2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Agglomerative Clustering of Defects in Ultrasonic Non-destructive Testing using Hierarchical Mixtures of Independent Component Analyzers

Addisson Salazar, Jorge Igual, Luis Vergara

Abstract— **This paper presents a novel procedure to classify materials with different defects, such as holes or cracks, from mixtures of independent component analyzers. The data correspond to the ultrasonic echo recorded after an impact by several sensors on the surface of the material. These signals are modelled by independent component analysis mixture models (ICAMM) for every kind of defect. After the ICAMM model is estimated for every defect, these are merged according to a distance measure that is obtained from the Kullback-Leibler divergence. The hierarchy obtained from the impact-echo data and the learning process allow different kinds of defective materials to be grouped consistently.**

I. Introduction

In the impact-echo technique, a material is excited by a hammer impact, which produces a response that is sensed by a mono or multi-sensor system that is located on the surface of the material. In this paper, we consider a multi-channel configuration with sensors located at different sides of a parallelepiped-shaped material. This configuration allows the microstructure material response to be measured from different planes in order to obtain a more complete examination of the underlying wave propagation phenomenon.

The impact-echo signals contain backscattering from grain microstructure as well as information about flaws in the inspected material [1]. The physical phenomenon of impact-echo corresponds to wave propagation in solids. When a disturbance is applied suddenly at a point on the surface of a solid, the disturbance propagates through the solid as three different types of waves: P-wave (normal stress), S-wave (shear stress), and R-wave (surface or Rayleigh) [2]. After a transient period in which the first waves arrive, wave propagation becomes stationary in resonant modes that vary depending on the defects inside the material.

Independent Component Analysis (ICA) decomposes the mixed observations in a linear transformation of statistically independent variables, estimating the mixing matrix and the independent sources up to a permutation, sign and amplitude indeterminacy. ICA not only decorrelates the sensor observations as principal component analysis PCA does, but it also reduces the higher-order statistical dependencies among them.

There are relatively few applications of ICA in the field of non-destructive testing (NDT); see, e.g., [3] or [4]. The main difficulties of the application of ICA to vibration signals were analysed in [5]. They include: scaling and labelling indeterminacies of the sources; the dynamic nature of the mechanical systems, which requires a convolutive mixture of sources to be described; the physical relevance of the source meaning; determining the exact number of sources a priori; the problem of handling signals that are distributed in time and space; and the requirement of the system invertibility.

We present here a new application that consists on the agglomerative clustering of defective materials attending to their similarities. A general procedure to identify defects using an ICA mixture model ICAMM can be found in [6]. The ICAMM extends the linear ICA method by learning multiple ICA models and weighting them probabilistically [7][8], as many other mixture models do, e.g., the famous Gaussian Mixture Model that assumes each generator is a Gaussian density.

ICAMM allows independent components with data densities with nonlinearities and non-Gaussian distributions to be modelled. We will use a version that includes non-parametric density estimation, semi-supervision learning, use of any ICA algorithm in parameter updating, and correction of the posterior probability after training due to residual dependencies in the data [9].

Clustering techniques have been extensively studied in many different fields for a long time. They can be organized in different ways according to several theoretical criteria. However, a rough widely accepted classification of these techniques is: hierarchical and partitional clustering; see for instance [10][11]. Both clustering categories provide a division of the data objects. The hierarchical approach also yields a hierarchical structure from a sequence of partitions performed from singleton clusters to a cluster including all data objects (agglomerative or bottom-up strategy) or vice versa (divisive or top-down strategy). This structure consists of a binary tree (dendrogram) whose leaves are the data objects and whose internal nodes represent nested clusters of various sizes. The whole node of the dendrogram represents

This work has been supported by Spanish Administration under grant TEC 2005-01820.

Salazar, J. Igual, L. Vergara, and A. Serrano are with Universidad Politécnica de Valencia, Departamento de Comunicaciones, Camino de Vera s/n, 46022, Valencia, Spain. (e-mails: asalazar@dcom.upv.es, jigual@dcom.upv.es, lvergara@dcom.upv.es, arsercar@teleco.upv.es).

This work has been supported by Universitat Politècnica de València under grant SP20120646; Generalitat Valenciana under grant ISIC/2012/006; and Spanish Administration and European Union FEDER Programme under grant TEC2011-23403 01/01/2012

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

the whole data set. The internal nodes describe the extent that the objects are proximal to each other; and the height of the dendrogram usually represents the distance between each pair of objects or clusters, or an object and a cluster.

Fig. 1 represents an example of data clustering. Fig. 1a shows a partitional clustering and Fig. 1b and Fig. 1c show two representations of hierarchical clustering obtained from the clusters in Fig. 1a. Fig. 1b shows data merging order only inside the larger clusters, whereas the dendrogram of Fig. 1c shows the sequence of merging steps for all the clusters and the distance at which they were mixed.

A review of the clustering algorithms should include the following types of algorithms: hierarchical; squared error-based (vector quantization); mixture density-based; graph theory-based; combinatorial search technique-based; fuzzy; neural network-based; and kernel-based. In addition, some techniques have been developed to tackle sequential, large-scale, and high-dimensional data sets [12].

The advantages of hierarchical clustering include embedded flexibility regarding the level of granularity and the ability to deal with different types of attributes. The disadvantages of hierarchical clustering are the difficulty of scaling up to large data sets, the vagueness of stopping criteria, and the fact that most clustering algorithms cannot recover from poor choices when merging or splitting data points [13].

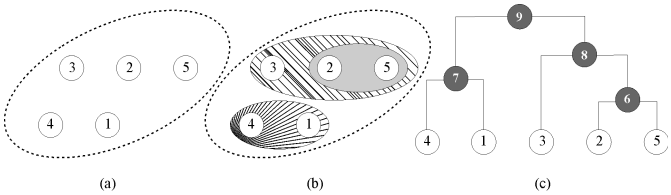

Fig. 1. Representations of partitional and hierarchical clustering: (a) partitional, (b) hierarchical, (c) dendrogram

A proximity or similarity measure is the basis for most clustering algorithms. This measure between clusters at one level in the hierarchy (also referred to as distance) is used to determine which of them will be merged. The distance between two clusters can be estimated between pairs of data objects of each of the clusters or between probabilistic relationships of the data densities of the two clusters. The following general recurrence formula for estimating a function distance $D(*,*)$ was proposed in [14]:

$$D\left(C_l,\left(C_i,C_j\right)\right)=\alpha_i D\left(C_l,C_i\right)+\alpha_j D\left(C_l,C_j\right)+ \\ +\beta_i D\left(C_i,C_j\right)+\gamma\left|D\left(C_l,C_i\right)-D\left(C_l,C_j\right)\right| \quad (1)$$

Equation (1) describes the distance between a cluster l and a new cluster formed by the merging of two clusters i and j. By manipulating the coefficients α_i,α_j,β, and γ, several hierarchical algorithms of clustering based on distances between data objects can be derived. Note that if $\alpha_i=\alpha_j=1/2$, $\beta=0$, and $\gamma=-1/2$, (1) is $D\left(C_l,\left(C_i,C_j\right)\right)=\min\left(D\left(C_l,C_i\right),D\left(C_l,C_j\right)\right)$, which corresponds to the single linkage method. In the case that $\alpha_i=\alpha_j=\gamma=1/2$ and $\beta=0$; (1) becomes $D\left(C_l,\left(C_i,C_j\right)\right)=\max\left(D\left(C_l,C_i\right),D\left(C_l,C_j\right)\right)$, which corresponds to the complete linkage method [11].

The probabilistic approaches to hierarchical clustering consider model-based criteria or Bayesian hypotheses to decide on merging clustering rather than using an ad-hoc distance metric. Basically, there are two approaches to derive the hierarchy: hierarchical generative modelling of the data or hierarchical ways of organizing nested clusters. Methods of the first approach include the following hierarchical generative models, for instance: Gaussian-based [15], diffusion-based [16], and mutation process-based [17]. The first two methods can be used for inference, and the last one can be used for semi-supervised learning. In [18], an agglomerative algorithm to merge of Gaussian mixtures is presented. It considers a virtual sample generated from the model at a level and uses EM (expectation maximization) to find the expressions for the mixture model parameters for the next level that best explain the virtual sample. Methods of the second approach include: agglomerative model merging, which is based on marginal likelihoods in the context of HMM [19]; a method to compute the marginal likelihood for c and $c-1$ clusters for use in an agglomerative algorithm [20]; clustering of multinomial feature vector data considering subsets of features having common distributions [21]; probabilistic abstraction hierarchies in which each node contains a probabilistic model with the most similar models as neighbouring nodes (estimated by a distance function) [22]; agglomerative algorithm for merging time series based on greedily maximizing marginal likelihood[23][24]; and using marginal likelihoods to decide which clusters to merge, when to stop, and when to avoid overfitting by testing a Bayesian hypothesis [25]. The model of this last method can be used to compute the predictive distribution of a test point and the probability of it belonging to any of the existing clusters in the tree.

Work on clustering that is related with hierarchies that are derived from independent component analysis (ICA) can be found in a hierarchical latent variable model for data visualization proposed in [28]. In this model, the form of the latent variable model is closely related to probabilistic principal component analysis (PPCA) [27][26]. The construction of the hierarchical tree proceeds top-down. At the top level of the hierarchy, a single visualization plot corresponding to a single model is defined. This model is partitioned into "clusters" at the second level of the hierarchy considering a probabilistic mixture of latent variable models. Subsequent levels, which are obtained using nested mixture representations, provide successively refined models of the data set [28].

ICA model-based hierarchies have also been explored. For instance, in [30], a method for capturing nonlinear dependencies in natural images for image segmentation and denoising is presented. It makes use of lower level linear ICA representation and a subsequent mixture of Laplacian distributions for learning the nonlinear dependencies.

We will use a hierarchical clustering of agglomerative type based on the assumption that distribution in each class of defect comes from an ICA model; i.e., the feature vector that defines our observations can be model as a linear combination of independent sources, where each defects has its own mixing matrix and sources. After the ICAMM model is estimated, a hierarchical procedure to merge them is used to identify similar defects.

This approach can be related to tree-dependent component analysis (TCA), which finds "clusters" of components such that the components are dependent within a cluster and independent between clusters [31]. Topographic independent component analysis (TICA) is another method that considers the residual dependence after ICA. This method defines a distance between two components using higher-order correlations, and it is used to create a topographic representation [29].

II. ICAMM STATEMENT OF THE PROBLEM

In impact-echo testing (see Fig. 2), the wave path propagation is divided into two parts: impact to point flaws (represented by *f* points in the figure) and point flaws to sensors. It is assumed that the set of point flaws builds defective areas with different geometries, such as cracks (small parallelepipeds), holes (cylinders), and multiple defects (combination of cracks and holes).

Depending on the kind of defective area, the spectrum measured by the sensors changes, which allows the kind of defect condition of the material to be discerned. It is demonstrated in [6] that the spectrum of different kinds of defective materials can fit into different ICA mixture models.

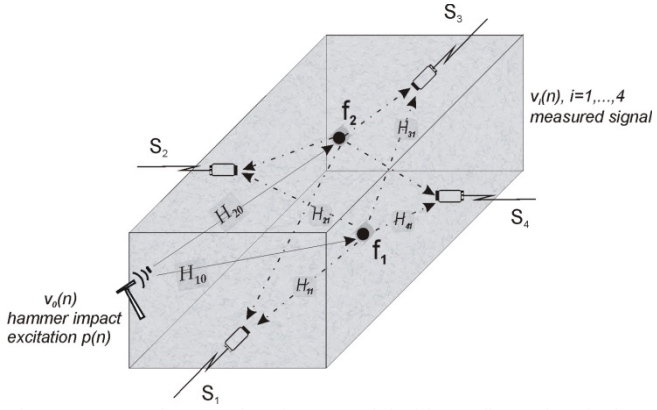

Fig. 2. Impact-echo procedure for a material with two flaw points f. The impact is due to the hammer and the sensors S register the echoes.

In ICA mixture modelling, it is assumed that feature (observation) vectors $\mathbf{x}_k$ corresponding to a given class C_k $(k = 1 \ldots K)$ are the result of applying a linear transformation defined by matrix $\mathbf{A}_k$ to a (source) vector $\mathbf{s}_k$, whose elements are independent random variables, plus a bias vector $\mathbf{b}_k$, i.e.,

$$\mathbf{x}_k = \mathbf{A}_k \mathbf{s}_k + \mathbf{b}_k \qquad k = 1, \ldots, K \tag{1}$$

This indicates that, in principle, a different ICA model should be required for every specific defect (defective zone with particular geometry), every specific deployment of the sensors, and every specific impact location. Thus, we can formulate the problem of classification of materials with different quality conditions, which are inspected by impact-echo in the ICAMM framework.

Although we could use the spectrum of the recorded signals in the estimation of the ICA model, in this paper we will follow a different approach. Instead of working with the raw data coming from the sensors, we obtain a feature vector that represents important time and frequency characteristics of the signals. In this way, we reduce the dimensions of the data vector since we do not have to work with the spectrum of the signals but with some signature of it.

Before feature extraction, the signals of the accelerometers were normalized using the maximum of the impact signal amplitude. The features extracted from the impact-echo signals are the following:

- Principal frequency $f_{\max}$:
$$f_{\max} \,/\, \left|\mathcal{F}\{x(t)\}_{f_{\max}}\right| \geq \left|\mathcal{F}\{x(t)_f\}\right|, \forall_f$$
- Principal frequency amplitude $A_{f_{\max}}$:
$$A_{f_{\max}} = \left|\mathcal{F}\{x(t)\}_{f_{\max}}\right|$$
- Centroid frequency f_c:
$$f_c = \frac{\int_{f_1}^{f_2} f \cdot \left|\mathcal{F}\{x(t)\}\right| df}{\int_{f_1}^{f_2} \left|\mathcal{F}\{x(t)\}\right| df}$$
- Signal power P:
$$P = \frac{\int_0^T |x(t)|^2 \, dt}{T}$$
- Principal frequency attenuation $\beta_{f_{\max}}$:
$$x_{f_{\max}}(t) = \mathcal{F}^{-1}\{\mathcal{F}\{x(t)\} \cdot BPF(f_{\max} \pm \Delta)\};$$
$$envelope\left(x_{f_{\max}}(t)\right) = A_{f_{\max}} e^{-t\beta_{f_{\max}}}$$
- Total signal attenuation β:
$$envelope(x(t)) = Ae^{-t\beta}$$
- Initial value of the attenuation curve P_0:
$$P_0 = 10\log(A)$$

where $x(t)$ is the recorded signal; $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ are the Fourier and the inverse Fourier transforms, respectively; and $BPF(f_{\max} \pm \Delta)$ is a narrow band pass filter centred in $f_{\max}$.

Combining these values, we obtain the feature vector $\mathbf{x}_k$ where k indicates the unknown class of the vector. At this point, we have to estimate the ICAMM parameters (1), i.e., the set of mixing matrices $\mathbf{A}_k$, sources $\mathbf{s}_k$ and bias terms $\mathbf{b}_k$ for every kind of defect $k = 1, \ldots, K$.

Assuming independent feature vectors in the training set, we may write the log-likelihood of the observations in the form:

$$L(\mathbf{X}/\boldsymbol{\Psi}) = \log p(\mathbf{X}/\boldsymbol{\Psi}) = \sum_{n=1}^{N} \log p\left(\mathbf{x}^{(n)}/\boldsymbol{\Psi}\right) \qquad (2)$$

where $\boldsymbol{\Psi}$ is a compact notation for all the unknown parameters $\mathbf{W}_k = \mathbf{A}_k^{-1}$, $\mathbf{b}_k$ for all the classes. We summarize the ICAMM algorithm that can be found in [9] which is based on a non-parametric source pdf estimation, supervised-unsupervised learning, and possibility of selecting a particular ICA algorithm:

0 Initialize $i = 0$, $\mathbf{W}_k(0)$, $\mathbf{b}_k(0)$.

1 Compute
$\mathbf{s}_k^{(n)}(i) = \mathbf{W}_k(i)\left(\mathbf{x}^{(n)} - \mathbf{b}_k(i)\right) \quad k = 1..K \quad n = 1...N$

2 Directly use $p\left(C_k/\mathbf{x}^{(n)}, \boldsymbol{\Psi}\right)(i) = p\left(C_k/\mathbf{x}^{(n)}, \boldsymbol{\Psi}\right)$ for those $k-n$ pairs with knowledge about $p\left(C_k/\mathbf{x}^{(n)}, \boldsymbol{\Psi}\right)$.
Compute

$$p\left(C_k/\mathbf{x}^{(n)}, \boldsymbol{\Psi}\right)(i) = \frac{|\det \mathbf{W}_k(i)|.p\left(\mathbf{s}_k^{(n)}(i)\right)}{\sum_{k'=1}^{K} |\det \mathbf{W}_{k'}(i)| p\left(\mathbf{s}_{k'}^{(n)}(i)\right)} \qquad k = 1...K$$

for the rest of $k-n$ pairs. Use

$$p\left(s_{km}^{(n)}\right) = a \cdot \sum_{n' \neq n} e^{-\frac{1}{2}\left(\frac{s_{km}^{(n)} - s_{km}^{(n')}}{h}\right)^2} \quad m = 1..M \quad k = 1..K$$

to estimate the marginals $p\left(\mathbf{s}_k^{(n)}(i)\right)$.

3 Use the selected ICA algorithm to compute the increments $\Delta_{ICA}^{(n)}\mathbf{W}_k(i)$ corresponding to the observation $\mathbf{x}^{(n)}, n = 1, \ldots N$, that would be applied in $\mathbf{W}_k(i)$, in an "isolated" learning of class C_k. Compute the total increment by

$$\Delta\mathbf{W}_k(i) = \sum_{n=1}^{N} \Delta_{ICA}^{(n)}\mathbf{W}_k(i) \cdot p\left(C_k/\mathbf{x}^{(n)}, \boldsymbol{\Psi}\right)(i)$$

Update
$\mathbf{W}_k(i+1) = \mathbf{W}_k(i) + \alpha \cdot \Delta\mathbf{W}_k(i) \quad k = 1...K$.

4 Compute $\Delta\mathbf{b}_k(i)$ using

$$\Delta\mathbf{b}_k(i) = \sum_{n=1}^{N}\left[-diag\left[\mathbf{f}\left(\mathbf{s}_k^{(n)}\right)\right]\mathbf{w}_{km}(i) \cdot p\left(C_k/\mathbf{x}^{(n)}, \boldsymbol{\Psi}\right)(i)\right].$$

Use $f\left(s_{km}^{(n)}\right) = \frac{1}{h^2}\left[\frac{\sum_{n' \neq n} s_{km}^{(n')} \cdot e^{-\frac{1}{2}\left(\frac{s_{km}^{(n)} - s_{km}^{(n')}}{h}\right)^2}}{\sum_{n' \neq n} e^{-\frac{1}{2}\left(\frac{s_{km}^{(n)} - s_{km}^{(n')}}{h}\right)^2}} - s_{km}^{(n)}\right]$ to estimate $\mathbf{f}\left(\mathbf{s}_k^{(n)}\right)$.

Actualize $\mathbf{b}_k(i+1) = \mathbf{b}_k(i) + \beta \cdot \Delta\mathbf{b}_k(i) \quad k = 1...K$, or simply re-estimate

$$\mathbf{b}_k(i+1) = \frac{\sum_{n=1}^{N} \mathbf{x}^{(n)} p\left(C_k/\mathbf{x}^{(n)}, \boldsymbol{\Psi}\right)(i)}{\sum_{n=1}^{N} p\left(C_k/\mathbf{x}^{(n)}, \boldsymbol{\Psi}\right)(i)} \qquad k = 1...K$$

5 Go back to step 1, with the new values
$\mathbf{W}_k(i+1)$, $\mathbf{b}_k(i+1)$ and $i \rightarrow i+1$
until convergence

III. Hierarchical ICA Mixtures

Once the ICAMM models have been obtained for every kind of defect, we can start the agglomerative clustering of them [32].

The conditional probability density of an observation vector $\mathbf{x}$ for cluster $C_k^h, k = 1,2,...,K-h+1$ in layer $h = 1,2,...,K$ is $p\left(\mathbf{x}/C_k^h\right)$. At the first level, $h = 1$, it is modelled by the K ICA mixtures obtained in the previous section, i.e., $p\left(\mathbf{x}/C_k^1\right)$ is:

$$p\left(\mathbf{x}/C_k^1\right) = \left|\det \mathbf{A}_k^{-1}\right| p\left(\mathbf{s}_k\right), \ \mathbf{s}_k = \mathbf{A}_k^{-1}\left(\mathbf{x} - \mathbf{b}_k\right) \qquad (3)$$

At each consecutive level, two clusters are merged according to some minimum distance measure until only one cluster is reached at level $h = K$. For the distance measure, we use the symmetric Kullback-Leibler divergence between the ICA mixtures, which is defined for the clusters u, v by:

$$D_{KL}\left(C_u^h, C_v^h\right) = \int p\left(\mathbf{x}/C_u^h\right) \log \frac{p\left(\mathbf{x}/C_u^h\right)}{p\left(\mathbf{x}/C_v^h\right)} d\mathbf{x} + \int p\left(\mathbf{x}/C_v^h\right) \log \frac{p\left(\mathbf{x}/C_v^h\right)}{p\left(\mathbf{x}/C_u^h\right)} d\mathbf{x} \qquad (4)$$

For layer h=1, from (4) we can obtain the following:

$$D_{KL}\left(C_u, C_v\right) = D_{KL}\left(p_{\mathbf{x}_u}(\mathbf{x}) // p_{\mathbf{x}_v}(\mathbf{x})\right) = \int p_{\mathbf{x}_u}(\mathbf{x}) \log \frac{p_{\mathbf{x}_u}(\mathbf{x})}{p_{\mathbf{x}_v}(\mathbf{x})} d\mathbf{x} + \int p_{\mathbf{x}_v}(\mathbf{x}) \log \frac{p_{\mathbf{x}_v}(\mathbf{x})}{p_{\mathbf{x}_u}(\mathbf{x})} d\mathbf{x} \qquad (5)$$

For brevity we write $p_{\mathbf{x}_u}(\mathbf{x}) = p\left(\mathbf{x}/C_u^1\right)$) and omit the superscript $h = 1$. For simplicity, we impose the independence hypothesis and we suppose that both clusters

have the same number of sources M :

$$p_{\mathbf{x}_u}(\mathbf{x}) = \frac{\prod_{i=1}^{M} p_{s_{u_i}}(s_{u_i})}{|\det \mathbf{A}_u|}, \quad s_{u_i} = \mathbf{A}_{u_i}^{-1}(\mathbf{x} - \mathbf{b}_{u_i})$$
$$p_{\mathbf{x}_v}(\mathbf{x}) = \frac{\prod_{j=1}^{M} p_{s_{v_j}}(s_{v_j})}{|\det \mathbf{A}_v|}, \quad s_{v_j} = \mathbf{A}_{v_j}^{-1}(\mathbf{x} - \mathbf{b}_{v_j}) \tag{6}$$

where s_{u_i}, $i = 1,\ldots,M$ and s_{v_j}, $j = 1,\ldots,M$ are the i - and j -th elements of the source vectors s_{u_i} and s_v for the corresponding clusters C_u and C_v. $A_{u_i}^{-1}$ and $A_{v_j}^{-1}$ are the the i - and j -th rows of the demixing matrices A_u^{-1} and A_v^{-1} for the corresponding clusters C_u and C_v.

IV. Merging ICA Clusters with Kernel-based Source Densities

The pdf of the sources is approximated by a non-parametric kernel-based density for both clusters:

$$p_{s_{u_i}}(s_{u_i}) = \sum_{n=1}^{N} a e^{-\frac{1}{2}\left(\frac{s_{u_i} - s_{u_i}(n)}{\Delta}\right)^2}, \; p_{s_{v_j}}(s_{v_j}) = \sum_{n=1}^{N} a e^{-\frac{1}{2}\left(\frac{s_{v_j} - s_{v_j}(n)}{\Delta}\right)^2} \tag{7}$$

where $s_{u_i}(n)$ and $s_{v_j}(n)$ are the sources s_{u_i} and s_{v_j} at time n. Again for simplicity, we have assumed the same kernel function with the parameters a, Δ for all the sources and the same number of samples N for each one. Note that this corresponds to a mixture of Gaussian models where the number of Gaussians is maximum (one for every observation) and the parameters are equal. Reducing the pdf of the sources to a standard mixture of Gaussians with a different number of components and priors for each source does not help in computing the Kullback-Leibler distance because there is no analytical solution for it. Therefore, we prefer to maintain the non-parametric approximation of the pdf in order to model more complex distributions than a mixture of a small finite number of Gaussians, such as three or four.

The symmetric Kullback-Leibler distance between the clusters u, v can be expressed as:

$$D_{KL}(p_{\mathbf{x}_u}(\mathbf{x}) // p_{\mathbf{x}_v}(\mathbf{x})) = -H(\mathbf{x}_u) - H(\mathbf{x}_v) - \\ -\int p_{\mathbf{x}_u}(\mathbf{x}) \log p_{\mathbf{x}_v}(\mathbf{x}) d\mathbf{x} - \int p_{\mathbf{x}_v}(\mathbf{x}) \log p_{\mathbf{x}_u}(\mathbf{x}) d\mathbf{x} \tag{8}$$

where $H(\mathbf{x})$ is the entropy, which is defined as $H(\mathbf{x}) = -E[\log p_{\mathbf{x}}(\mathbf{x})]$, and the other terms are the cross-entropies $E_{\mathbf{x}_v}[\log p_{\mathbf{x}_u}(\mathbf{x})], E_{\mathbf{x}_u}[\log p_{\mathbf{x}_v}(\mathbf{x})]$.

The entropy for the cluster u can be calculated through the entropy of the sources of that cluster taking into account the linear transformation of the random variables and their independence (6):

$$H(\mathbf{x}_u) = \sum_{i=1}^{M} H(s_{u_i}) + \log|\det \mathbf{A}_u| \tag{9}$$

The entropy of the sources cannot be analytically calculated. Instead, we can obtain a sample estimate of $\hat{H}(s_{u_i})$ using the training data. Denote the i -th source obtained for the cluster u by $\{s_{u_i}(1), s_{u_i}(2), \ldots, s_{u_i}(Q_i)\}$, where Q_i is the number of observations. The entropy can be approximated as follows:

$$\hat{H}(s_{u_i}) = -\hat{E}[\log p_{s_{u_i}}(s_{u_i})] = -\frac{1}{Q_i}\sum_{n=1}^{Q_i} \log p_{s_{u_i}}(s_{u_i}(n)) \\ p_{s_{u_i}}(s_{u_i}(n)) = \sum_{l=1}^{N} a e^{-\frac{1}{2}\left(\frac{s_{u_i}(n) - s_{u_i}(l)}{\Delta}\right)^2} \tag{10}$$

The entropy of $H(\mathbf{x}_v)$ is obtained analogously to (10). Other possible approximations are available, for example to use synthetic data produced from the distributions of the sources instead of the data used to learn the parameters of the ICA mixture model.

Once the entropy is computed, we have to obtain the cross-entropy terms.

$$E_{\mathbf{x}_u}[\log p_{\mathbf{x}_v}(\mathbf{x})] = \int p_{\mathbf{x}_u}(\mathbf{x}) \log p_{\mathbf{x}_v}(\mathbf{x}) d\mathbf{x} = \\ \int p_{\mathbf{x}_u}(\mathbf{x}) \log \frac{\prod_{i=1}^{M} p_{s_{v_i}}(s_{v_i})}{|\det \mathbf{A}_v|} d\mathbf{x} \\ E_{\mathbf{x}_v}[\log p_{\mathbf{x}_u}(\mathbf{x})] = \int p_{\mathbf{x}_v}(\mathbf{x}) \log p_{\mathbf{x}_u}(\mathbf{x}) d\mathbf{x} = \\ \int p_{\mathbf{x}_v}(\mathbf{x}) \log \frac{\prod_{i=1}^{M} p_{s_{u_i}}(s_{u_i})}{|\det \mathbf{A}_u|} d\mathbf{x} \tag{11}$$

Considering the relationships $\mathbf{x} = \mathbf{A}_u \mathbf{s}_u + \mathbf{b}_u$, $\mathbf{x} = \mathbf{A}_v \mathbf{s}_v + \mathbf{b}_v$ and thus $\mathbf{s}_v = \mathbf{A}_v^{-1}(\mathbf{A}_u \mathbf{s}_u + \mathbf{b}_u - \mathbf{b}_v)$, we obtain for the first cross-entropy in (11):

$$\int p_{\mathbf{x}_u}(\mathbf{x}) \log p_{\mathbf{x}_v}(\mathbf{x}) d\mathbf{x} = \int p_{\mathbf{s}_u}(\mathbf{s}) \log \frac{\prod_{i=1}^{M}\sum_{n=1}^{N} a e^{-\frac{1}{2}\left(\frac{s_{v_i} - s_{v_i}(n)}{\Delta}\right)^2}}{|\det \mathbf{A}_v|} d\mathbf{s}, \tag{12}$$

with s_{v_i} being the i -th element of the vector $\mathbf{s}_v$, i.e., $s_{v_i} = [\mathbf{A}_v^{-1}(\mathbf{A}_u \mathbf{s} + \mathbf{b}_u - \mathbf{b}_v)]_i$.

Using (12), by applying the independence of the sources

for the cluster u, we obtain:

$$\int p_{\mathbf{x}_u}(\mathbf{x})\log p_{\mathbf{x}_v}(\mathbf{x})\,d\mathbf{x} = -\log\left|\det \mathbf{A}_v\right| + \int\prod_{j=1}^{M} p_{s_{u_j}}(s_j)\log\prod_{i=1}^{M}\sum_{n=1}^{N} a e^{-\frac{1}{2}\left(\frac{s_{v_i}-s_{v_i}(n)}{\Delta}\right)^2} d\mathbf{s} = -\log\left|\det \mathbf{A}_v\right| + \sum_{i=1}^{M}\int p_{s_{u_M}}(s_M)ds_M \ldots \int p_{s_{u_1}}(s_1)\log\sum_{n=1}^{N} a e^{-\frac{1}{2}\left(\frac{s_{v_i}-s_{v_i}(n)}{\Delta}\right)^2} ds_1 \tag{13}$$

Again, there is no analytical solution to (13), so we have to use numerical alternatives to approximate the cross-entropy. Following the same idea as above with the entropy, we can use the data corresponding to every source for cluster u in order to approximate the expectation of (13). Assuming that we have or can generate Q_i observations according to distribution $p_{s_{u_i}}(s_i), i=1,\ldots,M$, we can estimate

$$\int p_{s_{u_M}}(s_M)ds_M \ldots \int p_{s_{u_1}}(s_1)\log\sum_{n=1}^{N} a e^{-\frac{1}{2}\left(\frac{s_{v_i}-s_{v_i}(n)}{\Delta}\right)^2} ds_1 \simeq \simeq \frac{1}{\prod_{i=1}^{M} Q_i}\sum_{s_M=1}^{Q_M}\ldots\sum_{s_1=1}^{Q_1}\log\sum_{n=1}^{N} a e^{-\frac{1}{2}\left(\frac{s_{v_i}-s_{v_i}(n)}{\Delta}\right)^2} \tag{14}$$

With $s_{v_i} = (\mathbf{Bs})_i + \mathbf{c}_i, \mathbf{B} = \mathbf{A}_v^{-1}\mathbf{A}_u, \mathbf{c}_i = \mathbf{A}_v^{-1}(\mathbf{b}_u - \mathbf{b}_v)$, $\mathbf{s} = \left[s_1(k),\ldots,s_M(l)\right]^T, k\in[1,Q_1],\ldots,l\in[1,Q_M]$. Of course, the other term in (11) is obtained in a similar way, considering that now the expectations are obtained by averaging the pdf of the sources of the other cluster.

Taking into account all the terms in (5), the symmetrical Kullback-Leibler distance between clusters u, v can be computed numerically from the samples following the corresponding distribution $\{s_{u_i}(1), s_{u_i}(2),\ldots,s_{u_i}(Q)\}$, $i=1,\ldots,M$, $\{s_{v_j}(1), s_{v_j}(2),\ldots,s_{v_j}(Q)\}$, $j=1,\ldots,M$ (we assume that the number of samples per source is the same for all of them).

The computations can also be easily extended to the case where the number of sources in every class is not the same. In the case that the distributions are approximated by just a single Gaussian (keeping in mind that the ICA problem reduces to the PCA problem since there is an indetermination defined by an orthogonal matrix that is not identifiable) and the distance is obtained analytically for the first level of the hierarchy, the distance between two multivariate normal distributions of dimension M, $p_u(\mathbf{x}) = N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u), p_v(\mathbf{x}) = N(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v)$ would be

$$D_{KL}\left(p_u(\mathbf{x}) // p_v(\mathbf{x})\right) = tr\left(\boldsymbol{\Sigma}_u\boldsymbol{\Sigma}_v^{-1}\right) + tr\left(\boldsymbol{\Sigma}_v\boldsymbol{\Sigma}_u^{-1}\right) - 2M + tr\left[\left(\boldsymbol{\Sigma}_u^{-1} + \boldsymbol{\Sigma}_v^{-1}\right)(\boldsymbol{\mu}_u - \boldsymbol{\mu}_v)(\boldsymbol{\mu}_u - \boldsymbol{\mu}_v)^T\right] \tag{15}$$

where $tr(\mathbf{A})$ is the trace of matrix $\mathbf{A}$.

Once the distances are obtained for all the clusters, the two clusters with minimum distance are merged at a certain level. This is repeated in each step of the hierarchy until one cluster at the level $h = K$ is reached. To merge a cluster at level h, we can calculate the distances from the distances of level $h-1$. Suppose that from level $h-1$ to h the clusters C_u^{h-1}, C_v^{h-1} are merged in cluster C_w^h. Then, the density for the merged cluster at level h is:

$$p_h\left(\mathbf{x}/C_w^h\right) = \frac{p_{h-1}\left(C_u^{h-1}\right)p_{h-1}\left(\mathbf{x}/C_u^{h-1}\right) + p_{h-1}\left(C_v^{h-1}\right)p_{h-1}\left(\mathbf{x}/C_v^{h-1}\right)}{p_{h-1}\left(C_u^{h-1}\right) + p_{h-1}\left(C_v^{h-1}\right)} \tag{16}$$

where $p_{h-1}\left(C_u^{h-1}\right), p_{h-1}\left(C_v^{h-1}\right)$ are the priors or proportions of the clusters u, v at level $h-1$. The rest of the terms are the same in the mixture model at level h as at level $h-1$. The only difference from one level to the next one in the hierarchy is that there is one cluster less and the prior for the new cluster is the sum of the priors of its components and the density the weighted average of the densities that are merged to form it. Therefore, the estimation of the distance at level h can be done easily starting from the distances at level $h-1$ and so on until level $h=1$. Consequently, we can calculate the distances at level h from a cluster C_z^h to a merged cluster C_w^h that was obtained by the agglomeration of clusters C_u^{h-1}, C_v^{h-1} at level $h-1$ as the distance to its components weighted by the mixing proportions:

$$D_h\left(p_h\left(\mathbf{x}/C_w^h\right) // p_h\left(\mathbf{x}/C_z^h\right)\right) = \frac{p_{h-1}\left(C_u^{h-1}\right)\cdot D_{h-1}\left(p_{h-1}\left(\mathbf{x}/C_u^{h-1}\right) // p_{h-1}\left(\mathbf{x}/C_z^{h-1}\right)\right)}{p_{h-1}\left(C_u^{h-1}\right) + p_{h-1}\left(C_v^{h-1}\right)} + \frac{p_{h-1}\left(C_v^{h-1}\right)\cdot D_{h-1}\left(p_{h-1}\left(\mathbf{x}/C_v^{h-1}\right) // p_{h-1}\left(\mathbf{x}/C_z^{h-1}\right)\right)}{p_{h-1}\left(C_u^{h-1}\right) + p_{h-1}\left(C_v^{h-1}\right)} \tag{17}$$

As at level 1, we can obtain the decision rule to assign a new data to a cluster in the hierarchy at level h by applying Bayes' theorem:

$$\arg\max_{C_k} p_h\left(C_k/\mathbf{x}\right) = \frac{p_h\left(\mathbf{x}/C_k\right)p_h\left(C_k\right)}{\sum_{i=1}^{K-h+1} p_h\left(\mathbf{x}/C_i\right)p_h\left(C_i\right)} \tag{18}$$

V. Results: Non destructive testing (NDT)

The hierarchical clustering algorithm was applied to the field of quality control of materials. The objective was to automatically obtain an appropriate hierarchical classification of parallelepiped-shaped material evaluated in real experiments using the impact-echo technique [34].

The materials were pieces of aluminium alloy series 2000 of 0.07x0.05x0.22 m. (width, height, and length, respectively). Up to three defects per piece were drilled in different locations of each piece. The defects passed through the pieces and consisted of holes in the shape of cylinders 10 mm. Ø and cracks in the shape of parallelepipeds of 5x20 mm. cross-section.

The material was excited by an impact and its response was measured by a multichannel system of sensors (accelerometers). Fig.3 shows the setup of the impact-echo experiments, which includes sensor configuration, impact localization, supports of the piece, and coordinate axes.

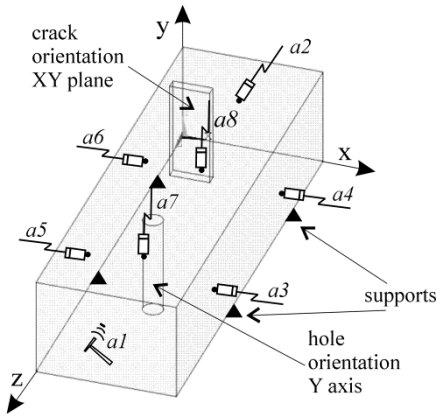

Fig.3. Impact-echo experiment for hierarchical classification

The example in Fig.3 contains two defects: one hole in the Y axis and one crack in the XY plane.

The experiments were performed using the following equipment:

(i) Instrumented impact hammer 084A14 PCB

(ii) 7 accelerometers (*a1-a7*) 353B17 PCB

(iii) ICP signal conditioner F482A18

(iv) Data acquisition module 6067E

(v) Notebook. The acquisition parameters were: sampling frequency=100,000 kHz, and observation time=50 ms.

The total number of experiments was 1200 executions of the impact-echo test from 60 specimens. The materials were from 5 categories. The first four categories corresponded to one-defect materials: one hole oriented in the X axis, one hole oriented in the Y axis, one crack oriented in the XY plane, and one crack oriented in the XZ plane. The fifth category corresponded to multiple defects.

Before feature extraction, the signals of the 7 accelerometers were normalized using the maximum of the impact signal amplitude. The feature vector has 7 dimensions: principal frequency; principal frequency amplitude and attenuation; centroid frequency; signal power; initial value of the attenuation curve and signal attenuation.

The dimensionality of the feature space was reduced from 49 features to 10 components by PCA for an explained variance greater than 92%.

The results of the hierarchical classification for the impact-echo application, using 0.3 as the supervision ratio in the ICAMM algorithm are shown in Fig.4. The accuracy of the classification for this case was the following (the values are in percentage):

(i) The bottom level— multiple defects, one hole in the X axis, one hole in the Y axis, one crack in the XY plane, and one crack in the ZY plane (100, 82.86, 71.74, 67.42, 65.83, respectively)

(ii) The intermediate level— multiple defects, holes, cracks (100, 88.37, 72.11, respectively)

(iii) The penultimate level— multiple defects, one defect (100, 88.34, respectively).

Considering the great complexity of the problem, the classification accuracy is high. Thus, the procedure was able to automatically learn the defect patterns of the materials and build a meaningful hierarchical structure (dendrogram) that allows the pieces were allocated in the right place of the classification tree.

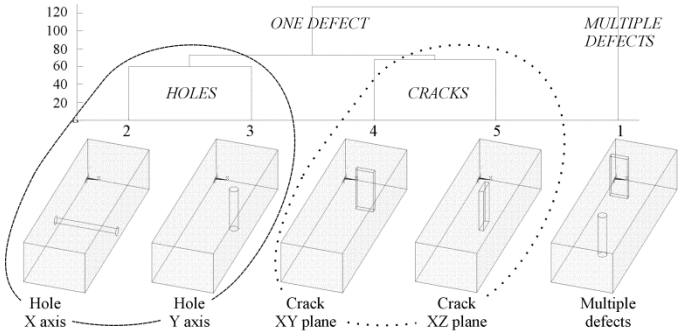

Fig.4. Hierarchy obtained for the impact-echo experiments. Meaningful groupings of the materials are found

The hierarchy obtained in Fig.4 can be interpreted in the following way. At the highest level of the hierarchy, the pieces are divided into two classes that represent the material condition: multiple defects and one defect. The intermediate levels of the hierarchy separate the specimens into three classes corresponding to the kind of defect in the materials: multiple defects, holes, and cracks. Finally, the lowest level of the hierarchy splits the materials into five classes that are related to the kind and orientation of the defects: multiple defects, one hole in the X axis, one hole in the Y axis, one crack in the XY plane, and one crack in the ZY plane. Therefore, the hierarchical levels would represent abstract conceptualizations about the condition of the material, i.e., general material condition, kind of defect, and defect orientation. This result is relevant taking into account the great difficulty of finding significant hierarchical patterns in this type of application.

VI. Conclusion

An application in material quality control using impact-echo testing has been presented. It is based on a method for agglomerative hierarchical clustering assuming an underlying ICA mixture model. The algorithm uses the ICAMM parameters estimated at the bottom of the hierarchy to create higher levels by grouping clusters. It is based on the symmetric Kullback-Leibler divergence between the clusters using the ICA parameters assuming non-parametric kernel-based source densities. Different structures of classification can be derived at the different levels of the bottom-up merging.

The results of application to NDT using impact-echo testing demonstrated that meaningful classification trees were obtained. From a feature space of temporal and frequency parameters extracted from the impact-echo signals at the lowest hierarchical level, defective materials were grouped consistently at higher levels. The groupings showed significant separation between materials with holes and cracks, and between materials with one defect and multiple defects. This kind of classification is very useful in certain industries (e.g., marble factories) where the knowledge about the kind of defect is critical in optimizing the manufacturing process of block cutting.

References

[1] M. Sansalone, W. Street, Impact-echo: Non-destructive evaluation of concrete and masonry. New York: Bullbrier Press, 1997.

[2] N.J. Carino, The impact-echo method: an overview, in: Structures Congress and Exposition (Ed. Chang, P.C.), Washington D.C., 2001, pp. 1-18.

[3] J. Igual, A. Camacho, L. Vergara, Blind Source Separation Technique for Extracting Sinusoidal Interferences in Ultrasonic Non-Destructive Testing, Journal of VLSI Signal Processing 38 (2004) 25-34.

[4] A. Salazar, L. Vergara, J. Igual, and J. Gosalbez, "Blind source separation for classification and detection of flaws in impact-echo testing," *Mechanical Systems and Signal Processing*, vol. 19, no. 6, pp. 1312-1325, 2005.

[5] J. Antoni, Blind separation of vibration components: Principles and demonstrations, Mechanical Systems and Signal Processing 19 (2005) 1116-1180.

[6] A. Salazar, L. Vergara, and R. Llinares, "Learning Material Defect Patterns by Separating Mixtures of Independent Component Analyzers from NDT Sonic Signals," *Mechanical Systems and Signal Processing*, vol. 24, no. 6, pp. 1870-1886, 2010.

[7] T. W. Lee, M.S. Lewicki, T.J. Sejnowski, ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10) (2000) 1078-1089.

[8] R. Choudrey, S. Roberts, Variational Mixture of Bayesian Independent Component Analysers, Neural Computation 15 (1) (2002) 213-252.

[9] A. Salazar, L. Vergara, A. Serrano, J. Igual, A general procedure for learning mixtures of independent component analyzers, Pattern Recognition 43 (2010) 69-85.

[10] B. Everitt, S. Landau, and M. Leese, *Cluster Analysis*. Fourth edition, Wiley, London: Arnold, 2001.

[11] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[12] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. on Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.

[13] D.T. Pham and A.A. Afify, "Clustering techniques and their applications in engineering," *Proc. of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 221, no. 11, pp. 1445-1459, 2007.

[14] G. Lance and W. Williams, "A general theory of classification sorting strategies: 1. Hierarchical systems," *The Computer Journal*, vol. 9, no. 4, pp. 373-380, 1967.

[15] C. Williams, "A MCMC approach to hierarchical mixture modelling," in *Proc. NIPS, vol. 13*, 1999, pp. 680-686.

[16] R.M. Neal, "Density modeling and clustering using Dirichlet diffusion trees," *Bayesian Statistics*, vol. 7, pp. 619-629, 2003.

[17] C. Kemp, T.L. Griffiths, S. Stromsten, and J.B. Tenenbaum, "Semi-supervised learning with trees," in *Proc. NIPS, vol. 17*, 2003.

[18] N. Vasconcelos and A. Lippman, "Learning mixture hierarchies," in *Proc. NIPS, vol. 12*, 1998, pp. 606-612.

[19] A. Stolcke and S. Omohundro, "Hidden Markov model induction by Bayesian model merging," in *Proc. NIPS, vol. 6*, 1992, pp. 11-18.

[20] J.D. Banfield and A.E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 43, pp. 803-821, 1993.

[21] S. Vaithyanathan and B, Dom, "Model-based hierarchical clustering," *Uncertainty in Artificial Intelligence*, vol. 16, pp. 599-608, 2000.

[22] E. Segal, D. Koller, and D. Ormoneit, "Probabilistic abstractions hierarchies," in *Proc. NIPS, vol. 15*, 2001, pp. 913-920.

[23] M.F. Ramoni, P. Sebastiani, and I.S. Kohane, "Cluster analysis of gene expression dynamics," *National Academy of Sciences*, vol. 99, pp. 9121-9126, 2003.

[24] N. Friedman, "Pcluster: Probabilistic agglomerative clustering of gene expression profiles," Technical Report 80, Herbew University, 2003.

[25] K.A. Heller and Z. Ghahramani, "Bayesian hierarchical clustering," in *Proc. ACM Int. Conf. Proc. Series, vol. 119, 22nd Int. Conf. on Machine Learning*, Bonn, Germany, 2005, pp. 297-304.

[26] M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443-482, 1999.

[27] M.E. Tipping and C.M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 61, Part 3, pp. 611-622, 1999.

[28] C.M. Bishop and M.E. Tipping, "A hierarchical latent variable model for data visualization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 281-293, 1998.

[29] A. Hyvärinen, P.O. Hoyer, and M. Inki, "Topographic Independent Component Analysis," *Neural Computation*, vol. 13, no. 7, pp. 1527-1558, 2001.

[30] H.J. Park and T.W. Lee, "Capturing nonlinear dependencies in natural images using ICA and mixture of Laplacian distribution," *Neurocomputing*, vol. 69, no, 13-15, pp. 1513-1528, 2006.

[31] F.R. Bach and M.I. Jordan, "Beyond independent components: trees and clusters," *Journal of Machine Learning Research*, vol. 3, pp. 1205-1233, 2003.

[32] A. Salazar, J. Igual and L. Vergara, "Learning hierarchies from ICA mixtures", IEEE IJCNN 2007, Orlando, USA.

[33] D.J. Mackay, *Information theory, inference, and learning algorithms*. Cambridge University Press, 2004.

[34] M. Sansalone and W.B. Streett, *Impact-echo: Non-destructive evaluation of concrete and masonry*. Bullbrier Press, USA, 1997.