantity in order to make a fair statistical description of the observations.

Fig. 6 shows a 400 \times 1000 pixels intensity image obtained by the ESAR sensor\(^{[47]} \) sensor in L-band and VV polarization. The area consists of agricultural parcels, a river (the dark curvilinear feature to the right), and a bright urban area to the lower right. The data have been equalized for visualization purposes, and we also show a regular grid comprised of squares of 25 \times 25 pixels that aids obtaining disjoint samples.

Assume we have collected \( N \) samples \( z_1, z_2, \ldots, z_N \) of the form \( z_j = z_{1,j}, z_{2,j}, \ldots, z_{n_j,j} \) for every \( 1 \leq j \leq N \) from areas which do not exhibit any evidence of not belonging to the Gamma distribution (at least, they do not show texture). These
samples might be of different sizes $n_1, n_2, \ldots, n_N$. Among the many possible approaches for estimating the number of looks, we will consider two of the simplest and most widely used in the literature. In both cases, we will assume that a $\Gamma(\mu_j, L)$ model holds for all the samples, i.e., that the mean may vary but not the number of looks.

In general, it is preferred to take several small disjoint samples with some spatial separation. This alleviates the problem of using correlated rather than independent observations.

The forthcoming results were obtained with $N = 12$ samples of equal size $n = 625$ each.

1. Weighted mean of estimates: Use each sample in Eqs. (23) and (24) to obtain the $N$ estimates $\hat{L}_1, \hat{L}_2, \ldots, \hat{L}_N$, then compute the final estimate by the weighted average

$$\hat{L}_{wme} = \frac{\sum_{j=1}^{N} n_j \hat{L}_j}{\sum_{j=1}^{N} n_j}$$

With the data collected over the image shown in Fig. 6 we obtained $\hat{L}_{wme} = 0.884$.

2. Regression: From Eq. (22) we have that $\text{Var}(Z) = \mu^2/L$, then the coefficient of variation $\text{CV}(Z) = \sqrt{\text{Var}(Z)}/\mu = L^{-1/2}$. With this, we may form the equation $\mu = \sqrt{L \text{Var}(Z)}$.

We now compute the sample mean $\bar{\mu}_j$ and variance $s_j^2$ of each sample $z_j$, and form the following regression model data:

$$\bar{\mu}_1 = \sqrt{L}s_1 + \varepsilon_1,$$
$$\bar{\mu}_2 = \sqrt{L}s_2 + \varepsilon_2,$$
$$\vdots$$
$$\bar{\mu}_N = \sqrt{L}s_N + \varepsilon_N$$

Notice that, although equivalent, we preferred to denote the sample mean as $\bar{\mu}_j$ instead of $\mu_j$ to elucidate the difference between simple mean and parameter estimation. Assuming further that the errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are iid zero mean Gaussian random variables with unknown variance, we are now in position to apply a simple regression model without intercept to estimate $L$.

Denote the dependent variable $\mathbf{s} = (\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_N)^T$, the independent variable $\mathbf{s} = (s_1, s_2, \ldots, s_N)^T$, and the errors $\mathbf{e} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^T$, where the superscript “T” denotes the transpose. The model is $\mathbf{s} = L^{1/2} \mathbf{1} + \mathbf{e}$, and the estimate of $L$ by regression is then given by

$$\hat{L}_e = \left( \frac{\mathbf{s}^T \mathbf{s}}{\mathbf{s}^T \mathbf{s}} \right)^2$$

Fig. 7 shows the points $(\bar{\mu}_j, s_j)$, the adjusted regression line, and its confidence band at the 95%. With this approach we obtained $\hat{L}_e = 0.756$.

Since the equivalent number of looks is assumed, in many cases, constant over the whole image, it is of paramount importance to make a careful estimation.

In the two following sections, the notion of “contrast” is the difference of statistical properties between two samples.

### 6.2 Edge detection by maximum likelihood

Besides providing estimators with excellent properties, the likelihood function can also be used to find edges. The approach proposed in Refs. [48–50] consists in forming a likelihood that models a strip of data, and finding the point where it is maximized. This point will, then, be the estimator of the edge position.

Fig. 8 illustrates the rationale behind this proposal. It shows strips of $10 \times 500$ pixels with samples from two $G_0$ distributions: $G_0(-1.5, 0.5, 3.0)$ in the left half, and $G_0(-1.5, -1.0, -5, 3)$ in the right half, with $\alpha_2 \in \{-2.0, -1.0, -5.0, -3.0\}$. The two halves seem distinguishable after image
equalization in the first three cases, but the reader must notice that they have the same unitary mean. With this, techniques that rely on the classical notion of contrast are doomed to failure.

Instead of relying on the mean, the Gambini Algorithm aims at finding the point where the total log-likelihood of the samples is maximum. It is sketched in Algorithm 1.

**Algorithm 1: The Gambini Algorithm for edge detection using maximum likelihood**

**Data:** A strip of data \( z = (z_1, z_2, \ldots, z_N) \) with possibly an edge; an estimate of the number of looks \( \hat{L} \)

**Result:** An estimate of the edge position

Obtain a strip of data \( z \)

**for** \( j \) **in each position**

form the left sample \( z_L = (z_1, z_2, \ldots, z_j) \);

form the right sample \( z_R = (z_{j+1}, z_{j+2}, \ldots, z_N) \);

estimate \( (\hat{\alpha}, \hat{\gamma})_L(z_L) \);

estimate \( (\hat{\alpha}, \hat{\gamma})_R(z_R) \);

Compute the total log-likelihood using Eq. (30):

\[ tll(j) = \ell((\hat{\alpha}, \hat{\gamma})_L; \hat{L}, z_L) + \ell((\hat{\alpha}, \hat{\gamma})_R; \hat{L}, z_R) \]

**end**

Find the point \( j^* \) which maximizes the total log-likelihood.

**6.3 Edge detection by distance maximization**

The previous approach at edge detection maximizes the total likelihood of the two strips. In this sense, the total log-likelihood is the function to be optimized.

Another approach consists in maximizing a distance between the models to the right and to the left of the moving point. The rationale is the same as before: the models will be maximally different when the samples are not mixed, i.e., at the point which separates them into two different parts.

Algorithm 2 sketches an implementation of this approach.

**Algorithm 2: The Gambini Algorithm for edge detection using a distance between models**

**Data:** A strip of data \( z = (z_1, z_2, \ldots, z_N) \) with possibly an edge; an estimate of the number of looks \( \hat{L} \)

**Result:** An estimate of the edge position

Obtain a strip of data \( z \);

**for** \( j \) **in each position**

form the left sample \( z_L = (z_1, z_2, \ldots, z_j) \);

form the right sample \( z_R = (z_{j+1}, z_{j+2}, \ldots, z_N) \);

estimate \( (\hat{\alpha}, \hat{\gamma})_L(z_L) \);

estimate \( (\hat{\alpha}, \hat{\gamma})_R(z_R) \);

Compute a distance between the models:

\[ s(j) = s(f((\hat{\alpha}, \hat{\gamma})_L; \hat{L}), f((\hat{\alpha}, \hat{\gamma})_R; \hat{L})) \]

**end**

Find the point \( j^* \) which maximizes the distance between the models.

The following points are noteworthy in Algorithm 2:

- The samples \( z_L \) and \( z_R \) only enter once, when computing the estimates \( (\hat{\alpha}, \hat{\gamma})_L \) and \( (\hat{\alpha}, \hat{\gamma})_R \) respectively. This leads to faster algorithms than those based on the total log-likelihood, which use the sample twice.
- If the distance \( s \) belongs to either the class

\[ (a) \mathcal{G}(-20, 19, 3) \quad \text{and} \quad (b) \mathcal{G}(-1.5, 0.5, 3.0) \\
(b) \mathcal{G}(-10, 9, 3) \quad \text{and} \quad (d) \mathcal{G}(-1.5, 0.5, 3.0) \]

\[ (c) \mathcal{G}(-5, 4, 3) \quad \text{and} \quad (e) \mathcal{G}(-1.5, 0.5, 3.0) \]

\[ (d) \mathcal{G}(-3, 2, 3) \quad \text{and} \quad (f) \mathcal{G}(-1.5, 0.5, 3.0) \]

Fig. 8 Strips of 10 \times 500 pixels with samples from two \( \mathcal{G} \) distributions
of $h$-$\phi$ divergences or geodesic distances can be turned into a test statistic. With this, there is an important additional piece of information: the answer to the question, “Is there enough evidence to consider $j^*$ an edge?”

- Again, any model is suitable for this approach.

Naranjo-Torres et al.\cite{42} used the Shannon geodesic distance, obtaining remarkable results. They also used this distance to measure the separability of classes.

### 6.4 Parameter estimation by distance minimization

In this section, “contrast” is the lack of adequacy of the model with respect to the observed data.

Nascimento et al.\cite{23} computed several of the distances and statistical tests presented in Sec. 4.2 Being the null hypothesis that two samples come from the same $\mathcal{G}^0$ law, they used these statistics as contrast measures.

The $h$-$\phi$ family of statistical tests is not the only way to measure the difference between samples. Still under the $\mathcal{G}^0$ model, Naranjo-Torres et al.\cite{42} computed the Shannon Geodesic distance and, using results from Ref. [41] derived another class of tests useful for edge detection. Frery and Gambini advanced these results\cite{43}, proposing tests based on the composition of marginal evidence. These tests proved their usefulness in the detection of edges, even in situations of very low contrast.

Noting that the histogram is a stable density estimator, Gambini et al.\cite{51} proposed an estimation technique which consists in minimizing a distance between the $\mathcal{G}^0$ density and a smoothed version of the histogram. The authors analyzed several smoothing techniques and distances obtained from the symmetrization of $h$-$\phi$ divergences. The proposal consists of using kernel smoothing, and the triangular distance.

Fig. 10 illustrates the procedure. The density that characterizes the model $Z \sim \mathcal{G}^0(-2,1,3)$ is shown in light blue, along with the histogram of a sample of size 500 from $Z$. The two black densities are the starting point (dashed line), and an intermediate step. The final estimate is the density in red.

Notice that this procedure may be used with any model.

### 6.5 Speckle reduction by nonlocal means

In this section, the “contrast” is inversely proportional to the similarity between two samples.

The Nonlocal Means approach is a new way of handling noise reduction\cite{52,53}. The idea is forming a large convolution mask that, differently from the classical location-invariant approach, is computed for every pixel to be filtered. The weight at position $(k, \ell)$ is proportional to a measure of similarity between the observation $z_{k,\ell}$ and $z_{i,j}$, the pixel to be filtered. Such similarity is usually computed using the information in surrounding estimation windows (patches) $\partial_{k,\ell}$ and $\partial_{i,j}$ around the pixels. The procedure typically spans a large (search) window, much larger than in typ-
Fig. 11 illustrates this idea with a search window of size 9 × 9 around the central pixel $z_{i,j}$, and patches of size 3 × 3.

The central weight of the convolution mask, $w_{i,j}$, is set to 1. The weight $w_{k,\ell}$ will be a measure of the similarity between the values around $z_{k,\ell}$ and $z_{i,j}$, highlighted with dark lines. Such similarity may be computed directly between the observations, i.e., $w_{k,\ell} = f(z_{i-m,j-n}, z_{k-m,\ell-n}; -1 \leq m, n \leq 1)$, or between descriptors of each patch: $w_{k,\ell} = g(f(z_{i-m,j-n}; -1 \leq m, n \leq 1), f(z_{k-m,\ell-n}; -1 \leq m, n \leq 1))$.

Teuber and Lang[54] studied the properties of weights designed for the multiplicative model and how they behave in the presence of additive noise. Penna and Mascarenhas[55] included a Haar decomposition of the weights.

Ferraioli et al. [56] took a different approach. Instead of measuring the distance between the patches samples, they assumed a Square-Root-of-Gamma distribution for the data. If $Z$ follows a Gamma distribution as in Eq. (19), then $Z^{1/2}$ is Square-Root-of-Gamma distributed. Furthermore, the ratio of two independent of such random variables follows an Amplitude $\mathcal{G}^{0}$ distribution. Using this property, the authors checked how pertinent is the hypothesis that the two samples belong to the same distribution and used this information to compute the weights.

![Fig. 11 Illustration of the Nonlocal Means approach](image)

The Nonlocal Means approach is flexible enough to allow using any conceivable model for the data, and many ways of comparing the samples. In Section 8.1, we comment on applications of this technique for PolSAR data.

### 7 Polarimetric Data

Polarimetric data are more challenging than their intensity counterparts. A fully polarimetric observation in a single frequency is a $3 \times 3$ complex Hermitian positive definite matrix. The basic model for such observation is the Wishart distribution, whose probability density function is

$$f(\mathbf{z}; \Sigma, L) = L^{3L/2} |\Sigma|^{L/2} \exp\{-L \text{Tr}(\Sigma^{-1} \mathbf{z})\}$$

where $L \geq 3$, $\Gamma_{3}(L) = \pi^{3} \prod_{i=0}^{2} \Gamma(L - i)$, and $\text{Tr}$ is the trace operator. We denote this situation $\mathbf{Z} \sim \mathcal{W}(\Sigma, L)$. Refs. [57,58] provide details and properties of this distribution.

The Wishart model is the polarimetric version of the Gamma distribution. Similarly to what was presented in Section 5, there are Polarimetric $\mathcal{K}$ and $\mathcal{G}^{0}$ distributions. The reader is referred to Ref. [59] for a survey of models for Polarimetric SAR data.

Consider the sample $\mathbf{Z} = \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ from the $\mathcal{W}(\Sigma, L)$ distribution. Its reduced log-likelihood is given by

$$\ell(\Sigma, \mathbf{Z}) = 3L \ln L + \frac{L}{n} \sum_{i=1}^{n} |\mathbf{Z}_i| - L \ln |\Sigma| - \ln \Gamma_{3}(L) - \frac{L}{n} \sum_{i=1}^{n} \text{Tr}(\Sigma^{-1} \mathbf{Z}_i)$$

The maximum likelihood estimator of $(\Sigma, \mathbf{Z})$, based on the sample $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ under the Wishart can be obtained either by maximizing Eq. (32), or by solving $\nabla \ell = 0$, which amounts to computing:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i,$$

and the solution of

$$3 \ln \hat{L} + \frac{1}{n} \sum_{i=1}^{n} \ln |\mathbf{Z}_i| - \ln |\hat{\Sigma}| - \psi^{(0)}(\hat{L}) = 0$$
where $\psi^{(0)}_3$ is the zero-order term of the $\nu$-th order multivariate polygamma function $\psi^{(0)}_3(L) = \sum_{i=0}^{2} \psi^{(i)}(L - 1)$, and $\psi^{(i)}$ is the ordinary polygamma function:

$$\psi^{(i)}(L) = \frac{\partial^{i+1} \ln \Gamma(L)}{\partial L^{i+1}}$$

Numerical software as, for instance, Ref. [60] provides these special functions.

Section 6.1 presented two simple alternatives for the estimation of $L$ using intensity data under the Gamma model. The same estimation is more delicate when dealing with polarimetric data. The reader is referred to the work by Anfinsen et al. [63] for a comprehensive discussion of this topic.

Ref. [24] is one of the pioneering works dealing with measures of contrast between Wishart matrices. Frery et al. [62] derived explicit expressions of several tests based on $h$-$\phi$ divergences between these models, and applied them to image classification.

Analogously to $h$-$\phi$ divergences, it is possible to define $h$-$\phi$ entropies. These entropies can also be turned into test statistics with known asymptotic distribution [69]. Frery et al. [68] derived several of these entropies under the Wishart model and studied their finite-size sample behavior. Nascimento et al. [64], using these results, proposed change detection techniques for PolSAR imagery.

It is noteworthy that both $h$-$\phi$ divergences and entropies, under the Wishart model, solely rely on two operations over the covariance matrix $\Sigma$: Inversion $\Sigma^{-1}$, and the determinants $|\Sigma|$ and $|\Sigma^{-1}|$. Coelho et al. [60] obtained accelerated algorithms for computing them by exploiting their property of being Hermitian positive definite.

The Wishart distribution characterized by Eq. (31) can be posed in more general way, namely

$$f(z; \Sigma, L) = \frac{L^{pL}|z|^{L-p}}{|\Sigma|^{L/2}\Gamma_p(L)} \exp\{-L\text{Tr}(\Sigma^{-1}z)\}$$

where $p$ is the number of polarizations. With this, the dual- and quad-pol cases are contemplated by setting $p = 2$ and $p = 4$, respectively.

8 Applications to Polarimetric Data

8.1 Speckle reduction by nonlocal means

We already described the Nonlocal Means approach for speckle reduction in Section 6.5.

Torres et al. [66] used test statistics for computing the weights for polarimetric data filtering. The proposed approach consists of:

1. Assuming the Wishart model for the whole image.
2. Computing the maximum likelihood estimator in each patch.
3. Computing a stochastic distance between the Wishart models indexed by these estimates; the authors employed the Hellinger distance.
4. Turning the test statistic into a $p$-value and then applying a soft function. This yields the weights that, after scaling to add 1, are applied as a convolution mask.

This technique produced good results concerning noise reduction and details and radiometric preservation.

Later, Deledalle et al. [67] extended this idea to other types of SAR data. Chen et al. [68] used a likelihood ratio test, while Zhong et al. [69] employed a Bayesian approach.

As previously said Nonlocal Means filtering constitutes a very general framework for noise reduction. In particular, patches can be made adaptive to obtain further preservation of details [70].

8.2 Classification by minimum distance

This section describes an application where “contrast” is a measure of how different two samples are.

Silva et al. [71] proposed the following approach for PolSAR image classification:

1. Obtain training samples from the image using all available information. Call these samples prototypes, and with them, estimate the parameters of Wishart distributions.
2. Apply transformations to the original image in order to make it a suitable input for a readily available segmentation software.
(3) Tune the software parameters to obtain a segmentation with an excessive number of segments.

(4) Estimate the parameters of a Wishart distribution with each of these segments.

(5) Assign each segment to the prototype with minimum distance.

The advantage of this approach over other techniques is that steps (2) and (3) above can be applied with any available software, regardless of their appropriateness to PolSAR data.

Negri et al.[30] used stochastic distances to enhance the performance of region-based classification algorithms based on kernel functions. They also advanced the knowledge about the influence of specifying an erroneous (smaller) number of classes on the overall procedure.

Gomez et al.[72,73] also used such distances for PolSAR image classification using diffusion-reaction systems and machine learning.

8.3 Edge detection

Sections 6.2 and 6.3 described two techniques for edge detection based on the maximization of, respectively, the total log-likelihood and a distance between the distributions of the data. This idea was extended to the more difficult case of finding edges between regions of polarimetric data.

Nascimento et al.[74] proposed edge detectors using test statistics based on stochastic distances and compared the results of several distances with those produced by the total likelihood.

Edge detectors based on stochastic distances provide the same or better results than the one based on the Wishart likelihood, and at a lower computational cost.

This work showed that each function to be maximized (negative total log-likelihood or any of the distances) provides results which, if fused, may lead to an improved edge detector. This is the avenue that Ref. [75] starts to explore.

8.4 Polarimetric decomposition

The “contrast” here is a measure of how different the original data and the transformed sample are.

With the aid of stochastic distances between Wishart models, Bhattacharya et al.[76] proposed a modification of the Yamaguchi four-component decomposition. This technique alleviates problems related to urban areas aligned with the sensor line of sight and was incorporated into the freely available PolSARpro v. 6 educational software distributed by ESA, the European Space Agency.[1]

8.5 Change detection

“Contrast” here is proportional to the change experienced between samples along time.

Consider the situation of having two samples: \(Z^{(1)} = Z_1, Z_2, \ldots, Z_m\) and \(Z^{(2)} = Z_{m+1}, Z_{m+2}, \ldots, Z_{m+n}\) from the \(W(\Sigma, L)\), and \(H_1 : Z_1, Z_2, \ldots, Z_m\) come from \(W(S_1, L_1)\) while \(Z_{m+1}, Z_{m+2}, \ldots, Z_{m+n}\) come from \(W(S_2, L_2)\). The maximum likelihood estimators defined in Eq. (33) and (34) for these two situations are, thus, \((\hat{\Sigma}^{(0)}, \hat{\lambda}^{(0)})(Z_1, Z_2, \ldots, Z_{m+n})\) under \(H_0\), and \((\hat{\Sigma}^{(1)}, \hat{\lambda}^{(1)})(Z_1, Z_2, \ldots, Z_m)\) and \((\hat{\Sigma}^{(2)}, \hat{\lambda}^{(2)})(Z_{m+1}, Z_{m+2}, \ldots, Z_{m+n})\) under \(H_1\).

The log-likelihoods ratio statistic defined in Section 4.1 becomes, under the Wishart model using Eq. (32):

\[
3 \left[ \frac{\hat{\lambda}^{(1)} \ln \hat{\lambda}^{(1)} + \hat{\lambda}^{(2)} \ln \hat{\lambda}^{(2)} - \hat{\lambda}^{(0)} \ln \hat{\lambda}^{(0)}}{m} \right] + \frac{\hat{\lambda}^{(1)} \ln \sum_{i=1}^{m} |Z_i| + \hat{\lambda}^{(2)} \ln \sum_{i=m+1}^{m+n} |Z_i| - \hat{\lambda}^{(0)} \ln \sum_{i=1}^{m+n} |Z_i| - \hat{\lambda}^{(1)} \ln |\hat{\Sigma}^{(1)}| - \hat{\lambda}^{(2)} \ln |\hat{\Sigma}^{(2)}| - \hat{\lambda}^{(0)} \ln |\hat{\Sigma}^{(0)}| - \ln \Gamma_3 \left( \hat{\lambda}^{(1)} \right) - \ln \Gamma_3 \left( \hat{\lambda}^{(2)} \right) - \ln \Gamma_3 \left( \hat{\lambda}^{(0)} \right) - \hat{\lambda}^{(1)} \ln \sum_{i=1}^{m} \text{Tr} \left( (\hat{\Sigma}^{(1)})^{-1} Z_i \right) - \hat{\lambda}^{(2)} \ln \sum_{i=m+1}^{m+n} \text{Tr} \left( (\hat{\Sigma}^{(2)})^{-1} Z_i \right) - \hat{\lambda}^{(0)} \ln \sum_{i=1}^{m+n} \text{Tr} \left( (\hat{\Sigma}^{(0)})^{-1} Z_i \right) \right]
\]

---

1 https://step.esa.int/main/toolboxes/polsarpro-v6-0-biomass-edition-toolbox/
Conradsen et al.\cite{77} and Nielsen et al.\cite{78} based their change detection techniques on this test.

Notice that, differently from all tests based on stochastic distances, Eq. (36) allows its generalization to contrasting $H_0$ against an alternative that includes several samples $Z^{(1)}, Z^{(2)}, Z^{(3)}, Z^{(4)}, \ldots$ in which at least one comes from a different model than the rest. Such extension amounts to adding more terms to Eq. (36) depending on $(\hat{\mathbf{X}}^{(3)}, \hat{\mathbf{L}}^{(3)}), (\hat{\mathbf{X}}^{(4)}, \hat{\mathbf{L}}^{(4)}), \ldots$

The recent availability of a plethora of multi-temporal data leads to considering tests that allow checking whether a change has occurred at any instant of the sequence of observations.

Nascimento et al.\cite{38}, using the results presented in Section 4.3, devised a test statistic based on entropies for such situation. The advantage of such a test over the ratio of likelihoods is twofold, namely, (i) it is more general, as it may comprise any $h$-entropy, and (ii) it is faster.

9 Deterministic Geodesic Distances

A polarimetric observation can be projected onto the surface of a 16 dimensional sphere. Ratha et al.\cite{79} used this mapping, known as Kennaugh representation, to measure geodesic distances between observations and prototypical backscatterers. We call these measures of dissimilarity deterministic because the reference, i.e., the prototypes, is a point rather than a model. The first application targeted unsupervised classification, but they were later used to devise new built-up\cite{80} and vegetation\cite{81} indexes. Such distances are also useful for unsupervised classification in an arbitrary number of classes\cite{79}. These ideas led to the proposal of a new decomposition theorem\cite{82}.

A recent development\cite{83} provides evidence that the distance between carefully selected samples and their closest elementary backscatterer follows a Beta distribution.

10 Summary, Conclusions and Research Avenues

This review presented, in an unified manner, applications of measures of contrast to problems that arise in SAR and PolSAR image processing and analysis. Those problems are: filter assessment (Section 2), edge detection (Sections 6.2, 6.3 and 8.3), parameter estimation (Section 6.4), noise reduction (Sections 6.5 and 8.1), segments classification (Section 8.2), polarimetric decomposition (Section 8.4), and change detection (Section 8.5).

Statistical tests derived from divergences, entropies, and geodesic distances are beginning to be exploited by the Remote Sensing community. These techniques have several important properties, among them excellent performance, even with small samples. Such feature might be useful in a problem that we did not tackle yet, namely, target detection.

Among the questions which, to this date, are open, one may list the following:

- Is there any gain in using other less known divergences?
- How is the behavior of change detection techniques based on quadratic differences of entropies other than the Shannon and Rényi, e.g., Arimoto and Sharma-Mittal?
- The asymptotic properties of these tests require the use of maximum likelihood estimators. How do they behave when other inference techniques, e.g., robust estimators, estimators by analogy, are used?
- How do these techniques perform when there is no explicit model, and it is inferred from the data with, for instance, kernel estimators?
- How do these test statistics perform under several types of contamination?

An important aspect that has not been included in this review is the spatial correlation. Such a feature is responsible for, among other characteristics, a higher-level perception of texture. We refer the reader to the work by Yue et al.\cite{84} for a modern model-based approach to texture description and simulation in SAR imagery.

The approach based on deterministic distances is fairly new, and it has already shown excellent performance. This kind of data representation in terms of their dissimilarity to prototypical backscatterers is flexible and extensible, as it accepts any number of reference points as input.

We consider that one of the most significant challenges in this area is integrating these tech-
niques with production software as, for instance, PolSARpro and SNAP.

As a final thought, one may ask why bothering with such modeling in a scenario where the Deep Learning approach is taking over. The answer is simple. The statistical approach provides more than results: if well employed, it produces answers that can be interpreted. This is a significant advantage, which, in turn, is not orthogonal to the other; statistical evidence may constitute an information layer to be processed by neural networks.

Reproducibility

We used Ref. [60] v. 3.1 to produce the plots and statistical analyses, on a MacBook Pro with 3.5 GHz Dual-Core Intel Core i7 processor, 16 GB of memory running Mac OS Catalina v. 10.15.2 operating system.

The code, data, and \LaTeX{} sources are available upon request to the author.

Acknowledgements

This work is only possible thanks to my coauthors’ talent, patience, and kindness. Listing all of them will make me incur unfair omissions, so I prefer to let the reader check on the cited papers the names of those who were instrumental for the development of this body of research.

References


[43] FRERY A C and GAMBINI J. Comparing samples from the $G_I^0$ distribution using a geodesic distance.[J]. TEST,


TORRES L, SANT’ANNA S J S, DA COSTA FREITAS C,


[83] FERNANDES D and FRERY A C. Statistical properties of geodesic distances between samples and elementary scatterers in PolSAR imagery[C]. 2019 TENGARSS, Kochi, India, 2019, in press.

Author Introduction

Alejandro C. Frery (S’92–SM’03) received a B.Sc. degree in Electronic and Electrical Engineering from the Universidad de Mendoza, Mendoza, Argentina. His M.Sc. degree was in Applied Mathematics (Statistics) from the Instituto de Matemática Pura e Aplicada (IMPA, Rio de Janeiro) and his Ph.D. degree was in Applied Computing from the Instituto Nacional de Pesquisas Espaciais (INPE, São José dos Campos, Brazil). He is currently the leader of LaCCAN – Laboratório de Computação Científica e Análise Numérica, Universidade Federal de Alagoas, Maceió, Brazil, and holds a Huashan Scholar position (2019–2021) with the Key Lab of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, Xi’an, China. His research interests are statistical computing and stochastic modeling.