Title: A theoretical framework for change detection based on a compound multiclass statistical model of the difference image

This paper appears in: IEEE Transactions on Geoscience and Remote Sensing

Date of Publication: 23 October 2017

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

Volume: 56, Issue: 02

Page(s): 1129 – 1143

DOI: 10.1109/TGRS.2017.2759663

A theoretical framework for change detection based on a compound multiclass statistical model of the difference image

Massimo Zanetti, *Student Member, IEEE,* and Lorenzo Bruzzone, *Fellow, IEEE*

Abstract

The change detection (CD) problem is very important in the remote sensing domain. The advent of a new generation of multispectral (MS) sensors has given rise to new challenges in the development of automatic CD techniques. In particular, typical approaches to CD are not able to well model and properly exploit the increased radiometric resolution characterizing new data as this results in a higher sensitivity to the number of natural classes that can be statistically modeled in the images. In this paper, we introduce a theoretical framework for the description of the statistical distribution of the difference image as a compound model where each class is determined by temporally-correlated class transitions in the bitemporal images. The potential of the proposed framework is demonstrated on the very common problem of binary change detection based on setting a threshold on the magnitude of the difference image. Here, under some simplifying assumptions, a multiclass distribution of the magnitude feature is derived and an unsupervised method based on the Expectation-Maximization (EM) algorithm and Bayes decision is proposed. Its effectiveness is demonstrated on a large variety of datasets from different MS sensors. In particular, experimental tests confirm that: 1) the fitting of the magnitude distribution significantly improves if compared to already existing models, 2) the overall change detection error is close to the optimal value.

Index Terms

Multispectral multitemporal images, change detection, change vector analysis, EM algorithm, Rician distributions.

I. Introduction

CHANGE detection (CD) is the process aimed at the identification of changes in multitemporal remote sensing data [1]. The increasing availability of multitemporal images makes the development of automatic and efficient change detection techniques of primary interest to address new challenges arising in environmental studies, Earth monitoring, damage assessment, etc, [2], [3].

Multispectral (MS) images provide information about the observed scene both in the spatial and the spectral domains. Given two coregistered multitemporal images acquired over the same geographical area, classical CD techniques assign to each pixel a label of type *unchanged* or *changed*. In the context of multiple CD, changed pixels are further divided into categories representing different kinds of changes [4], [5]. Being CD essentially a classification problem, there exist in literature both supervised and unsupervised approaches to address it. It is a matter of fact that, in MS images natural classes can be described in statistical terms. Indeed, radiance values show evident statistical accordance given their originating natural classes (not necessarily semantic classes). Post-classification methods first classify the two images separately and then perform CD by comparing the two classification maps [6]. The main issue is that accuracy is not very high, as this simple approach does not take into account the temporal dependence of the image pair. Therefore, compound-classification methods are introduced, that incorporate temporal dependency in the statistical formulation of the classification problem [7], [8]. Further improvements can be obtained by applying active-learning techniques for the training phase [9], or domain-transfer techniques if reference information is available only at one date [10].

The above-mentioned methods do not exploit a direct comparison of the single time images. In particular, the spectral difference at the pixel level is not explicitly considered. Nevertheless, the difference image carries useful information for the solution of the CD problem. In Change Vector Analysis (CVA), changes are associated to

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.

spectral variations of pixels after image differencing [11], [12]. The orientation of the spectral difference vectors can be used to separate between different kinds of changes [4]. However, much attention is devoted to CVA for the specific purpose of unsupervised binary change detection, i.e., discriminating between unchanged and changed pixels in a bitemporal image pair. Former change detection approaches assume the difference image as populated by two general classes representing the unchanged and changed pixels. Decision is then performed via statistical models by: 1) a-posteriori inference based on the Expectation-Maximization (EM) algorithm [13] also in the case of heterogeneous sensors [14], 2) significance test ot hypothesis test [15], [16], or, 3) local gradual descent [17]. In the context of CVA, the two-class model is used for the representation of the magnitude of the difference image [18]–[20]. Driven by the general intuition that in the difference image pixels having small magnitude are likely to be unchanged and pixels with high magnitude are changed, the magnitude information shows to be very informative. Indeed, many studies about binary change detection involve such feature, and the classical bi-modal behavior of the histogram of the magnitude of the difference image is usually interpreted as follows: the left mode (low magnitude) represents the population of unchanged pixels, whereas the right mode (high magnitude) represents the changed pixels. In [18], the distribution of the magnitude is empirically approximated by a mixture of two Gaussians, and inference is justified by means of Bayes decision. The binary decision is performed by thresholding the magnitude at the value that produces the minimum overall error of classification based on the Gaussian mixture (GM) model. Recently, it has been shown [19] that by assuming the Gaussian distribution of natural classes in the original images, the magnitude distribution can be approximated as Rayleigh for the unchange class and Rice for the change class. Such model is deeply investigated in [20], where a numerical method based on the EM algorithm for estimating the parameters of a Rayleigh-Rice (RR) mixture is developed. The parameter estimates are then used to compute a threshold by following a Bayes decision rule. Experiments on both real and synthetic datasets show that the RR model outperforms the classical GM. However, it is also observed that estimated densities do not fit accurately the magnitude distribution and the computed thresholds for binary decision are still different from the optimal ones. In this perspective, it seems that the two-class unchange/change model has some limitations due to the non-negligible effects caused by some subtle components that actually populate the mixture. Indeed, new generation sensors are characterized by increased radiometric resolution compared to the old ones. As a consequence, multispectral data provided by modern satellite missions present greater statistical variability and the typical simplifying two-class model [13], [17]–[20] for the difference image may no longer be valid to well describe the CD problem in last generation data. At the best of our knowledge, there is no study that extends the statistical interpretation of CVA to the multi-class case. Indeed, this paper addresses this specific problem.

The main motivation of the present paper is to introduce a theoretical framework for the description of the statistical distribution of the difference image and to propose a general definition of the change detection problem that extends CVA to the multiclass case. In the proposed model, both the unchange and the change classes are multiple, and their statistical distributions are explicitly derived starting from a well-known joint distribution model. On the one hand, the proposed model has strong connections with the supervised compound-classification approach to CD [7], [8], and allows for a formal extension of this method for the unsupervised study of the difference image. On the other hand, being a generalization of the statistical description of CVA, our model can be used to generalize the already existent methods based on CVA to the multiclass case. The novelty lies in the fact that the proposed model explains and justifies the multiclass nature of the difference image as a compound model where each class is determined by class transitions in the bitemporal images, not independently, but exploiting the temporal correlation of the image pair. This general model represents a valid theoretical starting point for the study and understanding of the CD problem in MS images. By leveraging on its general character, the proposed framework opens to several interpretations and applications.

To demonstrate the effectiveness and the generalization capabilities of the proposed model, a specific CD problem is studied from a theoretical viewpoint and a novel method is derived. More specifically, we address the simple (yet widely studied in the literature) problem of binary change detection based on setting a threshold on the magnitude of the difference image. It is a matter of fact that thresholding methods are very popular in literature. The use of classic methods on images with increased radiometric resolution showed lower accuracy than expected. By properly exploiting our theoretical study of the difference image a novel statistical model for the magnitude is derived, which is more appropriate for real data and does not constraint unchanged pixels to have the same statistical behavior. Then, a framework for binary decision based on Bayes theory is presented and a numerical method for parameter estimation and magnitude thresholding (which is a generalization of the one presented in

[20]) is proposed. Experiments involve multispectral images from Landsat-5, Landsat-7, Landsat-8 and Sentinel-2 satellites. Results show that, the proposed method better fits the real data and improves the detection performance of state-of-the-art methods. Moreover, in all experiments it is observed that the computed threshold is very close to the optimal one.

The paper is structured as follows. In Section II the statistical model describing the difference image is proposed and analyzed in detail. In Section III, firstly we derive the statistical distribution of the magnitude difference image in the multi-class case, and then, we propose an unsupervised method based on the EM algorithm and Bayes decision to perform binary change detection. Experimental results are presented in Section IV. Finally, Section V draws the conclusions of the work. In Appendix A we recall the mathematical notation and symbols used throughout the paper.

II. Statistical study of the difference image

In this section, we provide a theoretical derivation of the statistical distribution of the difference image starting from the hypothesis of Gaussian distribution of natural classes in the single time images. Based on this model, we give a formal definition of the multiple unchange and change classes. On the one hand, the proposed model allows for formally describing some approaches already present in literature. On the other hand, it also provides a framework for extending the statistical interpretation of change detection to a larger set of cases.

A. A compound multi-class mixture model of the difference image

Let $\mathbf{X}^t$, with $t = 1, 2$, be two multispectral images acquired over the same area at two different times. Each image has B bands, therefore each pixel value at location (i, j) is a B-dimensional vector $\mathbf{X}^t(i, j) \in \mathbb{R}^B$, where $t = 1, 2$. In a general statistical interpretation, the multispectral image formation is modeled as joint realization of certain B-dimensional random variables. To each pixel location (i, j) is assigned a random variable (X^t, Φ^t) where $X^t \in \mathbb{R}^B$ is associated to the observed spectral values of the pixel at time t, and $\Phi^t \in \{\phi_1, \dots, \phi_C\}$ is associated to its class label at time t. Hereafter, we will assume that these random variables are i.i.d., thus, the statistical distribution of the images can be fully described by defining one random variable (X^t, Φ^t) for each single time t. Two remarks about the notation used in this paper. First, probability density functions are denoted by $\mathrm{p}(.)$ for convenience. Second, given $t = 1, 2$ and $h = 1, \dots, C$, the probabilistic event $\Phi^t = \phi_h$ is written in a more compact way as ϕ_h^t.

A typical and commonly used [7]–[10] joint probabilistic model for X^1, X^2 assumes that the couple (X^1, X^2) depends on (Φ^1, Φ^2). By marginalizing w.r.t. the class variables, this leads to the following distribution

$$\mathrm{p}(x^1, x^2) = \sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2)\, \mathrm{p}(x^1, x^2 | \phi_h^1, \phi_k^2). \tag{1}$$

It is reasonable to assume that the realization of the variable X^t only depends on its associated class label Φ^t, for each $t = 1, 2$. Therefore, we slightly simplify our joint model by assuming the following conditional independences[1] [8]:

$$\begin{gathered} X^1 \perp X^2 \,|\, (\Phi^1, \Phi^2) \\ X^1 \perp \Phi^2 \,|\, \Phi^1 \\ X^2 \perp \Phi^1 \,|\, \Phi^2 \end{gathered} \tag{2}$$

With these assumptions we can split the joint conditional distribution appearing in (1) to obtain

$$\mathrm{p}(x^1, x^2) = \sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2)\, \mathrm{p}(x^1 | \phi_h^1)\, \mathrm{p}(x^2 | \phi_k^2). \tag{3}$$

In the most general case, the two images might not present the same set of observable classes: we assume that $\{\phi_1, \dots, \phi_C\}$ is the joint set of class labels which are observable in both images, where $C \geq 1$ is the total number

[1] Following a standard notation, for $A \perp B \,|\, C$ we say that A and B are conditionally independent given C. In words, the knowledge of C makes A and B independent.

of classes. The fact that some classes might not be observed in one of the single time images is formalized by setting to zero the corresponding class prior probabilities. For instance, if class ϕ_h is not observable at time $t = 1$, then $\mathrm{p}(\phi_h^1, \phi_k^2) = 0$ for all $k = 1, \dots, C$. A similar argument works for a class that is not observable at time $t = 2$. Therefore, we let $0 \leq \mathrm{p}(\phi_h^1, \phi_k^2) \leq 1$ and

$$\sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2) = 1 \tag{4}$$

in order for (3) to be consistent as a probability model. Now, let us assume that natural classes are distributed as multidimensional Gaussians in the two images. This is a common assumption in the literature when medium-high multispectral [21] or hyperspectral [22] images are considered, while it is more critical for Very High Resolution (VHR) images. A natural class that is observed at both times might be described by different statistical parameters in the two images due to seasonal effects, different radiometric conditions, co-registration errors, etc. Therefore, we model them as

$$X^t | \phi_h^t \sim \mathcal{N}(\mu_h^t, \Sigma_h^t) \tag{5}$$

where μ_h^t, Σ_h^t are the mean vectors and the covariance matrices of the class ϕ_h observed at time $t = 1, 2$.

Under the hypothesis of a joint model as in (3)–(5), the distribution of the difference $D := X^1 - X^2$ can be written as a mixture of (at most) C^2 Gaussian components (technical details are given in Appendix B). The density function of the difference turns out to be

$$\mathrm{p}(d) = \sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2) \, \mathcal{N}(d; \mu_h^1 - \mu_k^2, \Sigma_h^1 + \Sigma_k^2) \tag{6}$$

where the mixture terms are the class prior probabilities $\mathrm{p}(\phi_h^1, \phi_k^2)$.

B. Physical interpretation of the difference mixture model

The distribution (6) formalizes the intuitive concept that the difference image is populated by a set of classes, each one representing a possible class-by-class matching among the classes that populate the single time images. As mentioned above, we allow the same class ϕ_h to have different parameters μ_h^t and Σ_h^t in correspondence of the two acquisitions at times $t = 1, 2$. Every pixel that can be considered as drawn from the distribution $\mathcal{N}(\mu_h^1, \Sigma_h^1)$ in $\mathbf{X}^1$ and from the distribution $\mathcal{N}(\mu_h^2, \Sigma_h^2)$ in $\mathbf{X}^2$ is an unchanged pixel, as it belongs to the same natural class ϕ_h. It follows that each class ϕ_h has its own *unchange* behavior if and only if the class is observed in both images. Conversely, a pixel that can be considered as drawn from the distribution $\mathcal{N}(\mu_h^1, \Sigma_h^1)$ in $\mathbf{X}^1$ and from the distribution $\mathcal{N}(\mu_k^2, \Sigma_k^2)$ in $\mathbf{X}^2$, with $h \neq k$, is a changed pixel that changed its class from ϕ_h to ϕ_k. As a consequence, each class may have at most $C - 1$ different *change* behaviors.

In order to better understand which classes are changed and which ones are not changed, we can reason about priors. Each mixing term $\mathrm{p}(\phi_h^1, \phi_k^2)$ is exactly the joint probability that class ϕ_h is observed at time $t = 1$ and class ϕ_k is observed at time $t = 2$. These probabilities can be arranged in a useful way into a square matrix

$$\mathbf{Q} = \begin{pmatrix} \mathrm{p}(\phi_1^1, \phi_1^2) & \mathrm{p}(\phi_1^1, \phi_2^2) & \dots & \mathrm{p}(\phi_1^1, \phi_C^2) \\ \mathrm{p}(\phi_2^1, \phi_1^2) & \mathrm{p}(\phi_2^1, \phi_2^2) & \dots & \mathrm{p}(\phi_2^1, \phi_C^2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{p}(\phi_C^1, \phi_1^2) & \mathrm{p}(\phi_C^1, \phi_2^2) & \dots & \mathrm{p}(\phi_C^1, \phi_C^2) \end{pmatrix}. \tag{7}$$

If all the entries of the matrix $\mathbf{Q}$ are non-null, the distribution function $\mathrm{p}(d)$ has exactly C^2 mixture class components. If class ϕ_h is not observable at time $t = 1$, then the h-th row of the matrix is null. Similarly, if class ϕ_k is not observable at time $t = 2$, then the k-th column of matrix $\mathbf{Q}$ is null. Other simplifications of matrix $\mathbf{Q}$ can be done (i.e., setting to zero some of its entries) if prior knowledge about the studied scene is available. Just as an example, if we know that in the considered image pair it is impossible to find a change from class ϕ_h to class ϕ_k, then its corresponding prior probability can be set to zero, i.e., $\mathrm{p}(\phi_h^1, \phi_k^2) = 0$. All the non-null entries of the matrix $\mathbf{Q}$ define the prior probabilities of the mixture components that populate the difference image. Each of these mixture components describe the statistical distribution of the pixels that belong to class ϕ_h at time $t = 1$ and to class ϕ_k at time $t = 2$. It follows that, if $h = k$, then these pixels are not changed, whereas if $h \neq k$, then the pixels have changed their class from ϕ_h to ϕ_k. In both cases, their distributions are given explicitly. This interpretation allows us to give a formal definition of the unchange and change classes. in the difference image, as multiple classes.

Definition 1 (Unchange class)**.** *Each pixel that belongs to class ϕ_h both at times $t = 1$ and $t = 2$ is said to belong to the* unchange class *ω_h, where $h \in \{1, \dots, C\}$. The distribution of the unchange class ω_h in the difference image is Gaussian with parameters*

$$D|\omega_h \sim \mathcal{N}(\mu_h^1 - \mu_h^2, \Sigma_h^1 + \Sigma_h^2).$$

The set of all unchange classes is denoted by Ω_n, and it contains all classes that are associated to non-null diagonal entries of the matrix $\mathbf{Q}$*:*

$$\Omega_n := \{\omega_h : \mathrm{p}(\phi_h^1, \phi_h^2) \neq 0\}.$$

Definition 2 (Change class)**.** *Each pixel that belongs to class ϕ_h at time $t = 1$ and to class ϕ_k at time $t = 2$ is said to belong to the* change class *ω_{hk}, where $h, k \in \{1, \dots, C\}$ and $h \neq k$. The distribution of the change class ω_{hk} in the difference image is Gaussian with parameters*

$$D|\omega_{hk} \sim \mathcal{N}(\mu_h^1 - \mu_k^2, \Sigma_h^1 + \Sigma_k^2).$$

The set of all change classes is denoted by Ω_c, and it contains all classes that are associated to non-null off-diagonal entries of the matrix $\mathbf{Q}$*:*

$$\Omega_c := \{\omega_{hk} : \mathrm{p}(\phi_h^1, \phi_k^2) \neq 0, h \neq k\}.$$

As a simple corollary of the two definitions above, we have that the distribution of the difference image can be split into two parts as

$$\mathrm{p}(d) = K_n \, \mathrm{p}(d|\Omega_n) + K_c \, \mathrm{p}(d|\Omega_c), \tag{8}$$

where the first term describes the statistical distribution of the multiple unchange class:

$$\mathrm{p}(d|\Omega_n) := \frac{1}{K_n} \sum_{h=1}^{C} \mathrm{p}(\phi_h^1, \phi_h^2) \, \mathcal{N}(d; \mu_h^1 - \mu_h^2, \Sigma_h^1 + \Sigma_h^2), \tag{9}$$

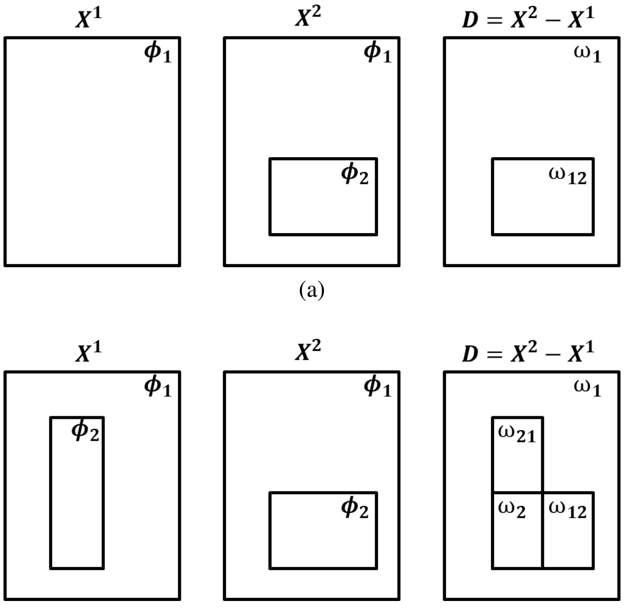
and the second term describes the statistical distribution of the multiple change class:

$$\mathrm{p}(d|\Omega_c) := \frac{1}{K_c} \sum_{\substack{h,k=1 \\ h \neq k}}^{C} \mathrm{p}(\phi_h^1, \phi_k^2) \, \mathcal{N}(d; \mu_h^1 - \mu_k^2, \Sigma_h^1 + \Sigma_k^2). \tag{10}$$

The constants K_n, K_c are defined in order to let probabilites $\mathrm{p}(d|\Omega_n)$ and $\mathrm{p}(d|\Omega_c)$ sum up to one, thus:

$$K_n := \sum_{h=1}^{C} \mathrm{p}(\phi_h^1, \phi_h^2), \qquad K_c := \sum_{\substack{h,k=1 \\ h \neq k}}^{C} \mathrm{p}(\phi_h^1, \phi_k^2). \tag{11}$$

Some important remarks are needed here. It is common practice in literature to model the change class as a mixture of different components [23]. However, the underlying assumption is always that the principal changes (the more statistically evident) are a small subset of all the ones that can be formulated theoretically. In particular this affects also the modeling of the unchange class, which is commonly reduced to be a single class. The increased radiometric resolution of last generation multispectral sensors typically results in more natural classes than the ones typically observed in older images. Thus, the mentioned approaches may not be appropriate anymore. A simple example about this distinction when $C = 2$, i.e., when images can be represented by two natural classes, is depicted in Figure 1. At the best of our knowledge, there is no attempt in literature to address the change detection problem by explicitly keeping into consideration the potential statistical variability of last generation multispectral data. In this work, we propose a general model that takes into account all the complexities arising by considering that both the unchange and change classes are multiple classes (it is worth noting that the proposed model also includes as



(a)

(b)

Fig. 1. Example of statistical dependency of the difference image with respect to the input pair when the number of natural classes is $C = 2$. (a) The natural class ϕ_2 is only observable in image $\mathbf{X}^2$, thus the resulting difference image has only one unchange behavior ω_{11} and one change behavior ω_{12} (from class ϕ_1 to ϕ_2). (b) Both natural classes ϕ_1, ϕ_2 are observable in the two images. Therefore, they have their own unchange statistical behaviors ω_1, ω_2, and mutual change behaviors ω_{12} (from class ϕ_1 to ϕ_2), and ω_{21} (from class ϕ_2 to ϕ_1).

particular cases some of the typical approaches to change detection already given in literature). With respect to the existent literature, the statistical model in (6) (and the discussion in this section), provide us additional quantitative information about:

1) the physical meaning of the classes that populate the difference image, and,
2) an explicit analytical description of their statistical distribution.

III. A method for binary change detection based on the magnitude of the difference image

In this section we consider a binary CD problem based on setting a threshold on the magnitude of the difference image. The proposed study starts from the model proposed in Section II and it firstly derives a general expression for the statistical distribution of the magnitude variable in the multi-class case. Then, a decision framework is defined following a multi-class maximum a-posteriori (MAP) approach, where the optimal value of the threshold depends on the parameters of the distribution. Lastly, with some simplifying assumptions w.r.t. general model, a fully unsupervised numerical method based on the EM algorithm [24] for parameter estimation of the derived mixture density is presented and the binary detection is performed based on the explicit calculation of the optimal value of the threshold.

A. The magnitude distribution in the multi-class case

In order to derive the statistical distribution of the magnitude of the difference image from (6), some simplifying assumptions are made. Firstly we assume similar radiometric conditions and no seasonal differences in the two images $\mathbf{X}^1, \mathbf{X}^2$. Therefore we expect the parameters of the same class to not vary between the two acquisitions, in such a way that

$$\begin{aligned} \mu_h^1 = \mu_h^2 &=: \mu_h \\ \Sigma_h^1 = \Sigma_h^2 &=: \Sigma_h \end{aligned} \tag{12}$$

for all $h = 1, \ldots, C$. Secondly, since our purpose is to exploit the magnitude information in the decision process, some regularity assumptions on the structure of the covariance matrices Σ_h are needed in order to derive the theoretical distribution of the magnitude as a Rice-like mixture type [25]. In practice, we assume that covariances are scalar matrices of the type

$$\Sigma_h = \sigma_h^2 \mathrm{I} \tag{13}$$

where I is the identity matrix of size $B \times B$ and σ_h are positive scalars representing the standard deviation of each class. It is worth noting that, it is common practice in the CD literature to assume that a change has small dimension in the feature space. Often, the feature space is reduced in such a way to obtain a lower dimensional space where the change is more evident. The analysis presented in the following holds true whenever the hypothesis of Gaussian distribution of classes in the images is matched. We retain Gaussian distribution of classes for a large class of transformations which is linear transformations. Particular cases are band selection and more in general linear projections, which are very popular in literature. Therefore, many of the methods for feature reduction used in the literature match with our theoretical setting. For simplicity, in the sequel we assume the feature space is 2-D (i.e., $B = 2$) but all the arguments presented can be easily generalized to any dimension. A few bands (typically the most representatives of the changes of interest) usually correspond to uncorrelated spectral ranges. Accordingly, we see that the assumption of diagonality of the matrices Σ_h is not very strict. The assumption of having the same variance in all dimensions is purely technical. Without this constraint, the analytical expression of the magnitude distribution becomes intractable from the computational viewpoint [19], [26]. To conclude our theoretical setting, following from the discussion in Section II we recall that spatial independence of pixels is assumed here. Therefore, each pixel value in the difference image is considered to be drawn from the distribution (6) independently from the others.

By taking into account (12), the difference image distribution $\mathrm{p}(d)$ can be written in the simpler form

$$\mathrm{p}(d) = \sum_{h,k=1}^{C} q_{hk} \, \mathcal{N}(d; \mu_h - \mu_k, \Sigma_h + \Sigma_k). \tag{14}$$

where the mixture terms q_{hk} are the entries of matrix Q given in (7), i.e., $q_{hk} := \mathrm{p}(\phi_h^1, \phi_k^2)$. Notice that each class component of the multiple unchange class is 0-mean and its covariance matrix is proportional (by factor 2) to the covariance matrix of the associated (unchanged) natural class

$$\mathrm{p}(d|\Omega_n) = \frac{1}{K_n} \sum_{h=1}^{C} q_{hh} \, \mathcal{N}(d; 0, 2\Sigma_h). \tag{15}$$

Regarding the class components of the multiple change class, they are mutually symmetric with respect to the origin

$$\mathrm{p}(d|\Omega_c) = \frac{1}{K_c} \sum_{\substack{h,k=1 \\ h \neq k}}^{C} q_{hk} \, \mathcal{N}(d; \mu_h - \mu_k, \Sigma_h + \Sigma_k). \tag{16}$$

Indeed, we have that $\mu_h - \mu_k = -(\mu_k - \mu_h)$ and, obviously, $\Sigma_h + \Sigma_k = \Sigma_k + \Sigma_h$.

These considerations enable us to describe the distribution $\mathrm{p}(d)$ when coordinates are changed to magnitude. Indeed, given the previous assumptions, the theoretical derivation of the distribution of the random variable $R := |D| = |X^1 - X^2|$ is possible (the operator $|.|$ denotes the Euclidean norm). It follows that the magnitude image $\mathbf{R} := \{|\mathbf{D}(i,j)| : (i,j)\}$ can be considered as a sample drawn from this distribution. Let us recall the general form of the Rician distribution that is given by

$$\mathcal{R}(\rho; \nu, \delta) = \frac{\rho}{\delta^2} \exp\left(-\frac{\rho^2 + \nu^2}{2\delta^2}\right) \mathrm{I}_0\left(\frac{\rho\nu}{\delta^2}\right) \qquad \rho \geq 0 \tag{17}$$

where $\mathrm{I}_0(.)$ is the 0-th order modified Bessel function of first kind [27]. By defining the magnitude variable as $\rho := |d|$, we have that the distribution of ρ follows

$$\mathrm{p}(\rho) = \sum_{h,k=1}^{C} q_{hk} \, \mathcal{R}(\rho; \nu_{hk}, \delta_{hk}) \tag{18}$$

where $\nu_{hk} = |\mu_h - \mu_k|$ and $\delta_{hk} = \sqrt{\sigma_h^2 + \sigma_k^2}$. Let us point out some important remarks here. Since $\nu_{hh} = 0$ for every $h = 1, \dots, C$, we easily have that any component related to the unchange class is actually Rayleigh. Moreover, the scale parameters of these Rayleigh distributions only depend on the variances of the associated unchanged classes. More specifically, $\delta_{hh} = \sigma_h \sqrt{2}$ for each h. The unchange class in the magnitude space is therefore described by

$$\mathrm{p}(\rho|\Omega_n) = \frac{1}{K_n} \sum_{h=1}^{C} q_{hh}\, \mathcal{R}(\rho; 0, \sigma_h \sqrt{2}). \tag{19}$$

All the other cross terms in (18) are mutually equal as $\nu_{hk} = \nu_{kh}$ and $\delta_{hk} = \delta_{kh}$ for each couple (h, k) such that $h \neq k$. Moreover, each non-centrality term is in general $\nu_{hk} \neq 0$, thus all components related to the change class are Rician. It follows that all the components of the multiple change class group together two-by-two and their mixture parameters are summed up

$$\mathrm{p}(\rho|\Omega_c) = \frac{1}{K_c} \sum_{\substack{h,k=1 \\ h<k}}^{C} (q_{hk} + q_{kh})\, \mathcal{R}(\rho; \nu_{hk}, \delta_{hk}). \tag{20}$$

This means that, in the magnitude space the change from class ϕ_h to ϕ_k (class ω_{hk}) and the opposite change (class ω_{kh}) are statistically equivalent. Therefore, the number (and the meaning) of change classes in the magnitude space is different from that in the difference space. In light of this, we change the notation for the unchange and change classes in the magnitude space by adding a superscript ρ, namely Ω_n^ρ and Ω_c^ρ.

Once the structure of the matrix $\mathbf{Q}$ is known, i.e., when the indexes of its non-null entries (and not necessarily their values) are known, the number of unchange and change class members in (18) can be easily obtained. Let us denote $C' := \#(\Omega_n^\rho)$ the number of members of the unchange class and $C'' := \#(\Omega_c^\rho)$ the number of members of the change class. On the one side, only classes that are observable in both images can have their unchange behavior, therefore we have in general

$$0 \leq C' \leq C. \tag{21}$$

On the other side, each natural class can have at most $C - 1$ different change behaviors. Thus, in view of the above mentioned statistical equivalence of change classes ω_{hk} and ω_{kh}, we have that

$$0 \leq C'' \leq \frac{C(C-1)}{2}. \tag{22}$$

For ease of notation, in the following the mixture indexes are re-ordered and mixture and shape parameters are re-named accordingly. We now write the distribution of the magnitude as

$$\mathrm{p}(\rho) = \sum_{h=1}^{H} \alpha_h\, \mathcal{R}(\rho; \nu_h, \delta_h) \tag{23}$$

where $H = C' + C''$ and, of course, $\sum_{h=1}^{H} \alpha_h = 1$. In the same manner we denote class labels as $\omega_h \in \Omega_n^\rho$ for $h = 1, \dots, C'$, and $\omega_h \in \Omega_c^\rho$ for $h = C' + 1, \dots, H$.

B. A Bayesian multi-class approach to binary decision

The theoretical model (23) describes the distribution of the magnitude of the difference image in the multi-class case and enables us to define a Bayesian framework for binary decision based on the maximum a-posteriori principle (MAP) and the thresholding of the magnitude variable. By following the MAP approach, for any given value of the target variable, one selects as the most likely originating class the one that maximizes the posterior probability of observing that value among all classes. It is well-known that this procedure minimizes the overall probability of committing classification errors. Therefore, given the mixture model (23) we define the MAP classification function

$$W[\rho] := \arg \max_{\omega_h \in \Omega^\rho} \; \alpha_h \; \mathrm{p}(\rho|\omega_h) \tag{24}$$

where $\Omega^\rho := \Omega_n^\rho \cup \Omega_c^\rho = \{\omega_1, \dots, \omega_H\}$ is the set of all classes. It is worth noting that, being each competing class either $\omega_h \in \Omega_n^\rho$ or $\omega_h \in \Omega_c^\rho$, this approach has an intrinsic binary interpretation. Indeed, each value $W[\rho]$ identifies (and associates to ρ) one member of the two multiple unchange or change classes. Thus, the binary classification based on the thresholding of the magnitude of the difference image can be defined by

$$\begin{aligned} W_n &= \{(i,j) : W[\mathbf{R}(i,j)] \in \Omega_n^\rho\} \\ W_c &= \{(i,j) : W[\mathbf{R}(i,j)] \in \Omega_c^\rho\}, \end{aligned} \tag{25}$$

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

The Bayes decision method presented here generalizes the two-class approach for binary change detection presented in [20], as this last one can be derived from (24) and (25) when two natural classes are considered (i.e., $C = 2$), with the additional assumption that one class is observable only in one of the two input images (i.e., $C' = C'' = 1$). An example of this situation is illustrated in Figure 1a, whereas in Figure 1b a more general case involving all possible class unchange and change behaviors is illustrated.

C. Basic assumptions for the binary change detection problem

The binary change detection problem is usually addressed in literature by assuming the unchange class to be a single class [13], [17]–[20]. Sometimes, such assumption is extended also to the change class [19], [20], while in other cases, the change class is modeled as multiple [4]. By a careful analysis of matrix $\mathbf{Q}$ and its relationship with the mixture components in (23), we see that:

- A model that assumes both the unchange and change classes to be single, is represented (in the simplest case) by a matrix of size 2×2. More specifically, the possible cases are:

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} q_{11} & 0 \\ q_{21} & 0 \end{pmatrix}, \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & 0 \end{pmatrix},$$

  where we recall that the changes ω_{12} and ω_{21} are statistically equivalent in the magnitude space, thus the last case relates to a single change.
- The assumption of having more than one change in the magnitude space puts a constraint on the number of natural classes in the model, which must be $C \geq 3$. The description of all possible cases becomes more and more complex when C increases.

As we can see, the distribution model can be arbitrarily complex if we do not have any prior knowledge about the classes involved. Some information about the structure of the matrix $\mathbf{Q}$ can be recovered in a supervised or semi-supervised way, by separately analyzing the single time images $\mathbf{X}^1, \mathbf{X}^2$ and their difference $\mathbf{D}$. However, this kind of approach may result too articulated for the specific purpose of binary change detection, whereas it seems more appropriate to be developed in the context of multiple change detection.

In formulating the a priori assumptions for the development of our method, we mainly consider that: 1) The magnitude variable has high informative content about the presence or absence of the change, not really about the type of change because of the compression from B-dimensional to 1-D space. 2) Simple models proved to be effective for addressing the specific binary change detection problem. 3) The solution of the EM algorithm is sub-optimal (local max), thus the introduction of several additional parameters will decrease the significance of the solution itself. In view of these points, let us consider the case where only one relevant change can be identified in the image pair, so that we can focus the attention on the multiple unchange class. Therefore, the matrix $\mathbf{Q}$ can be assumed of the form

$$\mathbf{Q} = \begin{pmatrix} q_{11} & & & & \\ & \ddots & & q_{hk} & \\ & & \ddots & & \\ & q_{kh} & & \ddots & \\ & & & & q_{CC} \end{pmatrix}, \tag{26}$$

where $h, k \in \{1, \ldots, C\}$ are fixed. We stress the fact that for the derivation of the statistical model in (23), some regularity assumptions on the natural classes are made, see (12) and (13). Such assumptions allows us to the derive a precise model for the magnitude distribution and they have the remarkably positive effect of letting many subtle natural classes collapse into a few dominant ones. Indeed, according to (15), *all* natural classes that are described by the same covariance matrix can be grouped into one single unchange class in the difference space as their distributions coincide. More specifically, let us assume that for a fixed $p \in \{1, \ldots, C\}$ there is a set of classes $\mathcal{B}_p \subset \{1, \ldots, C\}$ such that $\Sigma_h = \Sigma_p$, for all $h \in \mathcal{B}_p$. Then,

$$\sum_{h \in \mathcal{B}_p} q_{hh} \mathcal{N}(d; 0, 2\Sigma_h) = \mathcal{N}(d; 0, 2\Sigma_p) \sum_{h \in \mathcal{B}_p} q_{hh}. \tag{27}$$

This simple mathematical fact proves that the complexity of the unchange class (in terms of number of distinguishable class members) does not depend on the real number of natural classes, but only on those that have different covariance matrices. Notice in particular that, under the considered assumptions, the mean vector of each class (i.e., the feature vector that mainly characterizes the spectral signature of each class) does not play any role in the definition of the unchange class members.

Accordingly, we devise a distribution model that is able to represent the unchange multiple class in terms of dominant classes, while it does not increase too much the overall complexity of the associated estimation problem. Without putting a constraint on the *actual* number C of natural classes that populate the images, we just assume that two main groups $\mathcal{B}_p$ and $\mathcal{B}_q$, with $p, q \in \{1, \ldots, C\}$, can be formed out of all classes, which are homogeneous in terms of covariance matrices. Thus, we can write

$$\mathrm{p}(\rho|\Omega_n) = \mathcal{N}(d; 0, 2\Sigma_p) \sum_{h \in \mathcal{B}_p} \frac{q_{hh}}{K_n} + \mathcal{N}(d; 0, 2\Sigma_q) \sum_{h \in \mathcal{B}_q} \frac{q_{hh}}{K_n}. \tag{28}$$

In principle, this simplification could be extended to any number of groups. Nonetheless, we will show in the experimental part of this work how the simple choice of assuming two dominant classes proved to be sufficiently general to provide a change detection performance which is nearly optimal.

D. An unsupervised method for binary change detection based on the EM algorithm

The magnitude of the difference image $\mathbf{R}$ is considered to be a set of samples independently drawn from distribution (23). In order to apply the MAP approach for decision, parameters of this distribution must be numerically estimated. In the theoretical model presented in Section III-A, the parameters are constrained by their dependency on the natural class parametrization. In order to partially recover some flexibility that we lost in imposing conditions such as in (12) and (13), we implement a version of the EM algorithm without parameter constraints. In addition to greater flexibility, this also gives the advantage of implementing an unconstrained optimization. Given this, from the a priori model (26) coupled with (28) and applied to the distribution (23), we can write the probability distribution of the magnitude of the difference image as

$$\mathrm{p}(\rho) = \alpha_1 \mathcal{R}(\rho; 0, \delta_1) + \alpha_2 \mathcal{R}(\rho; 0, \delta_2) + \alpha_3 \mathcal{R}(\rho; \nu, \delta) \tag{29}$$

where $\alpha_3 = 1 - \alpha_1 - \alpha_2$. For notation simplicity we omitted the subscripts of the parameters of the Rician component (the third term in the summation). Let $\Psi = (\alpha_1, \alpha_2, \delta_1, \delta_2, \nu, \delta)$ be the vector of shape and mixture parameters involved in (29), then the dependency of this probability distribution with respect to the parameters can be expressed by writing $\mathrm{p}(\rho|\Psi)$. For any $r \in \mathbf{R}$ we define

$$\mathrm{p}(\omega_h|r, \Psi') := \frac{\alpha'_h \, \mathrm{p}(r|\omega_h, \Psi')}{\mathrm{p}(r|\Psi')} \tag{30}$$

to be the posterior probability that the sample r originated in the h-th component of the population, given Ψ'.

1) Iterative formulas: By the same arguments developed in [20], it is possible to define an iterative version of the EM algorithm for the estimation of the parameters of (29) as follows

$$\begin{aligned}
\alpha_i^{k+1} &= N^{-1} \sum_{r \in \mathbf{R}} \mathrm{p}(\omega_i | r, \Psi^k) \\
(\delta_i^2)^{k+1} &= \frac{\sum\limits_{r \in \mathbf{R}} \mathrm{p}(\omega_i | r, \Psi^k) r^2}{2 \sum\limits_{r \in \mathbf{R}} \mathrm{p}(\omega_i | r, \Psi^k)} \\
\nu^{k+1} &= \frac{\sum\limits_{r \in \mathbf{R}} \mathrm{p}(\omega_3 | r, \Psi^k) \frac{\mathrm{I}_1\left(\frac{r\nu^k}{(\delta^k)^2}\right)}{\mathrm{I}_0\left(\frac{r\nu^k}{(\delta^k)^2}\right)} r}{\sum\limits_{r \in \mathbf{R}} \mathrm{p}(\omega_3 | r, \Psi^k)} \\
(\delta^2)^{k+1} &= \frac{\sum\limits_{r \in \mathbf{R}} \mathrm{p}(\omega_3 | r, \Psi^k) \left[r^2 + (\nu^k)^2 - 2r\nu^k \frac{\mathrm{I}_1\left(\frac{r\nu^k}{(\delta^k)^2}\right)}{\mathrm{I}_0\left(\frac{r\nu^k}{(\delta^k)^2}\right)} \right]}{2 \sum\limits_{r \in \mathbf{R}} \mathrm{p}(\omega_3 | r, \Psi^k)}.
\end{aligned} \tag{31}$$

for $i = 1, 2$, where $\mathrm{I}_1(.)$ is the 1-th order modified Bessel function of first kind [27]. At each iteration k, the parameter updating can be optimized by computing only once the fractional term involving $\mathrm{I}_1(.)$ and $\mathrm{I}_0(.)$ as the formulas are implemented by linear algebra operations like elementwise vector products and divisions. Therefore, the algorithm complexity is $O(n)$, where n is the number of pixels.

2) Initialization: Iterative versions of the EM algorithm are known to be typically slow [24]. Moreover, the proposed algorithm iteratively updates the parameters of three different mixture components, for a total of two mixture and four shape parameters. This makes the initialization of the algorithm a very crucial point, as a bad strategy may easily result in a stagnation of the parameters updates. In particular, after many numerical experiments (both on synthetic and real data) we observed that the typical approach of median thresholding [13], [20] cannot be used for the preliminary Maximum Likelihood (ML) estimate of the parameters.

For a correct initialization, we found that the typical local minimum that separates the two main modes of the histogram of the magnitude must be precisely identified in order to have a first separation of the samples into two approximate classes: W_n^0 (unchanged samples) and W_c^0 (changed samples). Under the hypothesis that this local minimum is the stationary point of the underlying pdf with less curvature, it can be found as the point of lower variation of the corresponding cdf. A more stable numerical computation can be done by calculating the maximum point of the derivative of the quantile function. More specifically, let $Q : [0, 1] \to [0, 1]$ be the quantile function of the magnitude histogram and let $t^0 := \max_{t \in [0,1]} \partial Q(t)$. Then, the two approximate classes are defined as

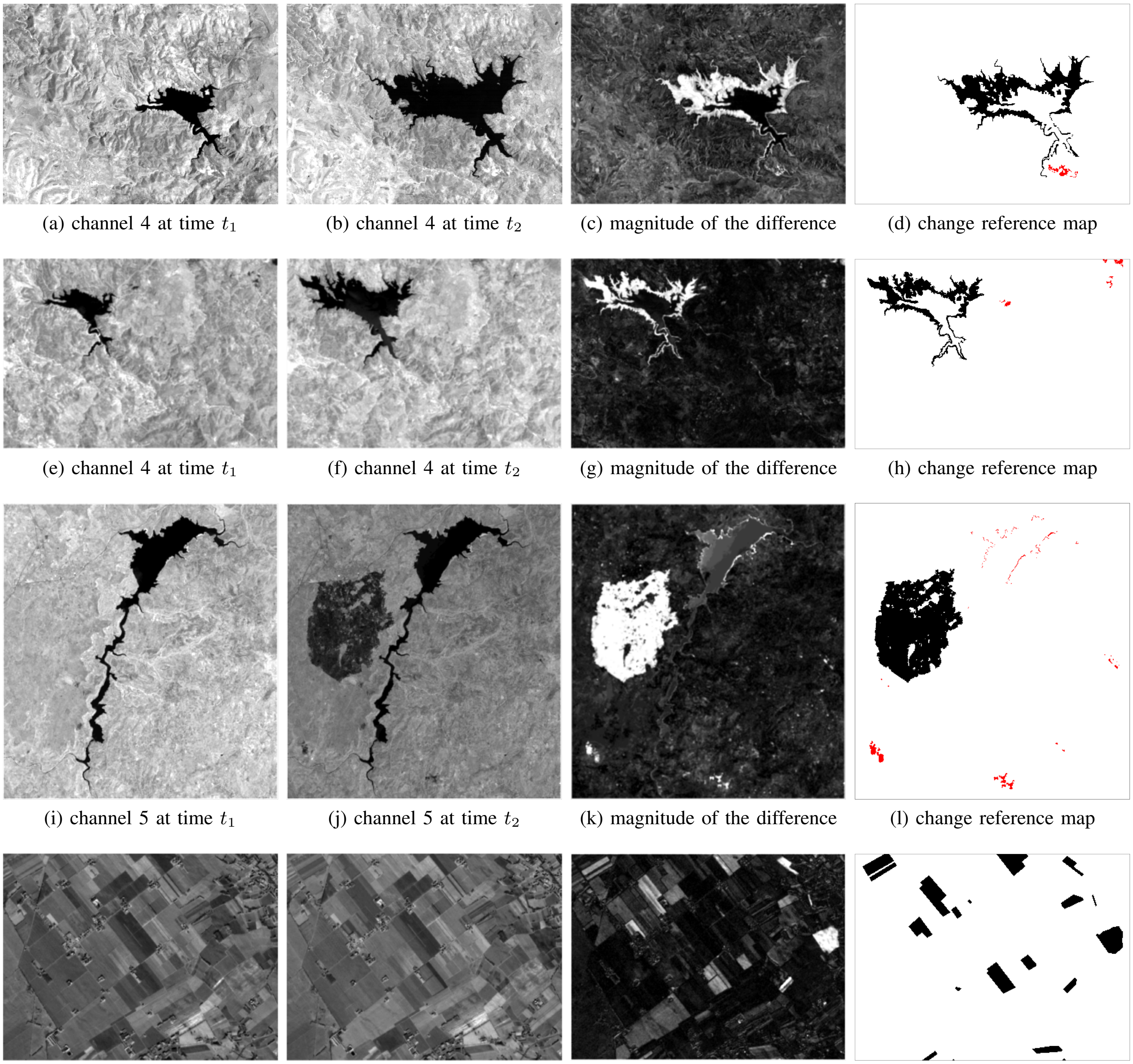
$$\begin{aligned}
W_n^0 &= \{(i, j) : \mathbf{R}(i, j) \leq t^0\} \\
W_c^0 &= \{(i, j) : \mathbf{R}(i, j) > t^0\},
\end{aligned} \tag{32}$$

Since in the target mixture the components that represent unchanged pixels are two, the samples in W_n^0 are further divided into two sets $W_{n,1}^0$, $W_{n,2}^0$ by median thresholding. Finally, shape and mixture parameters of the mixture are estimated via ML approach (refer to [28]) using the samples in the approximate classes. The shape parameters δ_i are estimated using the samples in $W_{n,i}^0$, with $i = 1, 2$, and ν, δ are estimated using the samples in W_c^0. The mixture parameters α_i are approximated by their corresponding proportions $\#(W_{n,i}^0)/\#(\mathbf{R})$, for $i = 1, 2$.

3) Convergence: The iterative algorithm defined by (31) is gradient-based, thus it can be stopped when slow variations of the objective energy (the ML function of the given set of magnitude samples) are observed. More specifically, the algorithm is stopped at the first iteration k such that the relative variation of the energy is less than 10^{-6} [20].

IV. Experimental results

In this section we present an extensive experimental analysis of the technique described in Section III for binary change detection based on the EM algorithm. Firstly, we describe the experimental setup, which is designed for comparing the proposed technique against the one recently proposed in [20] and other reference thresholding


(a) channel 4 at time t_1 (b) channel 4 at time t_2 (c) magnitude of the difference (d) change reference map

(e) channel 4 at time t_1 (f) channel 4 at time t_2 (g) magnitude of the difference (h) change reference map

(i) channel 5 at time t_1 (j) channel 5 at time t_2 (k) magnitude of the difference (l) change reference map

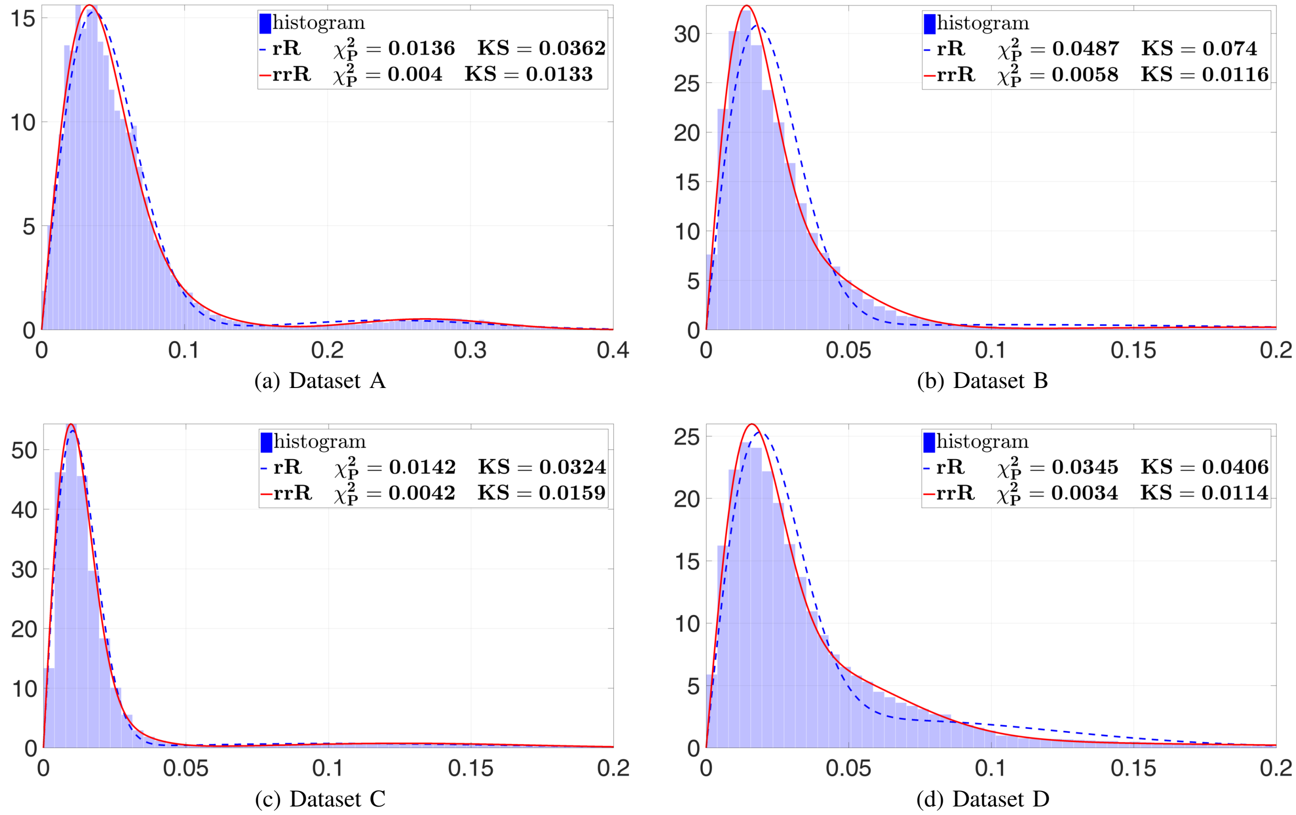
(m) channel 4 at time t_1 (n) channel 4 at time t_2 (o) magnitude of the difference (p) change reference map

Fig. 2. Illustration of the datasets analyzed in the experiments. The pictures show pre and post images in one band, the magnitude of the difference image (calculated using two bands) and a reference map of the changes (red pixels are minor changes with respect to the main changes that are represented by black pixels) for (a,b,c,d) dataset A, (e,f,g,h) dataset B, (i,j,k,l) dataset C, and (m,n,o,p) dataset D.

methods. Secondly, a brief description of the datasets analyzed in the tests (including image pairs from different multispectral sensors) is given. Then, the performance of the EM algorithm for parameter estimation is analyzed in detail, together with the fitting capability of the models. Lastly, the change detection is performed and the results are compared in order to assess the effectiveness of the proposed method with respect to the state-of-the-art techniques.

A. *Experimental setup*

The proposed algorithm is designed with the specific intent of improving the change detection performance of state-of-the-art statistical-based thresholding methods. It is shown in [20] that the distribution model based on a mixture of a Rayleigh and a Rician component, hereafter denoted **rR**, outperforms the Gaussian mixture model.



(a) Dataset A

(b) Dataset B

(c) Dataset C

(d) Dataset D

Fig. 3. Fitting capability of the considered methods. Estimated densities are superimposed to the histograms of the magnitude samples. In the legends also the distance metrics between the estimated densities and the histograms are given.

In the sequel, the proposed method, which is notated **rrR**, is tested against the **rR** and the well-known Kittler-Illingsworth (**KI**) method.

Let us present the outline of the experimental test. Firstly, the image pair is differenced and the magnitude image is calculated. Then, the magnitude values are used to estimate in an unsupervised way the parameters of both the **rR** and **rrR** statistical models via EM algorithm. Iteration details are given and the fitting performance of the estimated densities are analyzed. Lastly, the estimated parameters are used to perform Bayes decision according to the MAP approach. The binary change detection is performed by binarizing the magnitude image using the threshold value T that minimizes the overall error for all the **rR**, **rrR** and the **KI** cases. The change detection performance is evaluated in terms of missed/false alarms and overall errors (denoted MA, FA and OE, respectively). Results of fitting and change detection performance are also analyzed w.r.t. to the radiometric resolution of the images.

B. Datasets description

The datasets considered in this work consist of bitemporal image pairs from different multispectral sensors. The images are co-registered and radiometrically corrected. Images are followed by reference maps of the changes obtained by photo-interpretation. The datasets are illustrated in Figure 2 and briefly described in the following.

1) Dataset A: Images are acquired by the Thematic Mapper (TM) multispectral sensor of the Landsat-5 satellite. The scene represents an area including Lake Mulargia (Sardinia Island, Italy), at spatial resolution 30 m and radiometric resolution 8 bit. The images consist of 300×412 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. Changed pixels are 7694. The two bands that better represents the change of interest are TM4 and TM7 (channels 4 and 7), the near infrared (NIR) and the middle infrared (MIR) [19], respectively.

2) Dataset B: Images are acquired by the Enhanced Thematic Mapper Plus (ETM+) multispectral sensor of the Landsat-7 satellite. The scene represents the same subject of Dataset A (Lake Mulargia, Sardinia). Spatial resolution is 30 m and radiometric resolution is 8 bit. The images consist of 310×449 pixels, changed pixels are 5959. The dates of acquisition are May 2001 (t_1) and May 2003 (t_2). As for dataset A, bands ETM+4 (NIR) and ETM+7 (MIR) are selected as most representative of the change (lake enlargement).

3) Dataset C: Images are acquired by the Operational Land Imager (OLI) multispectral sensor of the Landsat-8 satellite. The investigated area includes Lake Omodeo and a portion of Tirso River (Sardinia Island, Italy). The images consist of 700×650 pixels at spatial resolution 30 m and radiometric resolution 12 bit. The dates of acquisition are 25th July 2013 (t_1) and 10th August 2013 (t_2). The main change is a fire occurred between August 7th and 9th in the south of Ghilarza village, spatial extension of the change is 36409 pixels. 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) [20], respectively.

4) Dataset D: Images are acquired by the Multispectral Instrument (MSI) of the Sentinel-2a satellite. The images consist of 300×400 pixels at spatial resolution 10 m and radiometric resolution 8 bit. The dates of acquisition are 31st December 2015 (t_1) and 20th January 2016 (t_2). The investigated area includes the surrounding of Tesis village (Pordenone, Italy). The main changes are all related to variations in soil conditions due to human activities on crops. Changed pixels are 7132. The two bands selected as most representative of the changes are bands 3 and 4, the red and the NIR channels, respectively.

C. EM algorithm and fitting performance

In the run of the iterative algorithm for parameter estimation of the **rR** mixture, the classical median thresholding technique for the initialization worked well. However, it resulted in never-ending iterations (the relative variation of the objective energy has never reached values below tolerance) in the **rrR** case, for all the three datasets. Therefore, the more appropriate initialization strategy based on the quantile function (described in Section III-D2) has been used. The initial approximate threshold values obtained in the calculations are $t^0 = 0.1477, 0.1091, 0.0559, 0.0100$ for Dataset A,B,C and D, respectively. These values are placed in the central portion of the histograms (cfr. with histograms in Figure 3) between the two principal modes. Notice that this does not hold true in the case of the median thresholding, indeed the corresponding values obtained in the computations are $t^{mid} = 0.2105, 0.1833, 0.2511, 0.0216$. As we can see, such values are significantly shifted on the right in the corresponding histograms. This proved to be a limitation for the initial estimation of the two Rayleigh components in the **rrR** mixture as the unchange class is sensibly over-estimated. Iteration details and the initial and final values of the estimated parameters are shown in Table I. As expected, the number of iterations and the time of computation increases in the case of **rrR**, as there are two additional parameters to be estimated in the model. Notice that, between the first an the last iteration, the parameters that are more significantly changed are the prior probabilities of classes, i.e., α in the **rR** case and α_1, α_2 in the **rrR** case.

Let us now study the fitting properties of the two methods. A qualitative analysis of the fitting performance can be done by looking at the plots of the estimated densities **rR** and **rrR**, which are superimposed to the histogram of the magnitude in Figure 3. The figure also provides quantitative evaluation of the fitting in terms of Chi square Pearson (χ^2_P) divergence and Kolmogorov-Smirnov (KS) distance between data and estimated densities. From the results we can see that the proposed **rrR** model fits the data in a satisfying way, and much better than **rR**. The fitting capabilities of the proposed statistical model increase with the radiometrical resolution of the images. This is demonstrated in an additional fitting test where the radiometric resolution of datasets C and D (from L8 and S2, respectively) is reduced from 12 bit to 8 bit by quantization. Table II reports the fitting quality in terms of χ^2_P divergence obtained for original images and the reduced ones. We see that when resolution changes from 8 bit to 12 bit, in both datasets the fitting of the **rrR** model improves much better than the **rR** model. In case of Dataset C, while **rR** improves by factor 2, the **rrR** model improves by factor 3. For Dataset D, the **rR** does not improve at all, while the **rrR** model improves by factor 2.

D. Change detection results

The whole theoretical framework presented in this work has been developed with the aim of better representing the typical distributions that characterize multispectral data acquired by new generation sensors. Following the expected (and confirmed) increased fitting performance of the resulting model, it was also expected an improvement in the change detection performance due to the increased capability of the model of representing the unchange multiple

TABLE I

EM algorithm iteration details and parameter estimations for both the **rR** and **rrR** mixtures.

mix	k	estimated parameters						secs
Dataset A								
		α	-	b	-	ν	σ	
rR	0	0.93	-	0.04	-	0.25	0.06	4.0
	60	0.92	-	0.04	-	0.23	0.07	
		α_1	α_2	δ_1	δ_2	ν	δ	
rrR	0	0.79	0.14	0.03	0.07	0.25	0.06	73.2
	1258	0.71	0.23	0.03	0.05	0.26	0.05	
Dataset B								
		α	-	b	-	ν	σ	
rR	0	0.96	-	0.02	-	0.20	0.06	3.6
	66	0.91	-	0.02	-	0.03	0.10	
		α_1	α_2	δ_1	δ_2	ν	δ	
rrR	0	0.90	0.06	0.02	0.05	0.20	0.06	13.7
	270	0.62	0.34	0.01	0.03	0.18	0.07	
Dataset C								
		α	-	b	-	ν	σ	
rR	0	0.92	-	0.01	-	0.12	0.04	10.9
	57	0.90	-	0.01	-	0.09	0.06	
		α_1	α_2	δ_1	δ_2	ν	δ	
rrR	0	0.87	0.05	0.01	0.03	0.12	0.04	65.7
	396	0.79	0.13	0.01	0.02	0.12	0.04	
Dataset D								
		α	-	b	-	ν	σ	
rR	0	0.98	-	0.03	-	0.21	0.06	2.15
	56	0.75	-	0.02	-	0.01	0.07	
		α_1	α_2	δ_1	δ_2	ν	δ	
rrR	0	0.90	0.08	0.02	0.08	0.21	0.06	6.01
	196	0.55	0.38	0.02	0.04	0.02	0.11	

TABLE II

Fitting performance of the statistical models **rR** and **rrR** at different radiometric resolutions. Fitting values are given in terms of χ^2_P divergence.

	Dataset C		Dataset D	
	rR	**rrR**	**rR**	**rrR**
8 bit	0.0214	0.0125	0.0375	0.0061
12 bit	0.0142	0.0042	0.0345	0.0034

class. This has been confirmed in the binary change detection phase of our tests. Table III reports the computed threshold value T and the change detection performance[2] in terms of missed/false alarms and overall errors (MA, FA and OE) for each dataset and each considered mixture model. The results can be compared with optimal performance based on a usual trial-and-error procedure applied to the reference maps of the changes.

As we can see, the proposed **rrR** mixture model for binary change detection proved to be sufficiently flexible

[2] Results of Table III can be converted to usual measurements "recall" and "precision" by the following transformations: $recall = 1 - MA/P$ and $precision = 1 - FA/(P - MA + FA)$ where P is the number of changed pixels.

TABLE III

Comparison of the change detection performance of the **KI**, **rR**, **rrR** methods with respect to optimal performance.

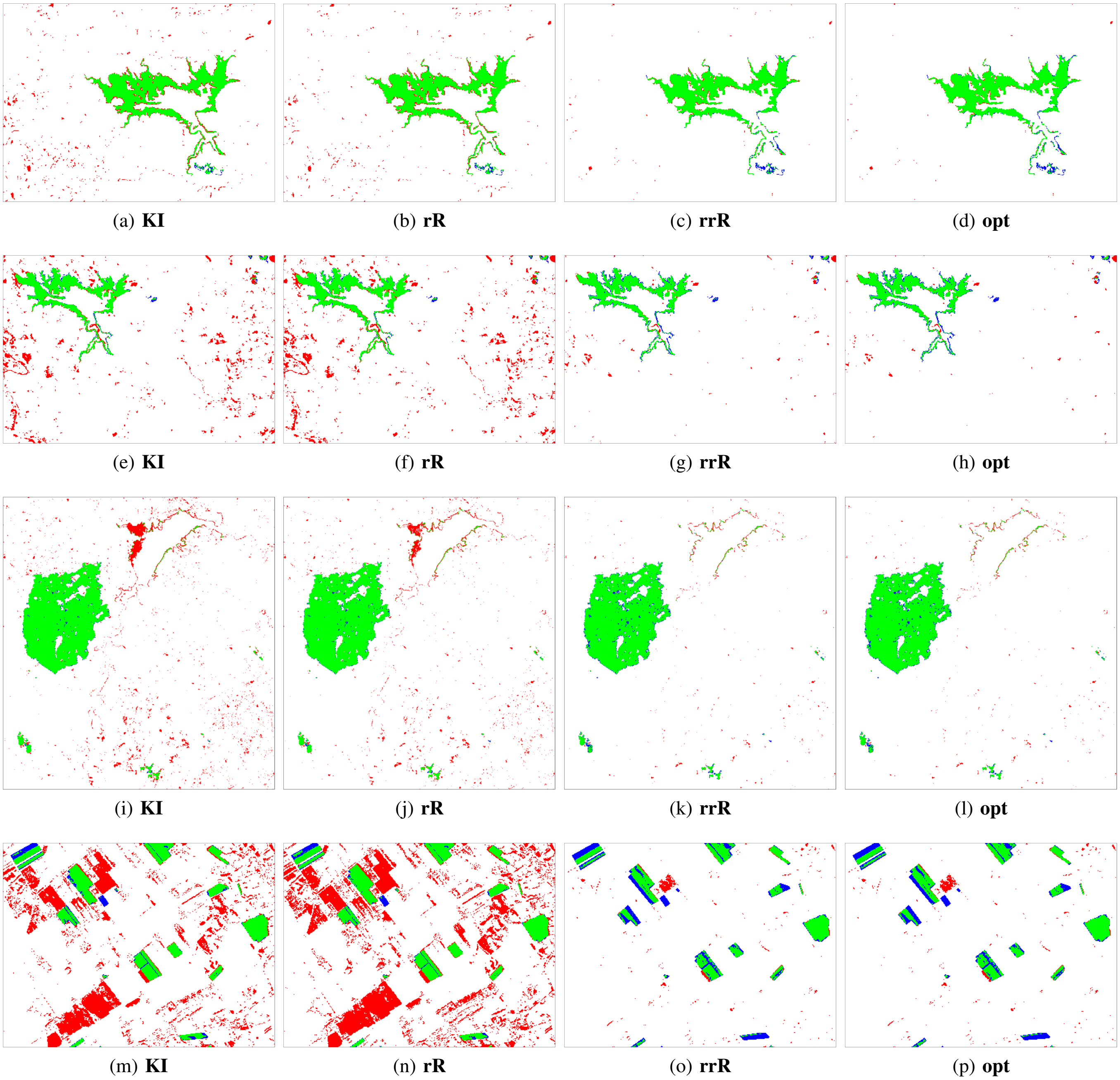
mix	T	MA (%)	FA (%)	OE (%)
	Dataset A (changed pixels 7694)			
KI	0.1237	112 (1.46)	2289 (1.97)	2401 (1.94)
rR	0.1332	139 (1.81)	1672 (1.44)	1811 (1.47)
rrR	0.1741	402 (5.22)	498 (0.43)	900 (0.73)
opt	0.1825	491 (6.38)	392 (0.34)	883 (0.71)
	Dataset B (changed pixels 5959)			
KI	0.0642	474 (7.95)	4946 (3.71)	5420 (3.89)
rR	0.0624	451 (7.57)	5421 (4.07)	5872 (4.22)
rrR	0.1028	975 (16.36)	801 (0.60)	1776 (1.28)
opt	0.1119	1115 (18.71)	567 (0.43)	1682 (1.21)
	Dataset C (changed pixels 36409)			
KI	0.0348	887 (2.44)	9593 (2.29)	10480 (2.30)
rR	0.0382	1064 (2.92)	6653 (1.59)	7717 (1.70)
rrR	0.0544	1992 (5.47)	1839 (0.44)	3831 (0.84)
opt	0.0553	2057 (5.65)	1746 (0.42)	3803 (0.84)
	Dataset D (changed pixels 7132)			
KI	0.0617	1050 (14.72)	13690 (12.13)	14740 (12.28)
rR	0.0527	837 (11.74)	19006 (16.84)	19843 (16.54)
rrR	0.1135	2557 (35.85)	1080 (0.96)	3637 (3.03)
opt	0.1146	2597 (36.41)	1025 (0.91)	3622 (3.02)

to well model the unchange and change classes in the datasets as it is able to return the same thresholds (with precision to the second decimal digit) that are obtained in the optimal (**opt**) case. A qualitative understanding of the performance is possible by looking at the change detection maps illustrated in Figure 4. With respect to the **rR** case, the proposed model presents much less false alarms at the expense of very few additional missed alarms. The performance of the **KI** method is very poor, confirming that a Rician-mixture based model is the more suited to fit the magnitude of the difference image in MS image analysis. The change detection results can be analyzed also w.r.t. the radiometric resolution. Indeed, images composing the dataset are from L5, L7, L8 and S2, thus with increasing radiometric resolution. We extract from Table III an interesting empirical observation: the average ratio between the OE of **rR** and the OE of **rrR** is higher in the group of 12 bit images (Datasets C and D) than in the group of 8 bit images (Datasets A and B). In fact we have: $0.5(1.70/0.84 + 16.54/3.03) \approx 3.75$ and $0.5(1.47/0.73 + 4.22/1.28) \approx 2.65$, respectively. Therefore, when radiometric resolution increases from 8 bit to 12 bit the proposed **rrR** method significantly improves the change detection performance with respect to **rR**.

V. Discussion and conclusion

The new generation of multispectral sensors mounted on satellite missions such as Landsat-8 and Sentinel-2 offers a unique opportunity of studying and monitoring the Earth surface by providing images at very short revisit time. Moreover, the spectral signature of the observed objects can be recorded at higher radiometric resolutions. This augments the statistical variability of multispectral data and typical approaches to change detection presented in literature that worked well on data acquired by older satellite missions are no more able to address the new challenges arising in this important field of remote sensing.

In order to fill this gap, in the first part of this work we presented a theoretical study of the change detection problem in a full multiclass framework. In particular, a statistical model of the difference image is derived starting from a few basic assumptions that are usually made in multispectral image analysis. The aim of the proposed study is that of defining a model with sufficient degrees of freedom for well representing the intrinsic multiclass nature of multispectral data also in the CD problem. It is worth noting that, in the proposed model no a-priori assumptions are made directly on the difference image. Instead, it explicitly describes the dependency between the



(a) **KI** (b) **rR** (c) **rrR** (d) **opt**

(e) **KI** (f) **rR** (g) **rrR** (h) **opt**

(i) **KI** (j) **rR** (k) **rrR** (l) **opt**

(m) **KI** (n) **rR** (o) **rrR** (p) **opt**

Fig. 4. Change detection maps obtained on (a-d) dataset A, (e-h) dataset B, (i-l) dataset C, and (m-p) dataset D. Blue pixels are missed alarms, red ones are false alarms and green ones are correctly detected changes.

distributions of natural classes in the bitemporal image pair and the statistical model of the difference image. Being a generalization of the statistical description of CVA, our model can be used to extend already existent methods based on CVA to the multiclass case. Another important remark is related to the hypothesis we made that natural classes are distributed as Gaussians in single time images. We recall that our theoretical framework can be extended to other distributions well-known in the image processing domain, as long as the difference of mixtures involving such distributions can be described analytically.

In the second part of the work the multiclass model of the difference image has been exploited in the context of binary change detection based on setting a threshold on the magnitude of the difference image. Here, a statistical distribution of the magnitude variable is described in general terms and binary decision is defined by means of Bayes theory. In the experimental part of this work, the parameters of the associated distribution are estimated from data by using an iterative version of the EM algorithm specifically devised under the assumptions that only one change has occurred and the unchange class is multiple. Experimental results on an extended dataset demonstrated the effectiveness of the proposed method with a remarkable improvement with respect to state-of-the-art approaches. On the one hand, the fitting capability of the proposed model is increased, as expected. On the other hand, in all considered datasets the change detection performance is nearly optimal, as confirmed by a comparison with the performance that can be obtained by exploiting the reference change maps.

Considering a more general situation where several changes have to be identified is one of the main possible scenarios where to apply the general framework presented in Section II. Of course, this involves more parameters and the complexity of the estimation problem increases. On the other hand, a straightforward way to use the proposed method in CVA with multiple changes is to apply it to the magnitude variable after fixing specific ranges of values for the direction variable. Further developments of the proposed thoeretical model can be considered taking into account: 1) the exploitation of the spatial contextual information (which is not considered in this work) to grant spatial homogeneity of the detected change patches, by using Markov Random Fields (MRFs) [13], or Spatially Variant Finite Mixtures (SVFM) [29]; 2) integration of the proposed analysis of the difference image with the compound-classification approach to CD [8], where the additional information about the physical meaning of the classes that originate after image differencing can be used to enforce the capability of the classifier to model temporal correlation in the data.

Appendix A
Notation

In this appendix we briefly recall the notation used throughout the paper and list the main symbols in Table IV. In the technical sections of this paper symbols may have superscripts $t = 1, 2$, where t refers to the time variable, meaning that for different values of t the symbols represent different variables. They also may have subscripts h, k (and others) that relate the specific symbols to particular classes h, k (and others).

TABLE IV
List of symbols used in this paper.

symbol	description
$\mathbf{X}$, $\mathbf{D}$, $\mathbf{R}$	multispectral and scalar images spatially referenced by spatial coordinates (i, j)
X, D, R	random variables associated to images and taking values x^t, d, ρ
Φ, Ω	random variables associated to class labels and taking values ϕ^t, ω
μ, Σ	parameters of multivariate Gaussian distributions (denoted by $\mathcal{N}$)
ν, δ	parameters of Rician distributions (denoted by $\mathcal{R}$)
p(.)	probability density function
$\mathbf{Q}$	matrix of prior probabilities $[\mathbf{Q}]_{h,k} = \mathrm{p}(\phi_h^1, \phi_k^2)$ where ϕ_h^t denotes the probabilistic event $\Phi^t = \phi_h$.
$W[\rho]$	MAP classification function
C	number of natural classes in single time images
α_i	mixture parameters (EM algorithm)
Ψ	parameter vector aggretaing variables $\alpha_1, \alpha_2, \delta_1, \delta_2, \nu, \delta$ (EM algorithm)

Appendix B
Statistical distribution of the difference of Gaussian mixtures

A derivation of the statistical distribution of the difference of two random variables $X^1, X^2 \in \mathbb{R}^B$ jointly distributed as

$$\mathrm{p}(x^1, x^2) = \sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2)\, \mathrm{p}(x^1|\phi_h^1)\, \mathrm{p}(x^2|\phi_k^2) \tag{33}$$

where $\mathrm{p}(x^1|\phi_h^1) = \mathcal{N}(x^1; \mu_h^1, \Sigma_h^1)$ and $\mathrm{p}(x^2|\phi_h^2) = \mathcal{N}(x^2; \mu_h^2, \Sigma_h^2)$ is possible by exploiting the characteristic function of linear combinations of random variables. For a random variable X admitting density function $\mathrm{p}(x)$, the characteristic function (CF) of X is the inverse Fourier transform of its density function:

$$\varphi_X(t) := E\left[e^{it^T X}\right] = \int_{\mathbb{R}^B} e^{it^T x}\, \mathrm{p}(x)\, dx. \tag{34}$$

The CF completely defines the probability distribution of the variable X. For a Gaussian distributed random variable X, its characteristic function can be described analytically in a closed form. Let us assume $X \sim \mathcal{N}(\mu, \Sigma)$, then we have that $\varphi_X(t) = \exp\{it^T\mu - \frac{1}{2}t^T\Sigma t\}$, where i is the imaginary constant: $i^2 = -1$.

Let us now consider the above mentioned random variables X^1, X^2, and let $V := a_1X^1 + a_2X^2$ be a linear combination of them with coefficients $a_1, a_2 \in \mathbb{R}$. By plugging (33) into (34) with $x = (x^1, x^2)$ and using the properties of the exponentials, the CF of V can be calculated as

$$\begin{aligned}
\varphi_V(t) &= \varphi_{a_1X^1+a_2X^2}(t) \\
&= E\left[\exp\left\{a_1t^TX^1\right\}\exp\left\{a_2t^TX^2\right\}\right] \\
&= \int\!\!\int \exp\left\{a_1t^Tx^1\right\}\exp\left\{a_2t^Tx^2\right\}\mathrm{p}(x)\,dx \\
&= \sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2) \int \exp\left\{a_1t^TX^1\right\}\mathrm{p}(x^1|\phi_h^1)\,dx^1 \times \\
&\qquad \int \exp\left\{a_2t^TX^2\right\}\mathrm{p}(x^2|\phi_k^2)\,dx^2 \\
&= \sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2)\exp\left\{ia_1t^T\mu_h^1 - \frac{1}{2}a_1^2t^T\Sigma_h^1t\right\}\times \\
&\qquad \exp\left\{ia_2t^T\mu_k^2 - \frac{1}{2}a_2^2t^T\Sigma_k^2t\right\} \\
&= \sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2)\exp\left\{it^T(a_1\mu_h^1 + a_2\mu_k^2) + \right. \\
&\qquad \left. -\frac{1}{2}t^T(a_1^2\Sigma_h^1 + a_2^2\Sigma_k^2)t\right\}.
\end{aligned} \tag{35}$$

This analytical expression uniquely identifies the distribution of V as a mixture of Gaussians

$$\mathrm{p}(v) = \sum_{h,k=1}^{C} \mathrm{p}(\phi_h^1, \phi_k^2)\,\mathcal{N}(v; a_1\mu_h^1 - a_2\mu_k^2, a_1^2\Sigma_h^1 + a_2^2\Sigma_k^2) \tag{36}$$

having C^2 mixture components. The distribution of the difference $X^1 - X^2$ can be easily obtained from (36) with $a_1 = 1$ and $a_2 = -1$.

Acknowledgments

The authors are thankful to anonymous reviewers for their suggestions for improvement of this paper.

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] 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.

[3] F. Bovolo and L. Bruzzone, "The time variable in data fusion: A change detection perspective," *Geoscience and Remote Sensing Magazine, IEEE*, vol. 3, no. 3, pp. 8–26, 2015.

[4] 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.

[5] A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 463–478, 2007.

[6] A. Singh, "Digital change detection techniques using remotely-sensed data," *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[7] L. Bruzzone and S. B. Serpico, "An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 35, no. 4, pp. 858–867, 1997.

[8] 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.

[9] ——, "Detection of land-cover transitions in multitemporal remote sensing images with active-learning-based compound classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 5, pp. 1930–1941, 2012.

[10] ——, "Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 51, no. 1, pp. 300–312, 2013.

[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] 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.

[13] 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.

[14] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J.-Y. Tourneret, "A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 799–812, 2015.

[15] T. Aach and A. Kaup, "Bayesian algorithms for adaptive change detection in image sequences using markov random fields," *Signal Processing: Image Communication*, vol. 7, no. 2, pp. 147–160, 1995.

[16] S.-C. Liu, C.-W. Fu, and S. Chang, "Statistical change detection with moments under time-varying illumination," *IEEE Transactions on Image Processing*, vol. 7, no. 9, pp. 1258–1268, 1998.

[17] Z. Yetgin, "Unsupervised change detection of satellite images using local gradual descent," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 5, pp. 1919–1929, 2012.

[18] 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.

[19] 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.

[20] M. Zanetti, F. Bovolo, and L. Bruzzone, "Rayleigh-rice mixture parameter estimation via em algorithm for change detection in multispectral images," *Image Processing, IEEE Transactions on*, vol. 24, no. 12, pp. 5004–5016, 2015.

[21] C. H. Chen, *Image Processing for Remote Sensing*, 1st ed. Boca Raton, FL: CRC Press, 2007.

[22] N. Acito, M. Diani, G. Corsini, and S. Resta, "Introductory view of anomalous change detection in hyperspectral images within a theoretical gaussian framework," *IEEE Aerospace and Electronic Systems Magazine*, vol. 32, no. 7, pp. 2–27, 2017.

[23] S. Liu, L. Bruzzone, F. Bovolo, M. Zanetti, and P. Du, "Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 53, no. 8, pp. 4363–4378, 2015.

[24] 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.

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

[26] P. Beckmann, "Statistical distribution of the amplitude and phase of a multiply scattered field," *Journal of Research of the National Bureau of Standards, 66D*, vol. 3, pp. 231–240, 1962.

[27] G. N. Watson, *A treatise on the Theory of Bessel Functions*, 2nd ed. Cambridge University Press, 1966.

[28] 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.

[29] S. Sanjay-Gopal and T. J. Hebert, "Bayesian pixel classification using spatially variant finite mixtures and the generalized em algorithm," *IEEE Transactions on Image Processing*, vol. 7, no. 7, pp. 1014–1028, 1998.