Dictionary-based probability density function estimation for high-resolution SAR data

Vladimir Krylov, Gabriele Moser, Sebastiano B. Serpico, Josiane Zerubia

Dictionary-based probability density function estimation for high-resolution SAR data

Vladimir Krylov*[a,c], Gabriele Moser*[b], Sebastiano B. Serpico[b], Josiane Zerubia[c]
[a] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, 119991 Leninskie Gory, Moscow (Russia);
[b] Dept. of Biophysical and Electronic Engineering (DIBE), University of Genoa, Via Opera Pia 11a, I-16145, Genoa (Italy);
[c] EPI Ariana, UR INRIA Sophia Antipolis Méditeranée, 2004, Route des Lucioles, B.P.93, FR-06902, Sophia Antipolis (France).

ABSTRACT

In the context of remotely sensed data analysis, a crucial problem is represented by the need to develop accurate models for the statistics of pixel intensities. In this work, we develop a parametric finite mixture model for the statistics of pixel intensities in high resolution synthetic aperture radar (SAR) images. This method is an extension of previously existing method for lower resolution images. The method integrates the stochastic expectation maximization (SEM) scheme and the method of log-cumulants (MoLC) with an automatic technique to select, for each mixture component, an optimal parametric model taken from a predefined dictionary of parametric probability density functions (pdf). The proposed dictionary consists of eight state-of-the-art SAR-specific pdfs: Nakagami, log-normal, generalized Gaussian Rayleigh, Heavy-tailed Rayleigh, Weibull, K-root, Fisher and generalized Gamma. The designed scheme is endowed with the novel initialization procedure and the algorithm to automatically estimate the optimal number of mixture components. The experimental results with a set of several high resolution COSMO-SkyMed images demonstrate the high accuracy of the designed algorithm, both from the viewpoint of a visual comparison of the histograms, and from the viewpoint of quantitive accuracy measures such as correlation coefficient (above 99,5%). The method proves to be effective on all the considered images, remaining accurate for multimodal and highly heterogeneous scenes.

Keywords: Synthetic aperture radar (SAR) image, probability density function (pdf), parametric estimation, finite mixture models, stochastic expectation maximization (SEM)

1. INTRODUCTION

Synthetic aperture radar (SAR) is an active imagery system working in the microwave band (such as, for instance, the X-band that corresponds to frequencies in range from 7 to 12.5 GHz) thus operational regardless of weather conditions and time of the day. SAR images are becoming widely used nowadays in various applications, e.g. in flood/fire monitoring, agriculture assessment, urban mapping. Modern SAR systems, e.g. italien

COSMO-SkyMed or german TerraSAR-X, are capable of providing very high resolution images (up to 50 cm of land resolution). In the context of remote sensing, a crucial problem is represented by the need to develop accurate models for the statistics of the pixel intensities. Focusing on SAR [3][4][20][25] data, this modeling process turns out to be a crucial task, for instance, for classification [7] or for denoising [20] purposes.

*e-mail: V. Krylov - vl_krylov@mail.ru; G. Moser - gemini@unige.it

In this paper, we address the problem of probability distribution function (pdf) estimation in the context of SAR amplitude data analysis. Specifically, several different theoretic and heuristic models for the pdfs of SAR data have been proposed in the literature, and have proved to be accurate for specific land-cover typologies. However, the existing variety of parametric pdf families specific to some particular types of landcover makes the choice of a single pdf family a hard task. A remotely sensed image, in general, can depict a varied scene, jointly presenting several distinct land cover typologies. Moreover, high resolution images usually present additional complications in the form of more sophisticated histograms.

In this paper, we develop an algorithm based on the dictionary approach, already successfully adopted in [15]; this approach has shown very good results for lower resolution images, as reported in [15]. In this research, however, we are interested in high resolution images (from new SAR systems like COSMO-SkyMed and TerraSAR-X, launched starting from 2007); these images are likely to present more complicated properties, due to the higher level of details. Our aim is to further develop the dictionary-based approach and verify its accuracy for high resolution images.

Thus, we address the SAR statistics estimation problem by adopting a finite mixture model (FMM) [23][24] for the amplitude pdf, i.e., by postulating the unknown amplitude pdf to be a linear combination of parametric components, each one corresponding to a specific land cover type [5][22]. In order to take explicitly into account the possible differences in the statistics of the mixture components, we avoid choosing *a priori* a specific parametric family for each component, but we allow each of them to belong to a given dictionary of SAR-specific pdfs. Working within the FMM framework we further develop the approach in [15] by enlarging the dictionary and improving the estimation scheme.

Specifically, the proposed algorithm automatically integrates the procedures of selecting the optimal model for each component and parameter estimation, by combining the stochastic expectation maximization (SEM) algorithm [1][2][13] and the method-of-log-cumulants (MoLC) [18][19]. The former is a stochastic iterative estimation methodology, dealing with problems of data incompleteness, developed as an improvement of the classic expectation-maximization algorithm [6][23] in order to increase its capability to compute maximum likelihood (ML) estimates [29]. The latter is a recently proposed estimation approach originating from the adoption of the Mellin transform [26] (instead of the more common Fourier transform) to the computation of characteristic functions, and from the corresponding generalization of the concepts of moments and cumulants [21]. We adopt this method both for its good estimation properties [17][18][27] and because it turns out to be feasible and fast for all the parametric families in the dictionary [15][17]. Whereas the well-known ML and the method-of-moments (MoM) estimation strategies [11][18] involve numerical difficulties for several parametric families from the dictionary [16][18][20]. However, the contribution of this paper lies in the introduction of the new procedure of automatic estimation of the number of mixture components: contrary to [15], the designed scheme integrates this estimation into the SEM iterative procedure, thus avoiding the cumbersome repetition of SEM several times for mixtures with different number of components. We also introduce the algorithm for SEM initialization, which is justified by the properties of pdfs in our dictionary and also provides accurate results.

The proposed parametric approach is validated on several high resolution COSMO-SkyMed images. The experimental results show the capability of the developed algorithm to accurately model the amplitude distribution of all the considered images, both from a qualitative viewpoint (i.e., visual comparison between the data histogram and the estimated pdf) and from a quantitative viewpoint (i.e., correlation coefficient, Kolmogorov-Smirnov distance between the data histogram and the estimated pdf), thus showing its effectiveness and flexibility.

The paper is organized as follows: in Section 2 and its subsections we present the FMM-based estimation scheme, SEM approach, MoLC methods, number of components estimation and some other aspects of the developed

algorithm. Section 3 reports the results of the application of the proposed approach to the statistical modeling of the grey-level of several real SAR images, showing the method's capabilities of fitting the amplitude distribution more efficiently than previously proposed parametric models for SAR amplitude data. Finally, conclusions are drawn in Section 4.

2. DICTIONARY-BASED FMM APPROACH

2.1 Finite Mixture Model approach

In order to formalize the common scenario when several distinct land-cover typologies are present in the same SAR image, we assume a finite mixture model (FMM) [23][24] for the distribution of grey levels. Specifically, we model the SAR image as a set $\mathcal{I} = \{r_1, r_2, \ldots, r_N\}$ of independent and identically distributed (i.i.d.) samples drawn from the mixture pdf:

$$p_r(r|\theta) = \sum_{i=1}^{K} P_i p_i(r|\theta_i), \quad r \geq 0, \tag{1}$$

where $p_i(r|\theta_i)$ are pdfs dependant on vectors θ_i of parameters, taking values in a set $\Theta_i \subset \mathbb{R}^{\ell_i}$ and $\{P_1, P_2, \ldots, P_K\}$ is a set of mixing proportions $\sum_{i=1}^{K} P_i = 1$ with $0 \leq P_i \leq 1, \quad i = 1, 2, \ldots, K$. Thus the aim is to estimate the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_K; P_1, P_2, \ldots, P_K)$. This so-called "i.i.d. approach" is widely accepted in the context of estimation theory [7][9] and corresponds to discarding the contextual information associated to the correlation between neighboring pixels in the image during the estimation process, thus exploiting only the greylevel information.

Each component of the $p_i(r|\theta_i)$ in (1) is modeled by resorting to a finite dictionary $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ (see Table 1) of $M = 8$ SAR-specific distinct parametric pdfs $d_i(x|\alpha_i)$, parameterized by vectors $\alpha_i \in A_i, i = 1, \ldots, M$. For descriptions of distributions and their physical properties see [10]. The method is designed to integrate and automate the number of mixture component estimation, the selection for each mixture component of the optimal model inside the dictionary and the parameter estimation for each mixture component.

2.2 Stochastic Expectation Maximization

As discussed in [10][15], considering the variety of estimation approaches for FMMs a reasonable choice for our application is the stochastic expectation maximization (SEM) scheme [2]. Thanks to the stochastic sampling involved in this scheme, we gain numerical tractability along with the better exploring capabilities as compared to EM scheme, thus higher chances of finding the global maximum. However, the sequence of parameter estimates generated by SEM is a discrete time random process that does not converge pointwise nor almost surely; it has been proved to be an ergodic and homogeneous Markov chain converging to a unique stationary distribution, which is expected to be concentrated around the global maxima of the log-likelihood function [2].

SEM is an iterative estimation scheme dealing with the problem of *data incompleteness*. This incomplete data can in general be an unobserved part of the data or either corrupt data. The incompleteness issue is formalized by assuming a "complete" data vector x to be unavailable and observable only through an "incomplete" data vector $y = \Phi(x)$ obtained by a many-to-one mapping $\Phi : X \to Y \subset \mathbf{R}^m$. Thus, a given realization y of the

Table 1: Pdfs and MoLC equations for the parametric families included in the considered dictionary $\mathcal{D}$. Here $\Gamma(\cdot)$ is the Gamma function [26], $K_\alpha(\cdot)$ the αth order modified Bessel function of the second kind [26], $J_0(\cdot)$ is the zero-th order Bessel function of the first kind [26], $\Psi(\cdot)$ the Digamma function [15], $\Psi(\nu,\cdot)$ the νth order polygamma function [15] and $G_\nu(\cdot)$ is the specific integral function for GGR [16].

Family	Distribution function	MoLC equations
Log-normal	$f_1(r\|m,\sigma) = \frac{1}{\sigma r\sqrt{2\pi}} \exp\left[-\frac{(\ln r - m)^2}{2\sigma^2}\right],$	$\kappa_1 = \beta$ $\kappa_2 = V.$
Weibull	$f_2(r\|\eta,\mu) = \frac{\eta}{\mu^\eta} r^{\eta-1} \exp\left[-\left(\frac{r}{\mu}\right)^\eta\right],$	$\kappa_1 = \ln b + \Psi(1)c^{-1}$ $\kappa_2 = \Psi(1,1)c^{-2}.$
Fisher	$f_3(r\|L,M,\mu) = \frac{\Gamma(L+M)}{\Gamma(L)\Gamma(M)} \frac{L}{M\mu} \frac{\left(\frac{Lr}{M\mu}\right)^{L-1}}{\left(1+\frac{Lr}{M\mu}\right)^{L+M}},$	$\kappa_1 = \ln\mu + (\Psi(L) - \ln L) - (\Psi(M) - \ln M)$ $\kappa_2 = \Psi(1,L) + \Psi(1,M)$ $\kappa_3 = \Psi(2,L) - \Psi(2,M).$
Generalized Gamma (GGamma)	$f_4(r\|\nu,\sigma,\kappa) = \frac{\nu}{\sigma\Gamma(\kappa)} \left(\frac{r}{\sigma}\right)^{\kappa\nu-1} \exp\left\{-\left(\frac{r}{\sigma}\right)^\nu\right\},$	$\kappa_1 = \Psi(\kappa)/\nu + \ln\sigma$ $\kappa_2 = \Psi(1,\kappa)/\nu^2$ $\kappa_3 = \Psi(2,\kappa)/\nu^3.$
Nakagami	$f_5(r\|L,\mu) = \frac{2}{\Gamma(L)} \left(\frac{L}{\mu}\right)^L r^{2L-1} \exp\left(-\frac{Lr^2}{\mu}\right),$	$2\kappa_1 = -\ln\lambda + \Psi(L) - \ln L$ $4\kappa_2 = \Psi(1,L).$
K-root	$f_6(r\|L,M,\mu) = \frac{4}{\Gamma(L)\Gamma(M)} \left(\frac{LM}{\mu}\right)^{(L+M)/2} \times$ $\times r^{L+M-1} K_{M-L}\left[2r\left(\frac{LM}{\mu}\right)^{1/2}\right],$	$2\kappa_1 = -\ln\lambda + \Psi(L) - \ln L + \Psi(M) - \ln M$ $4\kappa_2 = \Psi(1,L) + \Psi(1,M)$ $8\kappa_3 = \Psi(2,L) + \Psi(2,M).$
Generalized Gaussian Rayleigh	$f_7(r\|\lambda,\gamma) = \frac{\gamma^2 r}{\lambda^2\Gamma^2(\lambda)} \times$ $\times \int_0^{\pi/2} \exp[-(\gamma r)^{1/\lambda}(\|\cos\theta\|^{1/\lambda} + \|\sin\theta\|^{1/\lambda})] d\theta,$	$\kappa_1 = \lambda\Psi(2\lambda) - \ln\gamma - \lambda G_1(\lambda)[G_0(\lambda)]^{-1}$ $\kappa_2 = \lambda^2\Psi(1,2\lambda) + \lambda^2 G_2(\lambda)[G_0(\lambda)]^{-1} -$ $-\lambda^2[G_1(\lambda)]^2[G_0(\lambda)]^{-2}.$
SαS	$f_8(r\|\alpha,\gamma) = r\int_0^{+\infty} \rho \exp(-\gamma\rho^\alpha) J_0(r\rho) d\rho,$	$\alpha\kappa_1 = \Psi(1)(\alpha-1) + \alpha\ln 2 + \ln\gamma$ $\kappa_2 = \Psi(1,1)\alpha^{-2}.$

incomplete data may have been generated by any realization $x \in \Phi^{-1}(y)$; this does not allow, for instance, a direct computation of an ML estimate. SEM tries to avoid these difficulty by iteratively and randomly sampling a complete data set and using it to compute a standard ML estimate.

Specifically, we regard the FMM approach as being affected by a data incompleteness problem, since it is not known from which of the available statistical populations (corresponding directly to mixture components) involved in (1) a given image sample has been drawn. This implicitly means that no training information about the possible land-cover types in the SAR image is exploited in the estimation process, i.e., the SAR amplitude pdf estimation problem is addressed in an unsupervised context. In this manner, the complete data for FMM is represented by the set $\{(r_i, s_i), i = 1,\ldots,N\}$, where r_i is the observed part (SAR amplitude) and s_i the lacking data (label). Given an FMM with K components, s_i takes value in $\{1,\ldots,K\}$ and denotes to which out of the K components the i-th pixel belongs.

The classic SEM procedure consists of 3 steps: 1)Expectation step (E-step), here the probabilities for missing information are calculated based on the information collected on the previous iteration, 2)Stochastic step (S-

step), where we sample the missing information with respect to the distributions estimated on E-step, and 3)Maximization step (M-step), where we estimate the parameters via maximization. For detailed description see in [2], [15], [10].

A crucial point in the FMM estimation is the choice of the number of components. Several different validation functionals have been proposed in the literature as criteria to select the best value K^* of the parameter K, e.g., discriminant analysis [14], the minimum message length [8]. Unfortunately both approaches appear to be far from efficient in our application due to the strong overlapping between the statistics of distinct components in real SAR data [10]. In [10], this estimation was done by rerunning the whole procedure for every possible value of K from 1 to predefined K_{max} and then choosing K with the highest likelihood. Here, we suggest a much faster approach: we initialize SEM with un upper bound $K_0 > K^*$ of the number of components; then, if after the t-th iteration the value P_i^{t+1} goes below some predefined threshold, we disregard this component and at the $(t+1)$-th iteration we work with $(K-1)$ components. The value of this threshold should be fairly low, it can be set equal to 0 for simplicity (it would mean that literally no gray values have been put to some particular component). Analytically, this approach does not violate the SEM procedure and it has provided accurate results. Thus we integrate the K-step where we control values P_i^{t+1} into SEM and further refer to this modification as KSEM.

One more critical issue of any nonconvergent iterative scheme is to define the number of iterations. As mentioned above, the SEM sequence of pdf estimates is expected to reach a stationary behavior, so in order to stop it we apply the following stationarity criterion:

$$\sum_{t=t_0-k+1}^{t} \left(\sum_{i=1}^{K} |P_i^t - P_i^{(t-1)}| \right) < \alpha, \tag{2}$$

where K is the number of components, P_i^t are the mixing proportions on the t-th iteration as in (1). So we control the proportion of pixels being reallocated into some other component during the previous k iterations and once this portion goes below some level α on the t_0-th iteration we stop the algorithm. The value of α can be tricky to estimate, it might need some tuning. So the reasonable idea would be to mix this stop condition with the one in [15] (predefinition of the maximum number of iteration) and iterate until one of them is fulfilled.

2.3 Parameter estimation. Method of Log-Cumulants

As mentioned above, the M-step of SEM at each iteration involves computation of the optimal parameter vector θ^{t+1} by an ML procedure. This approach turns out unfeasible for several SAR-specific pdfs, such as, e.g., the K distribution [20]. Hence, we avoid using ML estimates and at the M-step we adopt the MoLC approach [15][28], which has been proven to be a feasible and effective estimation tool for common SAR parametric models [28] and also for all pdfs in our dictionary [12][15].

MoLC has recently been proposed as a parametric pdf estimation technique suitable for distributions defined on $[0, +\infty)$, it has been widely applied in the context of SAR-specific parametric families for amplitude and intensity data modeling, e.g., the Nakagami and K distributions [28]. MoLC is based on the generalization of the usual moment-based statistics by using the Mellin transform [26] for the computation of characteristic and moment generating functions, instead of the common Fourier transform, and allows stating a set of equations relating the unknown parameters of a given parametric model with one or more logarithmic cumulants (*log-cumulants*). The solution of such equations allows computing the desired parameter estimates [28]. These

equation have one solution for any observed log-cumulants for all of the pdfs in $\mathcal{D}$, except for, in some cases, K-root and GGR. For further details about the mathematical formulations of MoLC we recommend [10],[28].

Thus, we substitute the M-step in the SEM by two steps: 1)the "MoLC step", where for every component we estimate the parameters for all M models from the dictionary, and then on the 2)"Model Selection step" (MS-step) where we pick the pdf that provides the highest value of likelihood.

Finally, in order to reduce the computation time of the proposed method, we apply a histogram-based version [15] of the algorithm. The idea is to group the pixels by values of their intensities (i.e. on 8 bpp image we get 256 groups) rather than working with every pixel separately.

The structure of the resulting algorithm is as follows:

- **E-step:** compute, for each greylevel z and i-th component, the posterior probability estimates corresponding to the current pdf estimates, i.e. ($z = 0, 1, \ldots, Z-1$, $i = 1, 2, \ldots, K$):

$$\tau_i^t(z) = \frac{P_i^t p_i^t(z)}{\sum_{j=1}^{K} P_j^t p_j^t(z)}; \tag{3}$$

- **S-step:** sample the label $s^t(z)$ of each greylevel z according to the current estimated posterior probability distribution $\{\tau_i^t(z) : i = 1, 2, \ldots, K\}$ ($z = 0, 1, \ldots, Z-1$);

- **MoLC-step:** for the i-th mixture component ($i = 1, \ldots, K$), compute the following histogram-based estimates of the mixture proportion and of the first three log-cumulants:

$$P_i^{t+1} = \frac{\sum_{z \in Q_{it}} h(z)}{\sum_{z=0}^{Z-1} h(z)}, \quad \kappa_{1i}^t = \frac{\sum_{z \in Q_{it}} h(z) \ln z}{\sum_{z \in Q_{it}} h(z)}, \kappa_{bi}^t = \frac{\sum_{z \in Q_{it}} h(z)(\ln z - \kappa_{1i}^t)^b}{\sum_{z \in Q_{it}} h(z)}, \tag{4}$$

  where $b = 2$ or $b = 3$, $Q_{it} = \{z : s^t(z) = \sigma_i\}$ is the set of grey levels assigned to the i-th component ($i = 1, 2, \ldots, K$); then, solve the corresponding MoLC equations (see Table 1) for each parametric family $f_j(\cdot|\alpha_j)$ ($\alpha_j \in A_j$) in the dictionary, thus computing the resulting MoLC estimate $\alpha_{ij}^t (i = 1, 2, \ldots, K,\ j = 1, 2, \ldots, M)$;

- **K-step:** if for some i ($i = 1, 2, \ldots, K$): $P_i^t < threshold$, then $K = K - 1$;

- **MS-step:** for the i-th mixture component, compute the log-likelihood of each estimated pdf $f_j(\cdot|\alpha_{ij}^t)$ (except, maybe, GGR or K-root if the previous step yielded no solutions for the corresponding MoLC equations [16][28]) according to the data assigned to the i-th component:

$$L_{ij}^t = \sum_{z \in Q_{it}} h(z) \ln f_j(z|\alpha_{ij}^t) \tag{5}$$

  and define $p_i^{t+1}(\cdot)$ as the estimated pdf $f_j(\cdot|\alpha_{ij}^t)$ yielding the highest value of L_{ij}^t ($i = 1, 2, \ldots, K,\ j = 1, 2, \ldots, M$).

In this paper, we suggest a novel initialization procedure for SEM. It can be done as in [15] via a random initialization, however this results in costly burn-in iterations. Here we propose initialization based on the form of the histogram. Thus, first we find the number and locations of the histogram modes. The common way to do it is smoothing: the choice of a specific histogram smoothing procedure depends on the complexity of the histogram

(specifically, on the level of noise-like behavior), however for all our test images the linear smoothing with size 10-20 provided accurate estimation of the number of modes. Then we initialize the components with respect to the positions of the modes: on the grey levels corresponding to every mode of the histogram we randomly initialize several components (depending on the level of competitiveness). This level of competitiveness directly corresponds to the complexity of this particular image: if we expect to find many distinct types of land-cover there, this parameter should be set higher, say 3-4; whereas for the most of the cases this parameter could be set 2 without any loss in accuracy. The higher this parameter is set the more "thoroughly" the algorithm will tries to find the components in the image. Such an approach is justified by the observation that all the distributions in our dictionary $\mathcal{D}$ have single-mode pdfs. We remind here that the initial estimate K of the number of components for KSEM should be an upper bound. With sufficient level of competitiveness the upper bound condition is obviously met.

3. EXPERIMENTAL RESULTS

3.1 Data set for experiments

The proposed KSEM algorithm for pdf estimation has been tested on a set of real SAR images, and compared with several common SAR-specific parametric pdf models. The tests were run on 16 bpp singlelook X-band SAR-images with 5 m spatial resolution, acquired in 2008 over the region of Piemonte (Italy) by one of the COSMO-SkyMed satellites (©Italian Space Agency, ASI):

- "River_ponds", 1400×1400 pixels;
- "Small_river", 300×500 pixels;
- "Mountain_lake1" and "Mountain_lake2", 400×400 pixels;
- "Mountain_town", 2200×1700 pixels;
- "Mountain", 3000×1400 pixels

A further test image was acquired by the RAMSES airborne sensor (700×700 pixels, 8 bpp); we refer to it as "Ramses" for simplicity. This image was provided to us by the French Space Agency (©ONERA-CNES).

3.2 Estimation results

We stress that some of the images exhibit bimodal histograms, whereas the other ones have fairly simple unimodal histograms. The proposed KSEM method has been applied to all the considered images and the resulting pdf estimates have been assessed both quantitatively, by computing their correlation coefficients with the image histograms (see Table 2), and qualitatively, by visually comparing the plots of the estimates and of the histograms (see Fig. 1).

When we deal with 16 bpp images, in order to avoid cumbersome computations, basing on the nature of the data, we apply the following trick: instead of working with the whole 100% bins of the histogram, we work

only with 99,9% of the data, i.e. the part located within the 99,9%th quantile. This almost does not affect the estimation accuracy sharply reducing the computation burden. In fact for all the images in our test set, the histogram was compactly concentrated close to the origin of coordinates, i.e., close to zero, so usually the 99,9%th quantile is well below 1500, thus redusing the number of bins under study from 65536 to around 1000, thus reducing the computation weight 65 times.

Table 2: Results of the aplication of KSEM to all the employed SAR images: correlation coefficient (ρ) and Kolmogorov-Smirnov distance ($KS-dist$) between the estimated pdf and the image histogram, the estimated number K of mixture components and the pdfs in the mixture. Results for single component approximation: the best fitting model and the correlation coefficient.

Image	KSEM				Best single component	
	ρ	$KS-dist$	K	Mixture	Best model	ρ
River_ponds	99,95%	0,0001	3	(f_1, f_3, f_4)	Weibull	99,07
Small_river	99,80%	0,0067	5	$(f_4, f_5, f_2, f_1, f_1)$	GGamma	98,74
Mountain_lake1	99,91%	0,0011	4	(f_2, f_4, f_4, f_3)	Lognormal	86,20
Mountain_lake2	99,90%	0,0008	4	(f_4, f_4, f_4, f_3)	GGamma	97,39
Mountain_town	99,89%	0,0049	4	(f_1, f_3, f_4)	Nakagami	97,86
Mountain	99,80%	0,0017	3	(f_1, f_4, f_3)	Lognormal	99,41
Ramses	99,62%	0,0070	4	(f_6, f_2, f_1, f_4)	GGR	94,21

The correlation coefficients between the resulting estimated pdfs and the image histograms are very high (always greater than 99,5%) for all the considered images, thus showing the effectiveness of the proposed method from the viewpoint of the estimation accuracy. The visual comparison between the pdf estimates and the corresponding image histograms confirms this conclusion, as shown, for example, in Fig. 1.

In order to further assess the capabilities of the method, a comparison has also been performed with several other standard parametric models for SAR amplitude data. Specifically, we present here the result of the best-performing model from our dictionary $\mathcal{D}$ (see Table 2), the corresponding parameters for all models from $\mathcal{D}$ have been estimated by MoLC. A comparison between KSEM and single component estimate shows that KSEM yields the pdf estimate with the highest correlation coefficients with the image histograms of all images. The result is especially significant for bimodal images (e.g. "Mountain_lake1"), when single component models fail altogether in accuracy.

4. CONCLUSIONS

In this paper, an efficient finite mixture model estimation algorithm has been developed for SAR amplitude data pdf. It integrates the previously existing dictionary-based approach [15] with an innovative initialization scheme and estimation procedure for the the number of components. In particular, the developed estimation strategy is explicitly focused on the context of SAR image analysis and correspondingly a set of eight theoretic or empirical models for SAR amplitude data (i.e., Nakagami, log-normal, generalized Gaussian Rayleigh, SαS generalized Rayleigh, Weibull, K-root, Fisher and generalized Gamma) has been adopted as a dictionary.

The numerical results of the application of the method to several real SAR images acquired by the COSMO-SkyMed and airborne RAMSES sensors prove the proposed KSEM algorithm to provide very accurate pdf

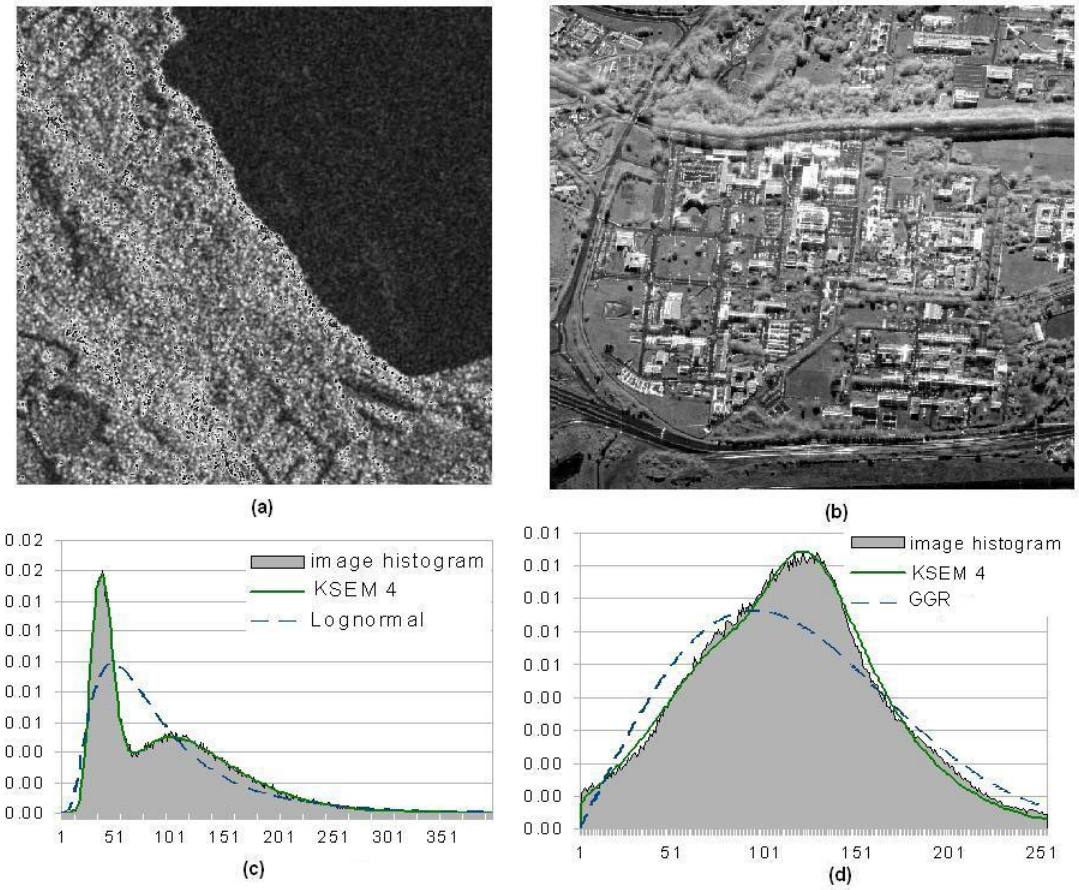

Figure 1: (a)"Mountain_lake1" (©ASI), (b) "Ramses" (©ONERA-CNES) images. Plot of the image histogram, KSEM pdf estimates and the best fitting pdf from the dictionary for the (c)"Mountain_lake1" image and (d)"Ramses" image.

estimates, both from the viewpoint of a visual comparison between the estimates and the corresponding image histograms, and from the viewpoint of the quantitative correlation coefficient between them. We stress, in particular, that the method proves to be effective on all the considered images, despite of their different statistics (i.e. histogram unimodality or multimodality), high heterogeneity. Correlation coefficients higher that 99% are obtained, in fact, in all cases, thus proving the flexibility of the method.

Specifically, the use of a mixture model is mandatory when dealing with multimodal statistics. Applied to "River_ponds", "Mountain_lake1" and "Mountain_lake2" images, which exhibit a bimodal histogram, the developed KSEM algorithm correctly detects positions and sizes of both statistical modes. On the other hand, the results provided by KSEM in case of unimodal histograms usually provide only minor improvement as compared to the best single-component parametric models included in the dictionary.

The proposed KSEM algorithm is completely automatic, by performing both the FMM estimation process and the optimization of the number of mixture components without any need for user interaction. In addition,

thanks to the specific histogram-based version of SEM it adopts, the computation time of KSEM is almost independent of the image size. These interesting operational properties, together with the estimation accuracy it provides for all the considered images prove KSEM to be a flexible and effective pdf estimation tool for high resolution or heterogeneous SAR data analysis. We stress that the possibility to refine or enlarge the dictionary with respect to specific sets of images further explains the potential of this method.

Acknowledgments

This research has been conducted within a collaboration between the research team ARIANA of the Institut National de Recherche en Informatique et Automatique (INRIA) Sophia Antipolis, France, and the Dept. of Biophysical and Electronic Engineering (DIBE) of the University of Genoa, Italy.

It was carried out with the financial support of French Space Agency (CNES), whose support is gratefully acknowledged. The authors would also like to thank the Italian Space Agency (ASI) for providing the COSMO-SkyMed images of Piemonte (©ASI, 2008) and CNES for providing the RAMSES image (©ONERA-CNES, 2004).

REFERENCES

[1] C. Biernacki, G. Celeux, and G. Govaert, *Strategies for getting highest likelihood in mixture models*, Research Report 4255, INRIA, September 2001.

[2] G. Celeux, D. Chauveau, and J. Diebolt, *On stochastic versions of the EM algorithm*, Research Report 2514, INRIA, March 1995.

[3] M. Cheney, *A mathematical tutorial on synthetic aperture radar*, SIAM Review **43** (2001), no. 2, 301–312.

[4] ———, *An introduction to synthetic aperture radar (SAR) and SAR interferometry*, pp. 167–177, in "Approximation theory X: wavelets, splines, and applications", C. K. Chui, L. L. Schumacher, and J. Stockler editors, Vanderbilt University Press, Nashville, TN, U.S.A., 2002.

[5] Y. Delignon and W. Pieczynski, *Modelling non-Rayleigh speckle distribution in SAR images*, IEEE Transactions on Geoscience and Remote Sensing **40** (2002), no. 6, 1430–1435.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data and the EM algorithm*, Journal of the Royal Statistical Society (1977), no. 39, 1–38.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd edition, Wiley, New York, 2001.

[8] M. A. T. Figueiredo and A. K. Jain, *Unsupervised learning of finite mixture models*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002), no. 3, 381–396.

[9] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd edition, Academic Press, 1990.

[10] V. Krylov, G. Moser, S. Serpico, and J. Zerubia, *Modeling the statistics of high resolution SAR images*, Research Report 6722, INRIA, November 2008.

[11] E. E. Kuruoglu and J. Zerubia, *Modelling SAR images with a generalization of the Rayleigh distribution*, IEEE Transactions on Image Processing **13** (2004), no. 4, 527–533.

[12] H.-C. Li, W. Hong, and Y.-R. Wu, *Generalized gamma distribution with MoLC estimation for statistical modeling of SAR images*, Proceedings of APSAR 2007. 1st Asian and Pacific Conference, 2007, pp. 525–528.

[13] P. Masson and W. Pieczynski, *SEM algorithm and unsupervised statistical segmentation of satellite images*, IEEE Transactions on Geoscience and Remote Sensing **31** (1993), no. 3, 618–633.

[14] U. Maulik and S. Bandyopadhyay, *Performance evaluation of some clustering algorithms and validity indices*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002), no. 12, 1650–1654.

[15] G. Moser, S. Serpico, and J. Zerubia, *Dictionary-based stochastic expectation–maximization for SAR amplitude probability density function estimation*, IEEE Transactions on Geoscience and Remote Sensing **44** (2006), no. 1, 188–199.

[16] G. Moser, J. Zerubia, and S. B. Serpico, *SAR amplitude probability density function estimation based on a generalized gaussian model*, IEEE Transactions on Image Processing **15** (2006), no. 6, 1429–1442.

[17] J.-M. Nicolas, *Introduction aux statistiques de deuxième espéce: application aux lois d'images RSO (introduction to second kind statistics: applications to SAR images laws)*, Research Report (in French) 2002D001, ENST, Paris, February 2002.

[18] ———, *Introduction aux statistiques de deuxième espéce: applications des logs-moments et des logs-cumulants a l'analyse des lois d'images radar*, Traitement du Signal (in French) **19** (2002).

[19] J.-M. Nicolas and F. Tupin, *Gamma mixture modeled with "second kind statistics": application to SAR image processing*, Proceedings of the IGARSS Conference, Toronto (Canada), 2002.

[20] C. Oliver and S. Quegan, *Understanding Synthetic Aperture Radar images*, Artech House, Norwood, 1998.

[21] A. Papoulis, *Probability, random variables, and stochastic processes*, 3rd edition, McGraw-Hill International Editions, 1991.

[22] M. Petrou, F. Giorgini, and P. Smits, *Modelling the histograms of various classes in SAR images*, Pattern Recognition Letters **23** (2002), 1103–1107.

[23] R. A. Redner and H. F. Walker, *Mixture densities, maximum likelihood, and the EM algorithm*, SIAM Review **26** (1984), no. 2, 195–239.

[24] ———, *Unsupervised learning of finite mixture models*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002), no. 3, 381–396.

[25] J.A. Richards and X. Jia, *Remote sensing digital image analysis*, Springer-Verlag, Berlin, 1999.

[26] I. Sneddon, *The use of integral transforms*, McGraw-Hill, New York, 1972.

[27] C. Tison, J.-M. Nicolas, and F. Tupin, *Accuracy of Fisher distributions and log-moment estimation to describe histograms of high-resolution SAR images over urban areas*, Proceedings of the IGARSS Conference, July 21-25, Toulouse (France), 2003.

[28] C. Tison, J.-M. Nicolas, F. Tupin, and H. Maitre, *A new statistical model for Markovian classification of urban areas in high-resolution SAR images*, IEEE Transactions on Geoscience and Remote Sensing **42** (2004), no. 10, 2046–2057.

[29] H. L. Van Trees, *Detection, estimation and modulation theory*, vol. 1, John Wiley & Sons., New York, 1968.