Title: Rayleigh-Rice Mixture Parameter Estimation via EM Algorithm for Change Detection in Multispectral Images

This paper appears in: IEEE Transactions on Image Processing

Date of Publication: 28 August 2015

Author(s): M. Zanetti, F. Bovolo, L. Bruzzone

Volume: 24, Issue: 12

Page(s): 5004 – 5016

DOI: 10.1109/TIP.2015.2474710

Rayleigh-Rice mixture parameter estimation via EM algorithm for change detection in multispectral images

Massimo Zanetti, Francesca Bovolo, *Senior Member, IEEE,* and Lorenzo Bruzzone, *Fellow, IEEE*

Abstract

The problem of estimating the parameters of a Rayleigh-Rice mixture density is often encountered in image analysis (e.g., remote sensing and medical image processing). In this paper we address this general problem in the framework of change detection (CD) in multitemporal and multispectral images. One widely used approach to change detection in multispectral images is based on Change Vector Analysis (CVA). Here, the distribution of the magnitude of the difference image can be theoretically modeled by a Rayleigh-Rice mixture density. However, given the complexity of this model, in applications a Gaussian-mixture approximation is often considered, which may affect the change detection results. In this paper we present a novel technique for parameter estimation of the Rayleigh-Rice density that is based on a specific definition of the Expectation-Maximization (EM) algorithm. The proposed technique, which is characterized by good theoretical properties, iteratively updates the parameters and does not depend on specific optimization routines. Several numerical experiments on synthetic data demonstrate the effectiveness of the method which is general and can be applied to any image processing problem involving the Rayleigh-Rice mixture density. In the change detection context, the Rayleigh-Rice model (which is theoretically derived) outperforms other empirical models. Experiments on real multitemporal and multispectral remote sensing images confirm the validity of the model by returning significantly higher change detection accuracies than those obtained by using state-of-the-art approaches.

Index Terms

Parameter estimation, EM algorithm, change detection, Rayleigh distribution, Rician distribution, change vector analysis, multispectral images, remote sensing.

I. Introduction

CHANGE Detection (CD) techniques are very important in many image processing application domains [1]. For example, in remote sensing, thanks to the increasing number of sensors providing multitemporal images, a huge amount of data suitable for CD is nowadays at disposal. This gives us the unique opportunity of monitoring the Earth surface by mapping changes occurred on the ground at very high temporal and spatial resolutions.

The approaches proposed in the literature to address CD can be divided into two main categories: supervised and unsupervised methods. Supervised methods [2] are based on classification procedures for which the availability of multitemporal reference data is necessary for the training phase of the classifiers. Since obtaining reference information is always a costly task, the use of supervised methods is rarely a practical solution. Unsupervised methods [3]–[9] which do not require any training set, are often preferred. The framework of unsupervised change detection encompasses a variety of automatic techniques for the detection of different kinds of change. According to the final application goals there are methods for the detection of multiple change information and methods for binary change detection. In multiple change detection the analysis requires the discrimination among all the possible classes of change. To this aim several methods have been proposed in the literature. Among them we recall Change Vector Analysis (CVA) [6], Compressed CVA (C^2VA) [7], Multivariate Alterate Detection (MAD) [8], Temporal Principal Component Analysis (T-PCA) [9]. In binary change detection the only interesting information is whether a pixel represents a change or not. In this context the semantic meaning of the change is ignored, and the labeling is carried out only for the two classes of unchanged and changed pixels (i.e., presence/absence of changes). Most of the

M. Zanetti and L. Bruzzone are with the Department of Information and Communication Technology, University of Trento, Trento I-38123, Italy. e-mail: {massimo.zanetti,lorenzo.bruzzone}@unitn.it.

F. Bovolo is with the Center for Information Technologies, Fondazione Bruno Kessler (FBK), Trento I-38123, Italy. e-mail: bovolo@fbk.eu.

techniques for unsupervised binary change detection are based on automatic statistical modeling and thresholding [3] and data clustering [4]. There are also data fusion based approaches [5].

In this work, the attention is focused on the CVA technique (and its derivations) [10] and the use of automatic statistical modeling and thresholding within the CVA framework. CVA demonstrated to be a valuable and flexible tool for the detection of changes in several contexts (e.g., Remote Sensing [11], Medical Diagnosis and Treatment [12]). This technique is based on the representation in polar coordinates of the difference image (which is obtained by subtracting two images representing the same scene at different times). In the polar feature space, pixels having high magnitude values are likely to be changed and their separation into different kinds of change can be performed by means of their direction values [6], [13]. As mentioned, the magnitude of the multispectral difference image carries information about presence/absence of changes. Thus, the magnitude variable can be employed to separate changed from unchanged samples, eliminate the latter and perform further analysis only on the former ones [13]. Usually, the information in the magnitude variable is extracted by means of thresholding procedures [14], [15]. In [3], [16], the statistical distribution of the magnitude as a mixture model representing the classes of unchanged and changed pixels is approximated by a Gaussian mixture, then decision is made using a Bayesian rule. However, recent studies [6] showed that the precise model of this distribution can be theoretically derived, thus opening the way to a theoretical well-founded method instead of an empirical one. The model relies mainly on two hypotheses: (1) natural classes are Gaussian distributed within each band of the multispectral images (a reasonable assumption for images obtained from passive sensors) and, (2) pixels are spatially independent (this assumption is usually done for remote sensing images at medium resolution). Under these hypotheses, the magnitude of the difference image can be theoretically described by a Rayleigh-Rice mixture density. For this reason we address the problem of defining an EM-type algorithm to the estimation of shape and mixture parameters of the Rayleigh-Rice density in order to accurately solve the binary CD problem on multispectral multitemporal images.

Problems involving the Rician distribution often arise in engineering applications, in particular in Remote Sensing [6] and Magnetic Resonance Imaging (MRI) [16]–[20]. Since when the Rician distribution was introduced for the modeling of the magnitude of Gaussian densities [21], many efforts have been made for developing algorithms for the estimation of its parameters. Important results have been achieved by using the method of moments (MOM) [19], [22], and Maximum-Likelihood (ML) approaches [22], [23]. Unfortunately, the MOM is shown to be inefficient at low signal-to-noise ratios (SNRs) [18], [20], while ML equations do not have in general a unique solution [24]. Because of the latter property, the solution of the ML problem becomes an optimization problem. Some papers propose adaptive techniques for selecting the initial starting values [18], [20], while in other cases slightly different Bayesian estimators are proposed in order to stabilize the problem [25].

In spite of these results, the problem of parameter estimation in the case of mixture densities including the Rician distribution is still poorly investigated. When the non-centrality of the signal is high, the Rician density is often approximated with a Gaussian one [16]. Other papers directly address the approximation of the parameters of a mixture model involving one (optional) Rayleigh and J Rician distributions all of them having a common scale parameter σ. In [26], the authors proposed a fitting procedure followed by an approximated Expectation-Maximization (EM) estimation of σ. In [27] an additional parameter is included in the model as missing information, leading to a substantial simplification in the maximization step. In practice, in [26] and [27] the parameter σ describes a common characteristic of the noise, which is supposed to appear with different non-centralities and same scale parameter. Therefore the estimation of σ from J Rician components makes the model robust. Other methods are based on local noise estimation [28] and wavelet-based noise estimation [29]. In many application problems, forcing the components of the Rayleigh-Rice mixture to have the same scale parameters is a strong assumption. In particular, in CD on remote sensing real images such similarity is rarely observed, and neither it has a theoretical justification. Therefore, an empirical use of the algorithms in [26], [27] to fit the distribution of the magnitude of the difference image, and therefore to solve the binary CD problem, is expected to present limitations.

In this paper we develop an EM-type novel method for the estimation of all shape and mixture parameters of a Rayleigh-Rice mixture density. Asymptotic properties of the considered statistical model enable us to define an iterative method based on subsequent updates of the parameters. By providing a variational interpretation of the EM algorithm as a problem related to a fixed point equation we both: (1) derive explicit formulas for implementing the updates, and (2) establish a lower bound on the speed of convergence of the iterations for reaching a maximum of the

expectation. The algorithm is robust and it can be initialized using standard techniques of preliminary thresholding. On the one hand, if compared to already existing methods, in our model all the statistical parameters of the mixture density are free (we let the two mixture components to have different scale parameters). As a result, a large variety of practical problems can be addressed. On the other hand, we emphasize that the algorithm does not need any optimization routine (differently from [26], [27]), thus many issues related to the choice of optimal maximization strategies and their impact on the solution are avoided.

The paper is structured as follows. Section II introduces the Bayesian framework for binary change detection based on the Rayleigh-Rice mixture density as a model describing the magnitude of the difference image. In Section III we first present an overview of the EM algorithm and then we provide explicit formulas for the approximation of mixture and shape parameters of the Rayleigh-Rice mixture density. Convergence analysis is also provided. The experimental results are given in Section IV where the proposed method is applied to synthetic and real change detection problems on multispectral images. Section V draws the conclusions of this paper. An Appendix including mathematical notions and technical results is also provided.

II. A PROBLEM OF BINARY CHANGE DETECTION

In this section, following [6], we recall how under reasonable assumptions the statistical distribution of the magnitude of a bi-temporal difference image acquired by passive sensors is described by a Rayleigh-Rice mixture density. The magnitude of the unchanged pixels follows a Rayleigh distribution, while the magnitude of the changed pixels follows a Rician distribution. Then, a framework for the classification of the pixels according to their class is given in terms of Bayesian decision theory.

A. The Rayleigh-Rice mixture model for the magnitude of the difference image

Let us consider two multispectral images X_1, X_2 acquired by passive remote sensing sensors at different times t_1, t_2, respectively, and representing the same geographical area. Let us assume that the two images are co-registered and radiometrically corrected, and that there has been (only) one relevant change in the scene between the two dates. Therefore, the pixels can be divided into two classes only: ω_n (unchanged pixels) and ω_c (changed pixels). The aim is to discriminate between changed and unchanged pixels in an unsupervised way. The detection of the changes occurred between t_1 and t_2 is based on the study of the so-called difference image

$$X_D := X_2 - X_1. \tag{1}$$

When images acquired by passive sensors are considered, the statistical distribution of natural classes within each spectral band can be reasonably modeled by Gaussian densities [30]. From now on, our analysis will be restricted to considering two bands among all the available ones. The following theoretical analysis holds for any multi-band image where classes can be modeled as Gaussian densities, both in the case where they are independent or jointly-distributed as Gaussian. In these cases we have that also the classes ω_n, ω_c in the difference image are Gaussian distributed, thus let us assume they are modeled by $\mathcal{N}(\mu_{b,n}, \sigma_{b,n})$ and $\mathcal{N}(\mu_{b,c}, \sigma_{b,c})$, for each band $b = 1, 2$.

In polar coordinates, the magnitude of pixels can be exploited for discriminating between unchanged and changed pixels, so we are interested in describing how this model can be represented when the difference image is transformed. Since we are considering two-band images, we have that every pixel $X_D(i, j)$ is a two-dimensional vector. Hence, it is uniquely determined by its magnitude $\rho(i, j)$ and direction $\theta(i, j)$ with respect to a fixed reference direction. Given the pixel in spatial position (i, j), if no change has occurred on the ground between the two dates t_1 and t_2, then the magnitude $\rho(i, j)$ is expected to be close to zero. Conversely, whenever a change has occurred, the magnitude is expected to be significantly different from zero.

We are now interested in modeling the theoretical distribution of the magnitude. To this aim, let us denote the random variable that describes this feature by ρ. Because of the above-mentioned assumption of normality of classes within bands, the distribution of ρ is theoretically given by the distribution of the magnitude of a two-dimensional point whose coordinates are Gaussian distributed random variables. The obtained model is quite complex, but with some additional reasonable assumption it can be greatly simplified [21]. A first crucial assumption is spatial independence of pixels. From the application viewpoint, this assumption is reasonable and widely supported in literature if optical images at medium spatial resolution (e.g., 30 mt) are considered [3], [31]. Then, since the two

images are co-registered and radiometrically corrected, we can assume that in those areas where pixels are not changed the distributions are not significantly different, therefore

$$\begin{aligned} \mu_{1,n} &= \mu_{2,n} = 0 \\ \sigma_{1,n} &= \sigma_{2,n} =: \sigma_n \end{aligned} \tag{2}$$

and we have that the magnitude of unchanged pixels is modeled by a Rayleigh distribution

$$\mathrm{p}\left(\rho|\omega_n\right) = \frac{\rho}{b_n^2}\exp\left(-\frac{\rho^2}{2b_n^2}\right) \qquad \rho \geq 0, \tag{3}$$

where the parameter $b_n = \sigma_n$. In the case of changed pixels, we still assume that the distributions have the same variance, but they can have different non-zero means. Hence,

$$\begin{aligned} \mu_{1,c} \neq \mu_{2,c} \qquad & \mu_{1,c}, \mu_{2,c} \neq 0 \\ \sigma_{1,c} = \sigma_{2,c} &=: \sigma_c. \end{aligned} \tag{4}$$

and the magnitude of changed pixels follows the more general Rician distribution, [21]

$$\mathrm{p}\left(\rho|\omega_c\right) = \frac{\rho}{\sigma_c^2}\exp\left(-\frac{\rho^2+\nu_c^2}{2\sigma_c^2}\right)\mathrm{I}_0\left(\frac{\rho\nu_c}{\sigma_c^2}\right) \qquad \rho \geq 0. \tag{5}$$

Here $\nu_c = \sqrt{\mu_{1,c}^2 + \mu_{2,c}^2}$ is the so called non-centrality parameter, and $I_0(.)$ is the 0-th order modified Bessel function of first kind [32]. In conclusion, the theoretical mixture density that models the distribution of the magnitude of pixels is given by

$$\mathrm{p}\left(\rho\right) = \mathrm{p}\left(\omega_n\right)\mathrm{p}\left(\rho|\omega_n\right) + \mathrm{p}\left(\omega_c\right)\mathrm{p}\left(\rho|\omega_c\right) \tag{6}$$

where $\mathrm{p}\left(\omega_h\right)$, $h = n, c$, are the prior probabilities of classes. According to the given assumptions, the magnitude image $|X_D| := \{\rho(i,j) : i,j\}$ can be considered as a set of i.i.d. samples drawn from the theoretical distribution (6). For images where the i.i.d. assumption is not reasonable, further modelization of the spatial-contextual dependence of pixels is required. For example this can be done using Markov Random Fields (MRFs) [3], [14].

B. A framework for automatic binary change detection

The theoretical formulation of the distribution of the pixel magnitude as a mixture model allows for a formal characterization of a threshold that separates pixels into two classes according to their magnitude, by means of the Bayesian decision theory. The two probability models involved in the mixture model, the Rayleigh and the Rician distributions, have got a non-empty intersection. Accordingly, it is not possible to exactly decide which class a pixel with a given magnitude belongs to. This means that, for every possible classification, there is always a classification error with probability e_n, e_c (as shown in Table I). It is well known that the overall error $e_n + e_c$ can be minimized by selecting the separating threshold as the solution $\rho = T$ (solving for ρ) of the equation

$$\frac{\mathrm{p}\left(\omega_c\right)}{\mathrm{p}\left(\omega_n\right)} = \frac{\mathrm{p}\left(\rho|\omega_n\right)}{\mathrm{p}\left(\rho|\omega_c\right)} \qquad \rho \geq 0 \tag{7}$$

that corresponds to the intersection of the two curves $\mathrm{p}\left(\omega_n\right)\mathrm{p}\left(\rho|\omega_n\right)$ and $\mathrm{p}\left(\omega_c\right)\mathrm{p}\left(\rho|\omega_c\right)$ which lies between the two modes (it is worth noting that this equation generally has more than one solution). This equation can be equivalently written as

$$\left(\frac{1}{2b_n^2} - \frac{1}{2\sigma_c^2}\right)\rho^2 + \log \mathrm{I}_0\left(\frac{\rho\nu_c}{\sigma_c^2}\right) = \frac{\nu_c}{2b_n^2} + \log\frac{\sigma_c^2\,\mathrm{p}\left(\omega_n\right)}{b_n^2\,\mathrm{p}\left(\omega_c\right)}. \tag{8}$$

The classification that corresponds to the minimum overall error is given by

$$\begin{aligned} W_n &= \{(i,j) : \rho(i,j) \leq T\} \\ W_c &= \{(i,j) : \rho(i,j) > T\}, \end{aligned} \tag{9}$$

where W_n, W_c are the sets of predicted unchanged/changed pixels, respectively.

In Section III we provide an EM-type algorithm for finding an accurate estimation of the parameters b_n, ν_c, σ_c and the prior probabilities $\mathrm{p}\left(\omega_n\right)$, $\mathrm{p}\left(\omega_c\right)$, in such a way that the change detection can be performed by thresholding the magnitude of the difference image at T, which is obtained by solving (8).

TABLE I
CLASSIFICATION ERROR IN A TWO-CLASS DECISION PROBLEM. ω_n UNCHANGED PIXELS, ω_c CHANGED PIXELS

		Predicted classes	
		ω_n	ω_c
Actual classes	ω_n	$-$	e_n
	ω_c	e_c	$-$

III. THE EM ALGORITHM FOR PARAMETER ESTIMATION OF THE RAYLEIGH-RICE MIXTURE DENSITY

Here we consider the problem of estimating the parameters describing the probability density function of a continuous non-negative random variable ρ. In the considered model, the population represented by ρ is a mixture of two components following a Rayleigh and a Rician distribution, respectively. Given a set of i.i.d. samples drawn from this distribution, the estimation problem is solved by following the principles of the EM algorithm (see [33], [34], for a comprehensive overview of the theory).
Firstly, an overview of the EM theory is given. Then, by exploiting the asymptotic properties of the considered model, an explicit iterative method for solving the estimation problem is presented. Being iterative and explicit, the proposed method does not require any optimization strategy. Detailed discussion on the convergence of the method is given.

A. EM approach to parameter estimation

Let us consider the family of density functions depending on the set of parameters $\Psi = (\alpha, \Theta)$, where $\Theta = (\theta_1, \theta_2)$ with $\theta_1 = b$ and $\theta_2 = (\nu, \sigma)$, given by

$$\mathrm{p}\left(\rho|\Psi\right) = \alpha_1 \, \mathrm{p}\left(\rho|\omega_1, \theta_1\right) + \alpha_2 \, \mathrm{p}\left(\rho|\omega_2, \theta_2\right), \tag{10}$$

where $\alpha_1 = \alpha$, $\alpha_2 = 1 - \alpha$ with $0 < \alpha < 1$, and

$$\mathrm{p}\left(\rho|\omega_1, \theta_1\right) = \frac{\rho}{b^2} \exp\left(-\frac{\rho^2}{2b^2}\right) \tag{11}$$

$$\mathrm{p}\left(\rho|\omega_2, \theta_2\right) = \frac{\rho}{\sigma^2} \exp\left(-\frac{\rho^2 + \nu^2}{2\sigma^2}\right) \mathrm{I}_0\left(\frac{\rho\nu}{\sigma^2}\right). \tag{12}$$

Let $\mathbf{x}$ be a set of N i.i.d. samples drawn from the distribution (10) determined by the set of parameters $\bar{\Psi}$, i.e., $\mathrm{p}\left(\rho|\bar{\Psi}\right)$. The aim of the EM algorithm is to estimate the real values $\bar{\Psi}$ using the samples $\mathbf{x}$. In particular, here we consider the case where each sample $x \in \mathbf{x}$ is unlabeled, thus the sampling is incomplete. We recall that a complete sample would be of the form $y = (x, \omega)$ where $\omega \in \{\omega_1, \omega_2\}$ is the *label* of the sample x (i.e., the index representing the population from which the observed value x comes from).
The EM algorithm gives, under certain hypotheses, an estimation of $\bar{\Psi}$ as a local maximizer $\hat{\Psi}$ of the so-called log-likelihood of the samples $\mathbf{x}$

$$L(\Psi) = \sum_{x \in \mathbf{x}} \log \, \mathrm{p}\left(x|\Psi\right). \tag{13}$$

The key point of the EM algorithm is [33]: given Ψ', then $L(\Psi) \geq L(\Psi')$ if Ψ maximizes the *conditional log-expectation*

$$Q(\Psi|\Psi') = \sum_{x \in \mathbf{x}} \sum_{h=1,2} \mathrm{p}\left(\omega_h|x, \Psi'\right) \log\left(\alpha_h \, \mathrm{p}\left(x|\omega_h, \theta_h\right)\right), \tag{14}$$

where for $h = 1, 2$ and for each $x \in \mathbf{x}$, the weight

$$\mathrm{p}\left(\omega_h|x, \Psi'\right) := \frac{\alpha'_h \, \mathrm{p}\left(x|\omega_h, \theta'_h\right)}{\mathrm{p}\left(x|\Psi'\right)} \tag{15}$$

is the posterior probability that x originated in the h-th component of the population, given Ψ' (see Figure 1). In its more general fashion, the EM algorithm is implemented as follows: an approximation Ψ^0 is firstly chosen, then for

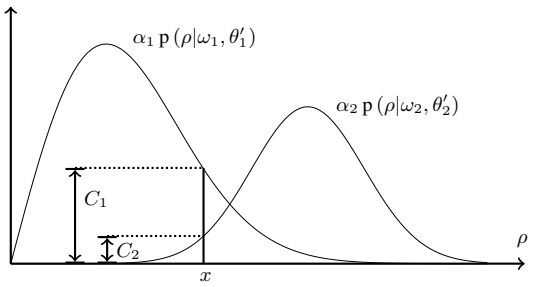


Fig. 1. Geometrical interpretation of the posterior probability that x originated in the h-th component given Ψ', as $\mathrm{p}\,(\omega_h|x,\Psi') = C_h/(C_1+C_2)$ for $h=1,2$.

$k=0,1,\ldots$ the conditional log-expectation $Q(\Psi|\Psi^k)$ is evaluated (E-step) and the next iterate is found (M-step) as $\Psi^{k+1} := \arg\max_\Psi Q(\Psi|\Psi^k)$. A recursive formula for updating the mixing proportions can be formally derived and the M-step can be also formulated as[1]

$$\begin{aligned} \alpha^{k+1} &:= N^{-1}\sum_{x\in\mathbf{x}} \mathrm{p}\,(\omega_1|x,\Psi^k) \\ \Theta^{k+1} &:= \arg\max_{\Theta} Q((\alpha^k,\Theta)|\Psi^k). \end{aligned} \tag{16}$$

Exploiting information theory, the EM algorithm to the parameter estimation of (10) can be further simplified. In the considered statistical model all parameters b,ν,σ are mutually independent (see Appendix A-C). Thus, following [34] the EM algorithm can be split into separated maximization steps

$$\begin{aligned} \alpha^{k+1} &:= N^{-1}\sum_{x\in\mathbf{x}} \mathrm{p}\,(\omega_1|x,\Psi^k) \\ b^{k+1} &:= \arg\max_{b} Q((\alpha^k,b,\nu^k,\sigma^k)|\Psi^k) \\ \nu^{k+1} &:= \arg\max_{\nu} Q((\alpha^k,b^k,\nu,\sigma^k)|\Psi^k) \\ \sigma^{k+1} &:= \arg\max_{\sigma} Q((\alpha^k,b^k,\nu^k,\sigma)|\Psi^k). \end{aligned} \tag{17}$$

Note that the difference between (16) and (17) is substantial. Without the parameter independence assumption, the iterative search of maximizers cannot be separated and the implementation must rely on ad-hoc optimization techniques. For example this is the case of [26] with $J=2$, where $\theta_1=(\nu_1,\sigma)$ and $\theta_2=(\nu_2,\sigma)$, the parameters of two Rician distributions, are obviously not independent. In this paper, we take advantage of (17) and we show that each partial maximization step can be performed by updating the corresponding variable according to an iterative rule.

B. Iterative equations for the EM algorithm

In this section the problem of numerically implementing (17) is addressed. By defining $\boldsymbol{\ell}(\Psi) := Q(\Psi|\Psi)$, the iterative procedure (17) can be instantiated by an iterative method attempting to solve

$$\nabla\boldsymbol{\ell}(\Psi) = 0. \tag{18}$$

[1]We present also this formulation of the EM algorithm since it helps us in making a clear distinction between our approach and the one in [26].

By writing gradient equations and performing math (see Appendix A-B for details) we get

$$\begin{aligned}
\frac{\partial \boldsymbol{\ell}}{\partial \alpha} &= \sum_{x \in \mathbf{x}} \frac{1}{\alpha} \mathrm{p}\left(\omega_1 | x, \Psi\right) - \frac{1}{1-\alpha} \mathrm{p}\left(\omega_2 | x, \Psi\right) \\
\frac{\partial \boldsymbol{\ell}}{\partial b} &= \sum_{x \in \mathbf{x}} \mathrm{p}\left(\omega_1 | x, \Psi\right) \left[\frac{x^2}{b^3} - \frac{2}{b}\right] \\
\frac{\partial \boldsymbol{\ell}}{\partial \nu} &= \sum_{x \in \mathbf{x}} \mathrm{p}\left(\omega_2 | x, \Psi\right) \left[\frac{x}{\sigma^2} \frac{\mathrm{I}_1\left(\frac{x\nu}{\sigma^2}\right)}{\mathrm{I}_0\left(\frac{x\nu}{\sigma^2}\right)} - \frac{\nu}{\sigma^2}\right] \\
\frac{\partial \boldsymbol{\ell}}{\partial \sigma} &= \sum_{x \in \mathbf{x}} \mathrm{p}\left(\omega_2 | x, \Psi\right) \left[\frac{x^2 + \nu^2}{\sigma^3} - \frac{2}{\sigma} - \frac{2x\nu}{\sigma^3} \frac{\mathrm{I}_1\left(\frac{x\nu}{\sigma^2}\right)}{\mathrm{I}_0\left(\frac{x\nu}{\sigma^2}\right)}\right]
\end{aligned} \tag{19}$$

where $\mathrm{I}_1(.)$ is the 1-st order modified Bessel function of first kind [32]. As we can see, gradient equations are highly non-linear. Formally, we derive a set of iterative equations for approximating the solution of (18) according to the method of subsequent approximations (see Appendix A-D for details). After some analytical manipulations we get the following iterative rules

$$\begin{aligned}
\alpha^{k+1} &= N^{-1} \sum_{x \in \mathbf{x}} \mathrm{p}\left(\omega_1 | x, \Psi^k\right) \\
(b^2)^{k+1} &= \frac{\sum\limits_{x \in \mathbf{x}} \mathrm{p}\left(\omega_1 | x, \Psi^k\right) x^2}{2 \sum\limits_{x \in \mathbf{x}} \mathrm{p}\left(\omega_1 | x, \Psi^k\right)} \\
\nu^{k+1} &= \frac{\sum\limits_{x \in \mathbf{x}} \mathrm{p}\left(\omega_2 | x, \Psi^k\right) \frac{\mathrm{I}_1\left(\frac{x\nu^k}{(\sigma^k)^2}\right)}{\mathrm{I}_0\left(\frac{x\nu^k}{(\sigma^k)^2}\right)} x}{\sum\limits_{x \in \mathbf{x}} \mathrm{p}\left(\omega_2 | x, \Psi^k\right)} \\
(\sigma^2)^{k+1} &= \frac{\sum\limits_{x \in \mathbf{x}} \mathrm{p}\left(\omega_2 | x, \Psi^k\right) \left[x^2 + (\nu^k)^2 - 2x\nu^k \frac{\mathrm{I}_1\left(\frac{x\nu^k}{(\sigma^k)^2}\right)}{\mathrm{I}_0\left(\frac{x\nu^k}{(\sigma^k)^2}\right)}\right]}{2 \sum\limits_{x \in \mathbf{x}} \mathrm{p}\left(\omega_2 | x, \Psi^k\right)}.
\end{aligned} \tag{20}$$

The above formulas fully determine our algorithm. In general, the method of subsequent approximations converges at least with linear speed provided the spectral radius of the Jacobian matrix of the iterative function, computed at the exact solution, is strictly less than one. Of course, convergence properties of the algorithm strongly depend on the choice of the first iterate.

C. Initialization of the algorithm

In order to increase the probability to converge to an optimal stationary point of the objective energy $\boldsymbol{\ell}(\Psi)$, an adequate initial approximation Ψ^0 must be found. A standard approach to this aim is based on a first raw classification of the data followed by maximum likelihood (ML) estimates of the parameters. Let us assume that the samples drawn from the mixture are divided into the two approximate classes W_1 and W_2, in such a way that $\mathbf{x} = W_1 \cup W_2$. ML estimates of shape parameters (in the following denoted by the ML superscript) are derived as the solutions of the so-called log-likelihood equations. Samples in W_1 are used to approximate b, whereas samples in W_2 are used to approximate ν, σ. For an explicit computation of their values we refer the reader to the existing literature [18], [20], [22], [25]. Once the ML estimates $b^{ML}, \nu^{ML}, \sigma^{ML}$ are computed, we use their values as the initial set of parameters Θ^0 for triggering the EM algorithm:

$$\Theta^0 = (b^{ML}, (\nu^{ML}, \sigma^{ML})). \tag{21}$$

As initial value for the mixing proportion we use the ratio (# denotes the number of samples)

$$\alpha^0 = \frac{\#W_1}{\#X}. \tag{22}$$

According to the method that is used to populate the approximate classes W_1, W_2, we have different initializations of the EM algorithm. The more these classes are good representatives of the true classes ω_1, ω_2, the more the ML parameter estimates of the two distributions $\mathrm{p}(\rho|\omega_h, \theta_h)$, $h = 1, 2$ will be close to the real values. A simple yet effective way to populate such approximate classes is given by thresholding the values of the samples $\mathbf{x}$ (in [3] a similar approach is used in the case of a mixture of Gaussian densities). Let us define, for any fixed value $T \geq 0$, the approximate classes W_1, W_2 as follows

$$\begin{aligned} W_1 &:= \{x \in X : x \leq T\}, \\ W_2 &:= \{x \in X : x > T\}. \end{aligned} \tag{23}$$

In terms of Bayes decision theory, the choice of the threshold T can be interpreted as an attempt of approximating a solution of (7). In this context, the problem of defining the approximate classes W_1, W_2 is turned into the problem of choosing the separating threshold T. At this stage, proper knowledge on the specific dataset should be used in order to simplify the task of computing T. On the one hand, such T should be able to give at least a coarse discrimination of the data, e.g., by properly exploiting the bi-modal behavior of the histogram. On the other hand, the computational complexity of this step, being itself a preliminary step, should be kept low. A choice often encountered in applications that meets the two above mentioned important requirements [3], is that of using

$$T = T_{mid} := \frac{\max X - \min X}{2}. \tag{24}$$

Whenever the two modes of the mixture described by (10) are well separated by T_{mid}, we expect to have sufficiently accurate preliminary ML estimates of the mixture parameters for triggering the EM algorithm. Of course, other strategies can be used.

D. Convergence analysis

In this section we analyze convergence properties of the parameter estimation algorithm defined by (20). For a better understanding of the quantitative analysis of the results, we specify that computations are performed using MATLAB® on a standard workstation. Hardware is Intel(R) Core(TM) i5-4750 CPU @3.20 GHz, 8.00 GB Ram.

According to [34], we aim at showing that the performance of the proposed EM-type iterative algorithm strongly depends on the *separability* of the two mixture components. The proposed iterative algorithm is run several times on synthetic samples $\mathbf{x}$ generated by the inverse transform sampling method applied to mixture densities of the type $\mathrm{p}(\rho|\Psi)$, where $\Psi = (\alpha, b, \nu, \sigma)$ are fixed sets of parameters. The size of each sample is $N = 10^4$. In order to parameterize the separability of the mixture components, in the tests all parameters are fixed except ν. Significance of the resulting estimates is ensured by stopping the algorithm at the iteration k such that the maximum relative error in approximating all the parameters is

$$\max_{i=1,\ldots,4} \left| \frac{\Psi_i - \Psi_i^k}{\Psi_i} \right| < 0.05.$$

For each test, we recorded the number of iterations k, the relative variation of the objective energy $|\boldsymbol{\ell}(\Psi^k) - \boldsymbol{\ell}(\Psi^{k-1})|/|\boldsymbol{\ell}(\Psi^{k-1})|$, the spectral radius of the Jacobian of the iterative function calculated at Ψ and the time of computation. The components of the Jacobian matrix are calculated by implementing (34).

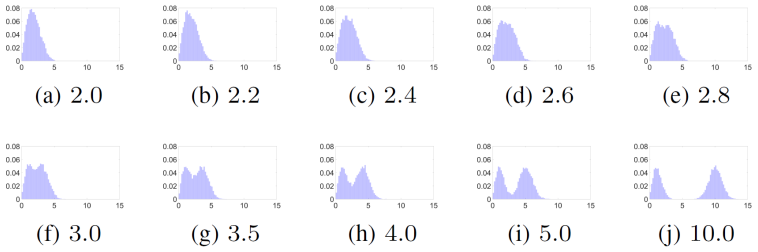
The results relate to two sets of parameters Ψ. In the first case the fixed values are $\alpha = 0.4$, $b = 1$, $\sigma = 1$, in the second case the Rician scale parameter is increased to $\sigma = 2$. In both cases the Rician non-centrality parameter ranges from $\nu = 2.0$ to $\nu = 10.0$. Results of computations (the mean of ten runs) are shown in Tables II and III. Histograms of the corresponding samples are given in Figures 2 and 3. Results are in agreement with the expected performance of the EM algorithm. Let us discuss more in detail the outcome of the experiments.

- An analysis of Figures 2 and 3 enables us to discriminate the separability of the corresponding mixtures. In the case of $\sigma = 1$, being also $b = 1$, the separation becomes more evident only for $\nu \geq 2.8$. In the case of $\sigma = 2$, the difference in the scale parameters allows for a better discrimination for any $\nu \geq 2.0$.
- As expected, the number of iterations for reaching sufficient approximations of the target parameters sensibly decreases as ν increases. In particular, in Table II the order of k moves from 10^3 ($\nu = 2.0$) to 10^0 ($\nu = 5.0, 10.0$), whereas in Table III this number is one order of magnitude less. This substantial difference is due

TABLE II

ITERATION DETAILS OF THE PROPOSED EM ALGORITHM ON RAYLEIGH-RICE MIXTURES WITH PARAMETERS $\alpha = 0.4$, $b = 1$, $\sigma = 1$ AND FOR DIFFERENT VALUES OF ν.

ν	k	$\frac{\lvert\boldsymbol{\ell}(\Psi^k)-\boldsymbol{\ell}(\Psi^{k-1})\rvert}{\lvert\boldsymbol{\ell}(\Psi^{k-1})\rvert}$	$rad(J\varphi(\Psi))$	$time(secs)$
2.0	3694	$1.07 \cdot 10^{-5}$	0.7728	35.93
2.2	2969	$2.26 \cdot 10^{-6}$	0.7419	30.13
2.4	1217	$1.14 \cdot 10^{-5}$	0.7034	13.65
2.6	800	$1.33 \cdot 10^{-5}$	0.6723	8.44
2.8	427	$9.14 \cdot 10^{-6}$	0.6521	4.84
3.0	266	$1.10 \cdot 10^{-5}$	0.6282	3.12
3.5	75	$2.03 \cdot 10^{-5}$	0.5963	1.07
4.0	22	$4.90 \cdot 10^{-4}$	0.5701	0.69
5.0	8	$1.29 \cdot 10^{-3}$	0.5426	0.55
10.0	1	$7.36 \cdot 10^{-5}$	0.5133	0.47



(a) 2.0 (b) 2.2 (c) 2.4 (d) 2.6 (e) 2.8

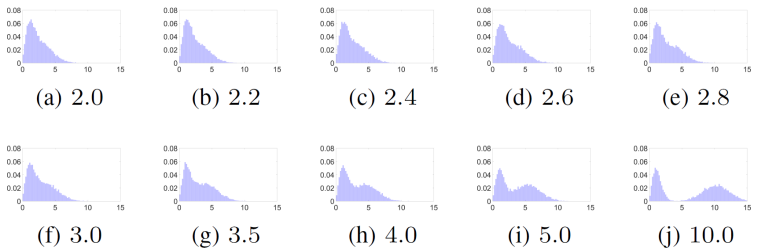
(f) 3.0 (g) 3.5 (h) 4.0 (i) 5.0 (j) 10.0

Fig. 2. Histograms of samples $\mathbf{x}$ generated from $\mathrm{p}\,(\rho|\Psi)$ with $\alpha = 0.4$, $b = 1$, $\sigma = 1$ and for different values of ν.

TABLE III

ITERATIONS DETAILS OF THE PROPOSED EM ALGORITHM ON RAYLEIGH-RICE MIXTURES WITH PARAMETERS $\alpha = 0.4$, $b = 1$, $\sigma = 2$ AND FOR DIFFERENT VALUES OF ν.

ν	k	$\frac{\lvert\boldsymbol{\ell}(\Psi^k)-\boldsymbol{\ell}(\Psi^{k-1})\rvert}{\lvert\boldsymbol{\ell}(\Psi^{k-1})\rvert}$	$rad(J\varphi(\Psi))$	$time(secs)$
2.0	352	$4.37 \cdot 10^{-6}$	0.8250	3.28
2.2	408	$5.90 \cdot 10^{-6}$	0.8424	2.99
2.4	304	$1.03 \cdot 10^{-5}$	0.8766	2.86
2.6	229	$2.34 \cdot 10^{-5}$	0.8827	2.21
2.8	186	$4.23 \cdot 10^{-5}$	0.8796	2.29
3.0	130	$9.09 \cdot 10^{-5}$	0.8625	1.72
3.5	91	$1.57 \cdot 10^{-4}$	0.8232	1.20
4.0	74	$1.06 \cdot 10^{-4}$	0.7781	1.13
5.0	28	$5.49 \cdot 10^{-4}$	0.6874	0.74
10.0	4	$5.41 \cdot 10^{-4}$	0.5447	0.54



(a) 2.0 (b) 2.2 (c) 2.4 (d) 2.6 (e) 2.8

(f) 3.0 (g) 3.5 (h) 4.0 (i) 5.0 (j) 10.0

Fig. 3. Histograms of samples $\mathbf{x}$ generated from $\mathrm{p}\,(\rho|\Psi)$ with $\alpha = 0.4$, $b = 1$, $\sigma = 2$ and for different values of ν.

to the fact that, in the second case, the separation between the two mixture components is possible because of the difference in the scale parameters.
- In all tests the spectral radius of the Jacobian of the iterative function is less than one, thus the algorithm enjoys at least linear convergence. Notice that for increasing ν, the spectral radius (which provides a quantitative estimation of the factor by which errors are reduced from one iteration to the next) decreases.
- Results from the tables suggest that a reasonable threshold value on the relative variation of the objective energy $\boldsymbol{\ell}(\Psi)$ for stopping the algorithm is $tol = 10^{-6}$.

IV. Experimental results on multispectral images

In this section, after presenting the datasets and the details of the experimental setup, we analyze the performance of the proposed method. Firstly, we give a quantitative measurement of data fitting by means of two divergence measures between the data and different statistical models estimated from the data. Then, the change detection performance is analyzed. A comparison in terms of change detection errors between the proposed method and the state-of-the-art one based on Gaussian mixtures is given.

A. Datasets description

The datasets considered in the experiments are both synthetic and real multispectral images. Real datasets (consisting of couples of multitemporal multispectral images) are accompanied by reference maps[2] of the changes, i.e., binary maps representing the classes ω_n, ω_c. Notice that minor changes (i.e., mapped changes that do not belong to the main change class of interest) are depicted in the reference maps in red (Figures 5c and 6c).

1) Dataset A (synthetic): This first dataset is a synthetically generated two-band difference image with statistical properties as defined in Section II-A (see Figure 4). Classes ω_n, ω_c are Gaussian distributed within each band with parameters: $\mu_{1,n} = \mu_{2,n} = 0$, $\sigma_n = 2.5$, $\mu_{1,c} = -50.0$, $\mu_{2,c} = -20.0$, $\sigma_c = 25$. The size of the image is 700×600 pixels. The proportions between the number of pixels in simulated classes and the total number of pixels (class priors) are

$$\mathrm{p}(\omega_n) = \frac{336000}{420000} = 0.8, \qquad \mathrm{p}(\omega_c) = \frac{84000}{420000} = 0.2.$$

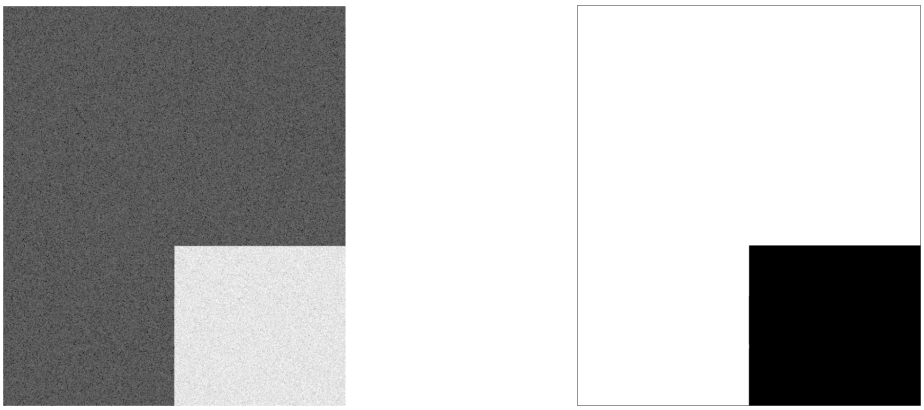
Simulated changed pixels are located in the bottom-right corner of the image.

2) Dataset B: The dataset is made up of two multispectral images acquired by the Thematic Mapper (TM) multispectral sensor of the Landsat 5 satellite (see Figure 5). The images are co-registered and radiometrically corrected. The scene represents an area including Lake Mulargia (Sardinia Island, Italy), at a resolution of 30 m. The image consists of 300×412 pixels (a total of 123600 pixels). The dates of acquisition are September 1995 (t_1) and July 1996 (t_2). Between the two dates one most relevant change, which is related to the extension of the lake surface, occurred in the study area. The scene presents 116120 unchanged and 7480 changed pixels. It follows that the prior probabilities of class are given by

$$\mathrm{p}(\omega_n) = \frac{116120}{123600} = 0.94, \qquad \mathrm{p}(\omega_c) = \frac{7480}{123600} = 0.06.$$

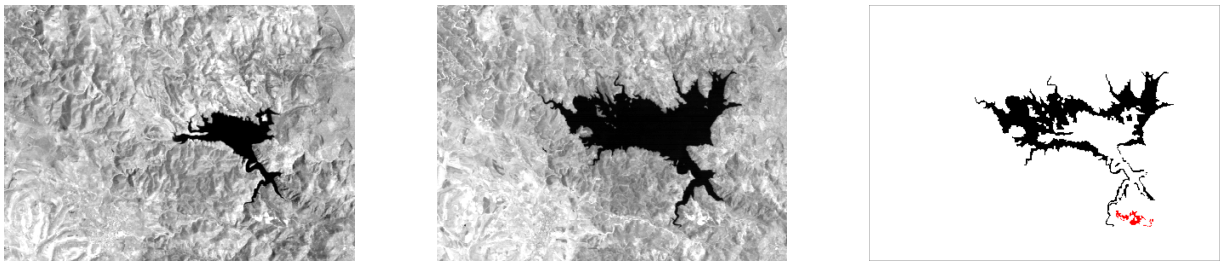
The two most representative bands of the changes are 4 and 7, the near infrared (NIR) and the middle infrared (MIR) (see [6] for details on the band selection). The ML parameter estimations of the normal distributions of classes ω_n, ω_c within each band of the difference image are $\mu_{1,n} = 3.57$, $\sigma_{1,n} = 10.26$, $\mu_{1,c} = -55.37$, $\sigma_{1,c} = 8.90$, and $\mu_{2,n} = 2.63$, $\sigma_{2,n} = 8.73$, $\mu_{2,c} = -40.84$, $\sigma_{2,c} = 10.67$. The numbers show that the initial assumptions (2) and (4) are approximately satisfied. The means $\mu_{1,n}$ and $\mu_{2,n}$ are both close to 0, $\sigma_{1,n}$ and $\sigma_{2,n}$ are very close each other and their mean can be used as an approximation of the variance of the unchanged pixels, i.e., $\sigma_n = 9.49$. A similar argument holds for the variance of the changed pixels, which can be approximated by $\sigma_c = 9.79$, the mean value of $\sigma_{1,c}$ and $\sigma_{2,c}$.

[2]Reference maps are obtained via photo-interpretation and they are used only for validation/comparison purposes.



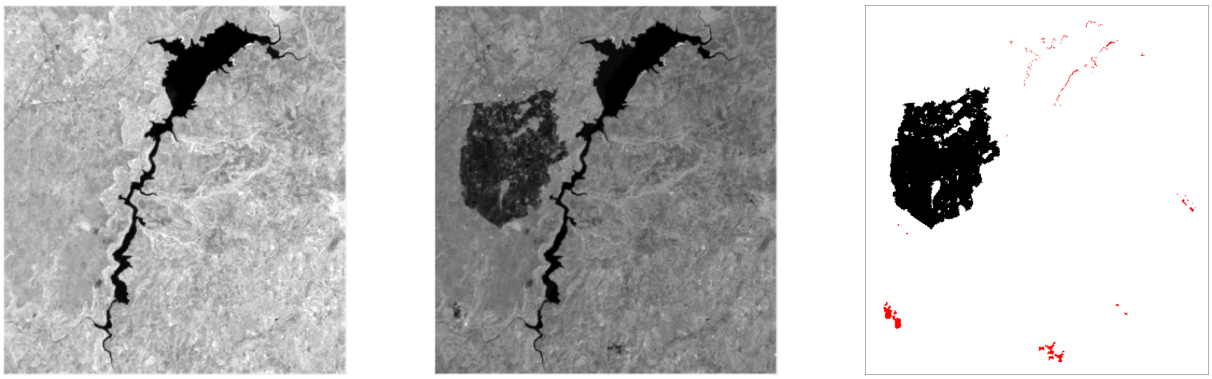
(a) (b)

Fig. 4. Dataset A. Synthetic two-band difference image. (a) Magnitude of the difference image, (b) map of simulated changed pixels (black).



(a) (b) (c)

Fig. 5. Dataset B: images of Lake Mulargia (Italy) acquired by the Thematic Mapper sensor of the Landsat 5 satellite: (a) channel 4 of the image acquired in September 1995; (b) channel 4 of the image acquired in July 1996; (c) change reference map indicating the enlargement of the lake (black) and an open quarry (red).



(a) (b) (c)

Fig. 6. Dataset C: images of Lake Omodeo and surrounding area (Italy) acquired by the Operational Land Imager sensor of the Landsat 8 satellite: (a) channel 5 of the image acquired in July 2013; (b) channel 5 of the image acquired in August 2013; (c) change reference map indicating the burned area extension (black) and other minor changes related to clouds and water (red).

3) Dataset C: The dataset consists of a couple of multispectral images acquired by the Operational Land Imager (OLI) multispectral sensor of the Landsat 8 satellite (see Figure 6). The investigated area includes Lake Omodeo and a portion of Tirso River (Sardinia Island, Italy). The image consists of 700×650 pixels (a total of 455000 pixels) at a resolution of 30 m. The dates of acquisition are 25th July 2013 (t_1) and 10th August 2013 (t_2). The change we are interested to estimate is a fire occurred between August 7th and 9th in the south of Ghilarza village. The post-event image is acquired just one day after the fire was extinguished. The area affected by the fire is mostly agricultural, with an extension of approximately 100 ha. The images are co-registered and radiometrically

corrected. According to the reference map, the scene presents 420227 unchanged and 34773 changed pixels, 1636 of them are related to small clouds and variations of the lake surface. The class prior probabilities are given by

$$\mathrm{p}\left(\omega_n\right) = \frac{420227}{455000} = 0.92, \qquad \mathrm{p}\left(\omega_c\right) = \frac{34773}{455000} = 0.08.$$

The two bands selected as most representative of the changes are bands 5 and 6, the near infrared (NIR) and the first short wavelength infrared (SWIR1). The choice is made by considering the band-pair that produced the smallest amount of overall errors in the change detection.

B. Experimental setup

Given two multispectral two-band images X_1, X_2, by following the framework given in Section II the magnitude of the difference image $|X_D|$ is modeled by the mixture density (6). The parameters of this density are estimated via EM algorithm. Then, a threshold for binary decision is calculated by following a Bayes rule and change detection is performed accordingly.

In the experiments presented in Section IV-C, the fitting of the proposed Rayleigh-Rice mixture model is tested against two empirical models for parameter estimation already present in literature: the first one is based on a Gaussian mixture, the second one is based on a mixture of Rician distributions with common scale parameter. It is worth noting that, the former represents the state-of-the-art in binary change detection based on the statistical modeling of the magnitude information, whereas the latter has been proposed by Maitra and Faden for the estimation of noise variance in MR images [26]. For clarity of notation, let us summarize the parameter notation used in the following:

Rayleigh-Rice mixture (RR):

$$\mathrm{p}\left(\rho|\omega_n\right) = \mathrm{Rayl}(b), \quad \mathrm{p}\left(\rho|\omega_c\right) = \mathrm{Rice}(\nu, \sigma).$$

Gaussian mixture (GG):

$$\mathrm{p}\left(\rho|\omega_n\right) = \mathcal{N}(\mu_1, \sigma_1), \quad \mathrm{p}\left(\rho|\omega_c\right) = \mathcal{N}(\mu_2, \sigma_2).$$

Rician mixture with common scale parameter (MF):

$$\mathrm{p}\left(\rho|\omega_n\right) = \mathrm{Rice}(\nu_1, \beta), \quad \mathrm{p}\left(\rho|\omega_c\right) = \mathrm{Rice}(\nu_2, \beta).$$

Iterative formulas for implementing parameter estimation in the GG model can be found in [7]. Implementation for parameter estimation in the MF case is presented in [26]. Initialization follows the principles outlined in Section III-C. Data fitting is measured in terms of the χ^2_P (χ^2–Pearson) divergence and the Kolmogorov-Smirnov (KS) distance between the data and the estimated densities.

In Section IV-D the binary change detection problem is solved using the framework described in Section II. The final parameter estimates of RR and GG are used to compute the magnitude thresholds T^{RR} and T^{GG} that correspond to the Bayes Decision Rule (BDR) of minimum overall error of classification, then change detection is performed by thresholding the difference image. By using a standard trial-and-error selection based on the reference map it has been possible to compute for all datasets the optimal threshold value T^{MOE} of minimum overall error (MOE). The results of change detection are compared to such optimal values. In order to make the results on the different datasets comparable, the detection errors are also given in terms of percentage according to $e_m := \frac{missed}{changed} \cdot 100\%$, $e_f := \frac{false}{unchanged} \cdot 100\%$ and $e_o := \frac{overall}{total} \cdot 100\%$. The steps of the whole procedure are summarized:

1) Populate the approximate classes W_1, W_2 and initialize the EM algorithm (in the MF case the algorithm is initialized as described in [26]).
2) Apply EM algorithm and assess fitting properties of the estimated mixture densities.
3) Calculate the threshold values according to both the BDR of minimum overall error and the optimal choice of MOE obtained via the reference map.
4) Populate the change detection classes W_n, W_c by thresholding the magnitude image.
5) Assess change detection performance in terms of false and missed alarms by comparing the estimated classes W_n, W_c with the true classes ω_n, ω_c.

C. Data fitting results

Let us present and discuss the results of EM parameter estimation and data fitting in the considered datasets. Numerical values of the estimated parameters and fitting performance in terms of statistical divergences χ^2_P and KS are showed in Table IV. For a qualitative understanding of the fitting of the estimated models, a plot of the histograms of the magnitude images with superimposed estimated densities is given in Figure 7.

In general, the real prior probability of the unchange class $\mathrm{p}(\omega_n)$ is always well approximated by α as no significant differences are observed among the trials. As expected, the fitting of RR is very precise in the case of Dataset A, as this dataset represents the ideal case of a difference image having the properties described in Section II. Nonetheless, the RR model results in the best fitting also in the real remote sensing datasets B and C, confirming that this model is much more suitable for representing the real distribution of the magnitude. The GG model is flexible enough to follow the bimodal behavior of the histogram, but it is never as precise as the RR. The limitations in using the MF model are evident. From the EM estimates of the RR model we can see that the scale parameters of the mixture components are significantly different in all datasets. This strongly affects the fitting of the MF model, which assumes the same scale parameter for both the mixture components. Data fitting measurements obtained with this model show that this strong assumption leads to very poor approximations of the real distributions (this can be seen very clearly from Figure 7). Thus, the model is inadequate to address the change detection problem. Notice also that the non-centrality parameter ν_1 converged approximately to 0 in all cases, confirming that the first component of the mixture (related to the distribution of unchanged pixels) is Rayleigh in real data.

Let us now check the consistency of the RR parameter estimates with the theoretical properties described in Section II-A that express the relationships between b_n, ν_c, σ_c (approximated by b, ν, σ) and the parameters of the normal distribution of classes within bands: σ_n, σ_c and $\mu_{b,c}$, for $b = 1, 2$. The values related to Dataset A match perfectly as the hypotheses are fulfilled by definition: we have $\sigma_n = b = 2.5$, $\sigma_c = 25$, $\sigma = 25.04$, and $\nu = 53.89$, $\sqrt{\mu^2_{1,c} + \mu^2_{2,c}} = 53.85$. In case of Dataset B, the Rayleigh parameter $b = 9.33$ matches with the variance $\sigma_n = 9.44$. The Rice parameters are slightly unmatched, compare $\sigma = 17.37$ with the variance $\sigma_c = 9.79$ and $\nu = 60.24$ with the non-centrality measurement $\sqrt{\mu^2_{1,c} + \mu^2_{2,c}} = 68.75$. A similar behavior is observed for Dataset C. The observed differences in real datasets are due to the Gaussian approximation of the distribution of classes within bands.

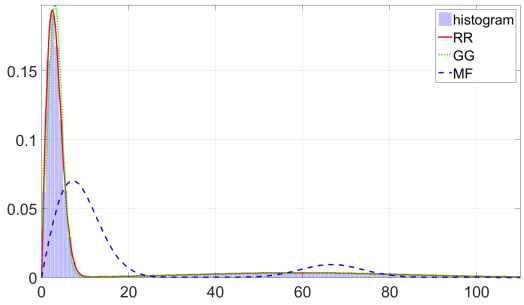
D. Change detection results

In this section the change detection results on the considered datasets are presented. The outcome of the experiments is detailed in Table V, where the CD performance is evaluated in terms of false and missed alarms and overall errors. It follows from the analysis in Section IV-C that the mixture component representing ω_n is better approximated by a Rayleigh density. The high non-symmetry of the Rayleigh component forces its Gaussian approximation to be slightly overestimated on the right side, thus, its right descending slope is steeper than the Rayleigh version. It follows that the decision threshold T^{GG} is smaller than T^{RR} in general. A first consequence of this is that the overall error of RR is always smaller than the error of GG. Moreover, the change detection by using the GG mixture presents less missed alarms, but many more false alarms with respect to the RR case. This is confirmed by looking at the results related to Dataset A (synthetic). It is not surprising that in this case the BDR threshold related to the RR model is almost identical to the optimal choice of minimum overall error: $T^{RR} = 10.12$ and $T^{MOE} = 10.40$. As expected, the threshold returned by the GG model, $T^{GG} = 8.82$, is smaller than T^{RR} and corresponds to an increasing overall error that moves from 0.19% (optimal case) to 0.29%. Notice that in this case the total number of wrongly detected pixels increases of approximately 1/3 (from 791 to 1211).

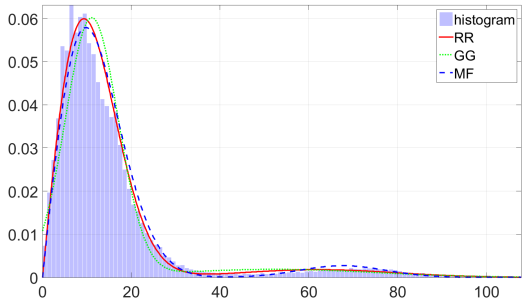
In the case of real datasets error percentages slightly increase, accounting for the approximation in assuming the Gaussian distribution of the classes within bands. Note that, though in some cases the BDR thresholds are not very close to the optimal values, the RR model always returned a better approximation of the optimal threshold with respect to the standard GG model. In Dataset B, the optimal threshold in terms of overall error corresponds to 770 overall errors, and the RR and the GG models returned 1820 and 3982 overall errors, respectively. Therefore, the overall error is more than halved. A similar result is obtained in the case of Dataset C, where again the RR model performed much better than GG. The number of overall errors is 4621 in the optimal case, whereas it is 8761 in case of RR and 13583 in the case of GG. Again we observe that the overall error is more than halved by using the proposed RR model. Figure 8 shows the change detection maps obtained by thresholding the magnitude of the difference image in the three datasets. Notice that in general the smaller number of missed alarms given by T^{RR}

TABLE IV
PARAMETER ESTIMATION VIA EM ALGORITHM AND DATA FITTING EVALUATION FOR THE THREE CONSIDERED MIXTURE DENSITIES.

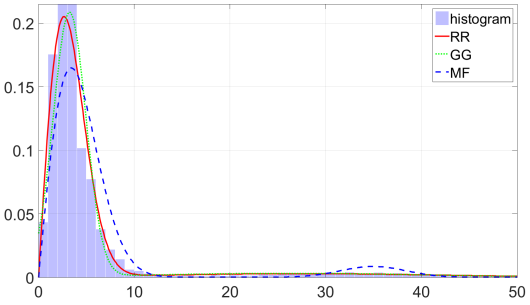
mix	estimated parameters					χ^2_P	KS
Dataset A							
RR	α	b	ν	σ		0.0002	0.0004
	0.80	2.50	53.89	25.04			
GG	α	μ_1	σ_1	μ_2	σ_2	0.0184	0.0561
	0.80	3.11	1.61	59.34	23.95		
MF	α	ν_1	ν_2	β		0.9045	0.5142
	0.83	$\approx$0	66.32	7.20			
Dataset B							
RR	α	b	ν	σ		0.0136	0.0362
	0.92	9.36	60.24	17.37			
GG	α	μ_1	σ_1	μ_2	σ_2	0.0420	0.0836
	0.90	11.10	5.97	52.73	22.57		
MF	α	ν_1	ν_2	β		0.0239	0.0590
	0.93	0.06	66.91	9.78			
Dataset C							
RR	α	b	ν	σ		0.0215	0.0400
	0.90	2.67	23.84	14.90			
GG	α	μ_1	σ_1	μ_2	σ_2	0.0500	0.0778
	0.89	3.25	1.70	26.35	14.13		
MF	α	ν_1	ν_2	β		0.1414	0.1880
	0.93	$\approx$0	34.82	3.41			



(a) Dataset A



(b) Dataset B



(c) Dataset C

Fig. 7. Histograms of the magnitude of the difference image and plot of the estimated densities.

TABLE V
PERFORMANCE OF CHANGE DETECTION BASED ON THE THRESHOLDING OF THE MAGNITUDE IMAGE FOR DIFFERENT VALUES OF THE THRESHOLD.

Threshold	missed (e_m)	false (e_f)	overall (e_o)
Dataset A			
$T^{RR} = 10.12$	699 (0.83%)	99 (0.03%)	798 (0.19%)
$T^{GG} = 8.82$	530 (0.63%)	681 (0.20%)	1211 (0.29%)
$T^{MOE} = 10.40$	735 (0.88%)	56 (0.02%)	791 (0.19%)
Dataset B			
$T^{RR} = 33.95$	36 (0.48%)	1784 (1.54%)	1820 (1.47%)
$T^{GG} = 28.02$	9 (0.12%)	3973 (3.42%)	3982 (3.22%)
$T^{MOE} = 47.11$	356 (4.76%)	414 (0.36%)	770 (0.62%)
Dataset C			
$T^{RR} = 9.92$	956 (2.75%)	7805 (1.86%)	8761 (1.93%)
$T^{GG} = 8.63$	727 (2.09%)	12856 (3.06%)	13583 (2.99%)
$T^{MOE} = 15.23$	1964 (5.65%)	2657 (0.63%)	4621 (1.02%)

results in a much less noisy change map compared to the one given by T^{GG}. In the optimal case the change maps are very clean. Since in these cases the number of missed alarms is the highest one, some details of the changes are lost (e.g., the thin boundaries of the lake in Dataset B and a portion of the fire in Dataset C). A possible improvement of the change detection maps could be obtained by using the distributions of the classes obtained by the proposed technique within a context-sensitive approach (e.g., based on MRFs [3]).

V. CONCLUSIONS

In this paper we addressed the problem often encountered in image analysis of the estimation of the parameters of a Rayleigh-Rice mixture density. The problem has been studied in the framework of Change Vector Analysis (CVA) for binary change detection in multitemporal and multispectral images. Here, under proper hypotheses, the distribution of the magnitude of the difference image can be theoretically modeled by a Rayleigh-Rice mixture density. The Rayleigh density describes the distribution of unchanged pixels, whereas the Rice density describes the distribution of the changed pixels. Parameter estimates are used to solve the binary change detection problem in a Bayesian context.

In the paper, a general implementation of an EM-type algorithm for the estimation of mixture and shape parameters of a Rayleigh-Rice mixture density is given. The proposed method enjoys good theoretical properties. First, statistical independence of parameters allowed us to define the algorithm in an iterative way, which results fast, easy to implement and not depending on specific optimization routines. Detailed analysis of accuracy and convergence properties of the algorithm is given. Second, in the considered model all the statistical parameters are free. We remark here that, because of its flexibility, the proposed method can be used for addressing a large variety of practical problems involving the Rayleigh-Rice density.

In the experimental part of the paper, the effectiveness of the Rayleigh-Rice model in solving the binary CD problem is demonstrated. Tests have been conducted on both synthetic and real datasets consisting of bi-temporal pairs of multispectral remote sensing images. Among other statistical models proposed in literature, the Rayleigh-Rice proved to be the one that better fits the distribution of the magnitude difference image in CVA. The change detection is significantly improved when compared to state-of-the-art method based on Gaussian modeling. In particular, the overall error of detection is always approximately halved. Moreover, even though the theoretical model is given for one single change class, the algorithm is robust to the presence of minor additional changes.

Future developments of this work include: (1) extend the use of the proposed model in the context of a multiple-change detection problem and assess the improvement in the change detection accuracy when different classes of change are considered; (2) integrate the class distributions obtained by the proposed method in a context-sensitive approach.

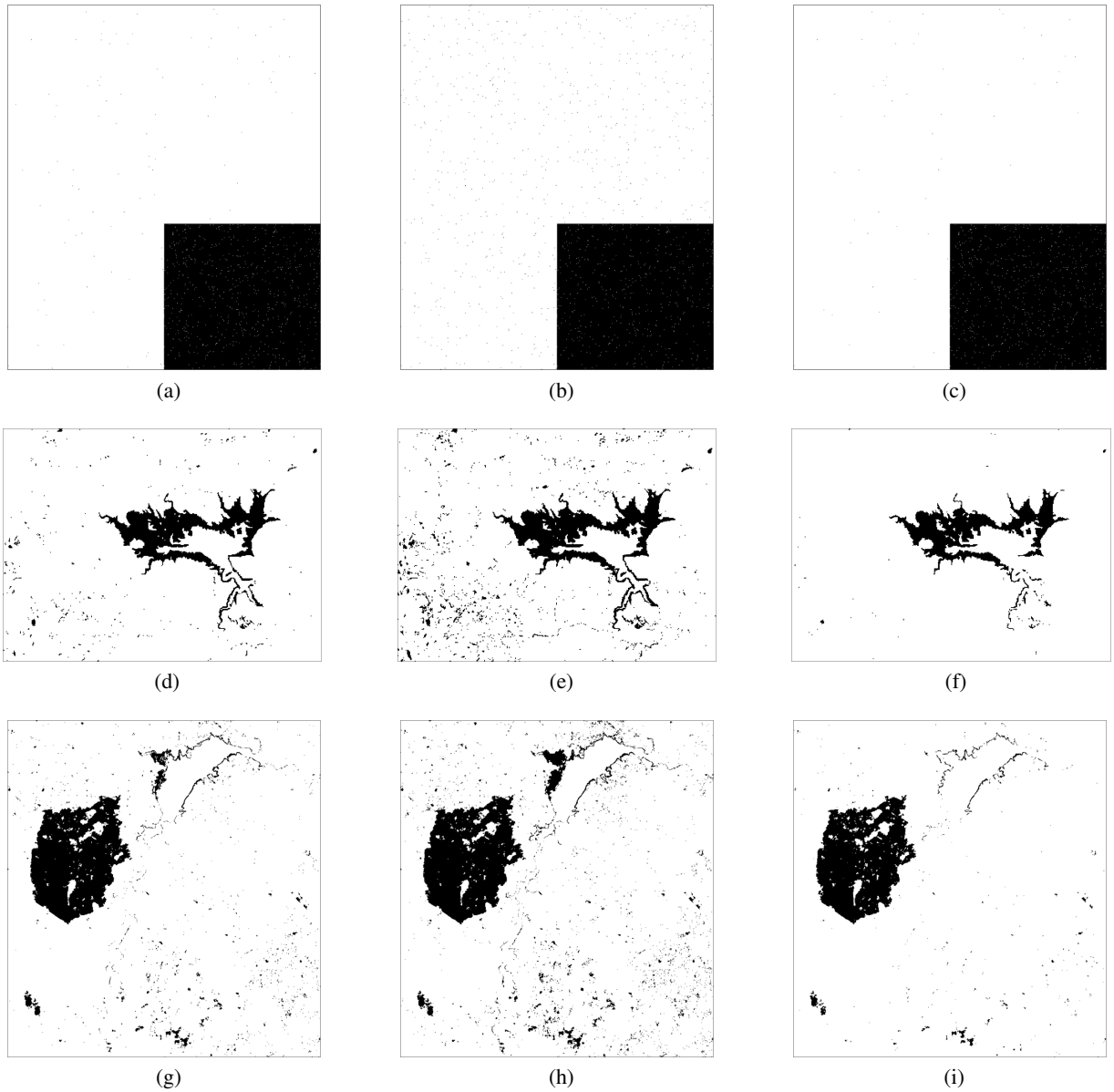


Fig. 8. Change detection maps obtained by thresholding the magnitude image using (a,d,g) T^{RR}, (b,e,h) T^{GG} and (c,f,i) T^{MOE}. In black are the estimated changed pixels W_c, in white the estimated unchanged pixels W_n.

VI. Acknowledgments

This work was supported by the Italian Ministry of Education, University, and Research (MIUR) under Grant 2012L48PE5 (PRIN 2012). The authors are grateful to the anonymous reviewers for their criticism, which helped us in improving the work.

Appendix A
Mathematical notions and technical results

A. Basic notions about modified Bessel functions

We recall here some facts about modified Bessel functions which are crucial for the development of the method that is presented in the paper. Let $I_m(x)$ be the m-th order modified Bessel function of first kind [32]. Derivatives

of $I_m(x)$ satisfy (among many other) the following recurrence rules:

$$\frac{d}{dx}I_0(x) = I_1(x), \tag{25}$$

$$\frac{d}{dx}(x^m I_m(x)) = x^m I_{m-1}(x). \tag{26}$$

In order to simplify the presentation of the results of next sections, we define two functions that appear frequently in the computations:

$$\mathrm{J}_1 := \frac{\mathrm{I}_1}{\mathrm{I}_0} \text{ and } \mathrm{J}_2 := {\mathrm{J}_1}^2 - 1. \tag{27}$$

B. Derivation of the conditional log-expectation

Let us derive the energy $\boldsymbol{\ell}(\Psi)$ with respect to its parameters to obtain (19). Computations for $\partial/\partial b$ are the same to that of $\partial/\partial\sigma$ if σ is replaced by b and $\nu = 0$. It is a trivial fact that by deriving $\boldsymbol{\ell}(\Psi)$ with respect to the Rice parameters the terms related to the Rayleigh distribution vanish. Therefore, without loss of generality we can consider

$$\boldsymbol{\ell}(\Psi) = \sum_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)\log\left(\left(1-\alpha\right)\mathrm{p}\left(x|\omega_2,\theta_2\right)\right). \tag{28}$$

Because of summation over x and the $\log$, we can write

$$\boldsymbol{\ell}(\Psi) = \sum_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)(\log A + \log B), \tag{29}$$

where $A := \frac{x}{\sigma^2}\exp\left(-\frac{x^2+\nu^2}{2\sigma^2}\right)$ and $B := \mathrm{I}_0\left(\frac{x\nu}{\sigma^2}\right)$. Derivations involving A and B can be computed by using (25) and by applying the chain rule. We get

$$\frac{\partial \log A}{\partial \nu} = -\frac{\nu}{\sigma^2}, \quad \frac{\partial \log A}{\partial \sigma} = \frac{x^2+\nu^2}{\sigma^3} - \frac{2}{\sigma},$$
$$\frac{\partial \log B}{\partial \nu} = \frac{x}{\sigma^2}\mathrm{J}_1\left(\frac{x\nu}{\sigma^2}\right), \quad \frac{\partial \log B}{\partial \sigma} = -\frac{2x\nu}{\sigma^3}\mathrm{J}_1\left(\frac{x\nu}{\sigma^2}\right).$$

Final derivatives are obtained by putting together the results.

C. Asymptotic analysis of the maximum-likelihood estimation

The Rayleigh-Rice mixture density $\mathrm{p}\left(\rho|\Psi\right)$ is twice differentiable with respect to its parameters and enjoys some regularity properties that allow us to write its information matrix (see Lemma 5.3 of [35] and its generalization to the multi-parameter case, p.125) as

$$I(\Psi)_{i,j} = -E\left[\frac{\partial^2}{\partial\psi_i\partial\psi_j}\log\,\mathrm{p}\left(\rho|\Psi\right)\Big|\Psi\right] \quad i,j = 1,\ldots,4 \tag{30}$$

where we recall that $\Psi = (\alpha, b, \nu, \sigma)$. Since the information matrix of the Rician distribution is diagonal [17], also $I(\Psi)$ is diagonal. Thus, the parameters of the mixture density are mutually orthogonal.

Orthogonality brings some nice statistical properties [36]: (1) asymptotic estimates of the parameters are independent; (2) the asymptotic standard error for estimating one parameter does not depend on the knowledge of the others; (3) the maximum likelihood estimate of ψ_i given ψ_j varies only slowly with ψ_j. All these properties are fundamental for the formalization of an algorithm that iteratively updates the ML estimates of the mixture parameters.

D. Iterative method for the solution of the non linear gradient system

Iterative formulas for approximating the solution of the gradient system (18) are given according to the method of subsequent approximations applied to a fixed point equation [37]. After some analytical manipulations of equations (19) we can see that a point Ψ^* satisfying $\nabla\boldsymbol{\ell}(\Psi^*) = 0$ also satisfies $\Psi^* = \varphi(\Psi^*)$, where $\varphi(\Psi)$, the so-called iterative function, is made up of four components:

$$\begin{aligned}
\varphi_1(\Psi) &= \frac{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_1|x,\Psi\right)}{\sum\limits_{x\in\mathbf{x}} 1} \\
\varphi_2(\Psi) &= \sqrt{\frac{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_1|x,\Psi\right)x^2}{2\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_1|x,\Psi\right)}} \\
\varphi_3(\Psi) &= \frac{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)\mathrm{J}_1\left(\frac{x\nu}{\sigma^2}\right)x}{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)} \\
\varphi_4(\Psi) &= \sqrt{\frac{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Theta\right)\left[x^2+\nu^2-2x\nu\mathrm{J}_1\left(\frac{x\nu}{\sigma^2}\right)\right]}{2\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Theta\right)}}.
\end{aligned} \tag{31}$$

A small remark is needed here. Since $0 < I_1(y)/I_0(y) < 1$ for every $y > 0$, the argument of the square root in (31) is always non-negative, hence φ_4 is well defined.
Given an initial point Ψ^0, the iterative method for solving the fixed point equation $\Psi = \varphi(\Psi)$ consists in applying the iterative rule

$$\Psi^{k+1} = \varphi(\Psi^k). \tag{32}$$

The method of subsequent approximations is shown to be convergent with at least linear speed whenever the spectral radius of the Jacobian matrix $J\varphi$, which in our case reduces to

$$J\varphi = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \dfrac{\partial\varphi_3}{\partial\nu} & \dfrac{\partial\varphi_3}{\partial\sigma} \\ 0 & 0 & \dfrac{\partial\varphi_4}{\partial\nu} & \dfrac{\partial\varphi_4}{\partial\sigma} \end{pmatrix}, \tag{33}$$

computed at the exact solution Ψ^*, is less than one. The computation of these non-zero derivatives can be performed by applying (26). We have

$$\begin{aligned}
\frac{\partial\varphi_3}{\partial\nu}(\Psi) &= \frac{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)\left\{-\frac{x^2}{\sigma^2}\mathrm{J}_2\left(\frac{x\nu}{\sigma^2}\right)-\frac{x}{\nu}\mathrm{J}_1\left(\frac{x\nu}{\sigma^2}\right)\right\}}{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)} \\
\frac{\partial\varphi_3}{\partial\sigma}(\Psi) &= \frac{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)\left\{\frac{2x^2\nu}{\sigma^3}\mathrm{J}_2\left(\frac{x\nu}{\sigma^2}\right)+\frac{2x}{\sigma}\mathrm{J}_1\left(\frac{x\nu}{\sigma^2}\right)\right\}}{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)} \\
\frac{\partial\varphi_4}{\partial\nu}(\Psi) &= \frac{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)\left\{\frac{x^2\nu}{\sigma^2}\mathrm{J}_2\left(\frac{x\nu}{\sigma^2}\right)+1\right\}}{\varphi_4(\Psi)\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)} \\
\frac{\partial\varphi_4}{\partial\sigma}(\Psi) &= \frac{\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)\left\{\frac{x^2\nu^2}{\sigma^3}\mathrm{J}_2\left(\frac{x\nu}{\sigma^2}\right)+\frac{x\nu}{\sigma}\mathrm{J}_1\left(\frac{x\nu}{\sigma^2}\right)\right\}}{\varphi_4(\Psi)\sum\limits_{x\in\mathbf{x}} \mathrm{p}\left(\omega_2|x,\Psi\right)}
\end{aligned} \tag{34}$$

The spectral radius $rad(J\varphi(\Psi))$ is defined as the maximum absolute value of the eigenvalues of $J\varphi(\Psi)$. The spectral radius is a measure of the rate at which the error in approximating the fixed point reduces between two consecutive iterations.

The fixed point iteration can be accelerated to quadratic convergence by using the more general Newton's method (or one of its many derivations)

$$\Psi^{k+1} = \Psi^k - H\boldsymbol{\ell}(\Psi^k)^{-1}\boldsymbol{\ell}(\Psi^k), \tag{35}$$

where $H(.)$ is the Hessian operator. However, a low number of iterations is at the cost of one matrix inversion for each iteration, which may not be desirable in applications. In particular, for CD in multispectral images the simple iteration (32) performed excellently without the need of accelerators.

REFERENCES

[1] R. J. Radke, S. Andra, O. Al-Kofani, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, March 2005.

[2] B. Demir, F. Bovolo, and L. Bruzzone, "Classification of time series of multispectral images with limited training data," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3219–3233, August 2013.

[3] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[4] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 772–776, 2009.

[5] S. L. Hégarat-Mascle and R. Seltz, "Automatic change detection by evidential fusion of change indices," *Remote Sensing of Environment*, vol. 91, no. 3, pp. 390–404, 2004.

[6] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, January 2007.

[7] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2196–2212, May 2012.

[8] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sensing of Environment*, vol. 64, no. 1, pp. 1–19, 1998.

[9] V. Ortiz-Rivera, M. Vélez-Reyes, and B. Roysam, "Change detection in hyperspectral imagery using temporal principal components," in *Proc. SPIE 6233 Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XII*, Orlando (Kissimmee), May 2006.

[10] S. Singh and R. Talwar, "Review on different change vector analysis algorithms based change detection techniques," in *Image Information Processing (ICIIP), 2013 IEEE Second International Conference on*, 2013, pp. 136–141.

[11] R. D. Johnson and E. S. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *International Journal of Remote Sensing*, vol. 19, no. 3, pp. 411–426, 1998.

[12] R. Simoes and C. Slump, "Change detection and classification in brain MR images using change vector analysis," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 7803–7807.

[13] S. Liu, L. Bruzzone, F. Bovolo, and P. Du, "Hierarchical unsupervised change detection in multitemporal hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 244–260, January 2015.

[14] L. Bruzzone and D. F. Prieto, "An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 452–466, April 2002.

[15] M. Baisantry, D. S. Negi, and O. P. Manocha, "Change vector analysis using enhanced PCA and inverse triangular function-based thresholding," *Defence Science Journal*, vol. 62, no. 4, pp. 236–242, 2012.

[16] A. Hennemuth, A. Seeger, O. Friman, S. Miller, B. Klumpp, and S. O. H. O. Peitgen, "A comprehensive approach to the analysis of contrast enhanced cardiac MR images," *IEEE Transactions on Medical Imaging*, vol. 27, no. 11, pp. 1592–1610, 2008.

[17] O. T. Karlsen, R. Verhagen, and W. M. Bovee, "Parameter estimation from rician-distributed data sets using a maximum likelihood estimator: Application to T1 and perfusion measurements," *Magnetic resonance in medicine*, vol. 41, no. 3, pp. 614–623, 1999.

[18] T. Benedict and T. Soong, "The joint estimation of signal and noise from the sum envelope," *IEEE Transactions on Information Theory*, vol. 13, no. 3, pp. 447–454, 1967.

[19] J. Sijbers, A. J. D. Dekker, J. V. Audekerke, M. Verhoye, and D. V. Dyck, "Estimation of the noise in magnitude MR images," *Magnetic Resonance Imaging*, vol. 16, no. 1, pp. 87–90, 1998.

[20] J. J. Sijbers, A. J. den Dekker, P. Scheunders, and D. V. Dyck, "Maximum-likelihood estimation of Rician distribution parameters," *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 357–361, 1998.

[21] S. O. Rice, "Mathematical analysis of random noise," *Bell System Technical Journal*, vol. 23, no. 3, pp. 282–332, 1944.

[22] K. K. Talukdar and W. D. Lawing, "Estimation of the parameters of the Rice distribution," *The Journal of the Acoustical Society of America*, vol. 89, no. 3, pp. 1193–1197, 1991.

[23] J. Sijbers and A. J. D. Dekker, "Maximum likelihood estimation of signal amplitude and noise variance from MR data," *Magnetic Resonance in Medicine*, vol. 51, no. 3, pp. 586–594, 2004.

[24] C. F. Carobbi and M. Cati, "The absolute maximum of the likelihood function of the Rice distribution: Existence and uniqueness," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 4, pp. 682–689, 2008.

[25] L. Lauwers, K. Barbé, W. V. Moer, and R. Pintelon, "Estimating the parameters of a Rice distribution: A Bayesian approach," in *IEEE Instrumentation and Measurement Technology Conference (I2MTC'09)*, May 2009, pp. 114–117.

[26] R. Maitra and D. Faden, "Noise estimation in magnitude MR datasets," *IEEE Transactions on Medical Imaging*, vol. 28, no. 10, pp. 1615–1622, 2009.
[27] R. Maitra, "On the Expectation-Maximization algorithm for Rice-Rayleigh mixtures with application to noise parameter estimation in magnitude MR datasets," *Sankhya B*, vol. 75, no. 2, pp. 293–318, 2013.
[28] J. Rajan, D. Poot, J. Juntu, and J. Sijbers, "Noise measurement from magnitude MRI using local estimates of variance and skewness," *Physics in medicine and biology*, vol. 55, no. 16, p. N441, 2010.
[29] R. D. Nowak, "Wavelet-based Rician noise removal for magnetic resonance imaging," *IEEE Transactions on Image Processing*, vol. 8, no. 10, pp. 1408–1419, 1999.
[30] C. H. Chen, *Image Processing for Remote Sensing*, 1st ed. Boca Raton, FL: CRC Press, 2007.
[31] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 609–630, 2013.
[32] G. N. Watson, *A treatise on the Theory of Bessel Functions*, 2nd ed. Cambridge University Press, 1966.
[33] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
[34] R. A. Redner and H. F. Walker, "Mixture densities, Maximum Likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, April 1984.
[35] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 1998, vol. 31.
[36] D. R. Cox and N. Reid, "Parameter orthogonality and approximate conditional inference," *J. Roy. Statist. Soc. Ser. B*, vol. 49, no. 1, pp. 1–39, 1987, with a discussion.
[37] R. L. Burden and J. D. Faires, *Numerical Analysis*, 9th ed. Boston, MA: Brooks/Cole, Cengage Learning, 2011.