Professor Gheorghe RUXANDA, PhD
Professor Ion SMEUREANU, PhD
The Bucharest Academy of Economic Studies

UNSUPERVISED LEARNING WITH EXPECTED MAXIMIZATION ALGORITHM

Abstract. *The paper presents the theoretical and numerical aspects of Expected Maximization algorithm for the particular case of estimating the mixture probability distribution parameters, respectively, for the case of incomplete data estimates, as well as for the particular case of unsupervised statistical learning. The probabilistic context of algorithm is defined and analyzed, its logic elements are deduced and rendered and the particular form of algorithm is presented in case of Gaussian conditioned distributions. Starting from the basic ideas of EM algorithm, it is justified the equivalence between the maximization of likelihood function logarithm and the maximization of conditioned entropy of prior distributions of statistical subpopulations of a mixture probability distribution, entropy estimated according to posterior distributions of these statistical subpopulations. A numerical illustration of using EM algorithm for unsupervised statistical learning processes, for the Gaussian mixture distributions case is also presented.*

Key words: *Learning Machines, Mixture Distributions, ML Estimators, EM Algorithm, Pattern Recognition, Cluster Analysis.*

JEL Classification: C01, C11, C13, C38, C45, C46, C81, C87

I. Introduction

EM Algorithm (**E**xpected **M**aximization) is a powerful algorithm of estimating the probability distribution parameters with the nature of certain *mixture probability distributions*, respectively, of estimating the probability distribution parameters including *latent* or *unobservable* random variables. The estimate of probability distribution parameters of this kind is known under the name of *incomplete data estimate*. The estimate problems of this type are frequently met in the context of *statistical learning theory,* especially, in the context of *pattern recognition* problems. In other words, EM algorithm is an iterative procedure useful to determine Maximum Likelihood estimators (ML), under the conditions of certain *implicit information,* represented by *hidden* or *missing* data. As it is known, getting maximum likelihood estimators is based on the idea of determining the

estimators such as the available information to get an as high as possible likelihood, with respect to the values of these estimators.

EM algorithm was proposed for the first time in 1977 year by Dempster, Laird and Rubin, at the same time with the publication of the paper *Maximum Likelihood from Incomplete Data via the EM Algorithm,* although the idea of this approach has appeared earlier in an unpublished paper of Stuart Russel. Afterwards, the concerns related to EM algorithm and to its possible applications have been largely expanded, this leading to the appearance of numerous variants of this algorithm (incremental, batch, online etc.).

EM algorithm can be used to solve numerous and various categories of problems, such as those related to: estimation of incomplete data models or which include unobserved variables, cluster analysis and unsupervised pattern recognition, estimation of mixture probability distribution parameters, estimation of dynamic system states, estimation of Markov chain model parameters with hidden information, Data Mining analyses, computer vision, reconditioning the multi frame super resolution images etc.

In most cases, where EM algorithm is used, the available information (sample of observations) is considered to come from a statistical population $\mathcal{P}$ structured under the form of a number of distinct statistical subpopulations $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$ without known a priori, which are the subpopulation, the observations come from. As statistical population $\mathcal{P}$ is made of several subpopulations, its probability distribution is a mixture probability distribution of subpopulations $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$ and its probability density is defined as a convex combination. The parameters to be estimated in this case are represented by the parameters of probability distributions of the K subpopulations, respectively, the expected values, variances and covariances of these subpopulations, as well as by the "weights" of each subpopulation in the total of statistical population $\mathcal{P}$. These "weights" are the coefficients of convex combination defining the density of mixture probability distribution. The information associated to parameters of "weight" type is the missing or hidden information. The *latent* character of this information results from the fact that, although each observation comes from a certain statistical subpopulation, although it contains information about the respective subpopulation, this information is not "visible", but it appears only in an implicit form, as we do not know exactly which is the subpopulation, the respective observation, comes from. Any observation coming from mixture probability distributions contains the "finger print " of particular distributions it comes from, but at latent level and in a hidden manner.

The mixture probability distribution is a *probabilistic model* associated to a statistical (total) population, structured under the form of a finite or infinite number of statistical distinct populations. The probability distributions which form a mixture, i.e. the probability distributions of statistical subpopulations, are called *elements* or *components* of mixture. In case when the probability distributions of subpopulations are continuous,

the mixture probability density is a convex combination of probability densities of subpopulations, i.e. a linear combination with non-negative coefficients, where the coefficients sum to 1. The coefficients of convex combination of probability densities of subpopulations are called mixture probability *weights.*

As for the mixture probability distributions, we mention two types of problems. The first problem is related to deducing the properties of statistical (total) population on the basis of studying the properties of subpopulations. The second problem refers to the inferences of the properties of statistical subpopulations, on the basis of the available observations of the total statistical population, without knowing the subpopulations these observations come from.

One of the important approaches related directly to inference is the *identification of the origin subpopulation* of available observations, i.e. the subpopulations which are the origin of observations, approach related to the field *of unsupervised learning* or *cluster analysis*. For instance, if we consider the problem of recognizing the hand-writing characters, each character is represented by a matrix of size $p \times q$ where each matrix element is equivalent to a pixel with a value equal to 0 or 1, value depending on the color of the respective pixel, for instance white or black. This means that the total number of characteristics of each subpopulation is $n = p \times q$ i.e. these characteristics can be represented by a $(p \times q)-$ dimensional vector, whose elements are of binary type. Each character is a statistical subpopulation, so that, if the total number of characters is K, we shall have a number of K $(p \times q)$-dimensional statistical subpopulations and the mixture probability will have a number of components equal to the total number of different characters. Each characteristic (pixel) has a probability distribution of Bernoulli type, so that, for each character, we have a number of $(p \times q)$- probability distributions.

A mixture probability distribution can be regarded as being the probability distribution of a *new random variable*, whose values are the values of other random variables, chosen randomly, from a given set of random variables, each random variable of this set, being associated to a certain statistical subpopulation. The random mechanism according to which the values of this new random variable are generated, can be considered as a random mechanism *working in two stages*: in the first stage, it is randomly chosen a random variable of the given set of random variables, and in the second stage, the value of this variable is randomly generated. Practically, in the first stage, a statistical subpopulation is randomly chosen, and in the second stage, an observation of this statistical subpopulation is randomly chosen. It follows that in case of this random mixed mechanism, two types of probability distributions are involved: the probability distribution associated to the first stage and defined besides the given set of random variables, respectively, the probability distributions of random variables of this set. The probability distribution associated to the first stage is the *prior distribution of subpopulations* and the

probability distributions associated to the second stage are the *conditioned distributions* or the *distributions of subpopulations*.

Intuitively, the statistical subpopulations constituting the mixture probability distributions get the nature of certain *causes* or *nature states*, and the observations generated from mixture probability distributions get the nature of certain *effects* determined by the respective causes, but without knowing which is the "contribution" of each cause to producing these effects. So, an observation generated by mixture probability distributions gets the nature of an *observed effect* and the main objective of statistical inference is to probabilistically evaluate the extend to which, the causes contribute to producing the respective effect, i.e. to identify the subpopulation, the respective observation comes from.

The most appropriate approach, in this context, is the *Bayesian approach*, where the fundamental part in probabilistically evaluating the contributions of causes to producing a certain effect, is played by the so – called *posterior probabilities* of causes. The estimate of posterior probabilities of causes or of nature states assumes to know two categories of information: *prior probabilities of causes* or *nature states*, probabilities which, in case of mixture probability distributions, are represented by mixture *weights* i.e. probabilities with which a certain effect in each nature states occurs, called also *observed probabilities*. In case of mixture probability distributions, the prior probabilities are represented by mixture *weights*, and the observed probabilities are represented by the probabilities of observations, at the level of each subpopulation and can be determined only if the probability laws of each subpopulation are known.

As for the nature of mixture probability distributions, we have to specify that the mixture probability distributions must not be confounded with the distribution of a weighted sum of random variables, which abstract the statistical subpopulations, as, in this latter case, the random mechanism of generating the values of random sum type variable is a mechanism working into a single stage: the values of random sum type variables directly result from the values taken by random variables of the respective sum. Moreover, if the probability distributions of random sum type variables are unimodal distributions, then, the probability distribution of weighted sum of random variables is also of unimodal type and directly results from the simultaneous probability law of the respective random variables. In exchange, the probability density of mixture probability distributions is defined via a convex combination of probability densities of the statistical subpopulations, which constitute the mixture, and the mixture probability distribution is always a *multimodal distribution*. The previous renditions of mixture probability distributions are valid also for the case where instead of random variables, vectors of random variables are taken into consideration.

In order to use EM algorithm, with a view to solving the problems related to the estimate of parameters of mixture probability distributions, it is necessary *to know the nature of probability distributions of statistical subpopulations* $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$ distributions

which get the nature of certain conditioned probability distributions. In cases when the characteristics of subpopulations are unknown, the probability densities of these subpopulations can be approximated by using *kernel* type techniques or *average and shifted histograms* type techniques. But, in many cases, when the probability laws of subpopulations are unknown, we can assume that these distributions are *normal distributions*.

Each iteration of EM algorithm includes *two* distinct computational stages, respectively, ***E****xpected stage* and ***M****aximization stage*. According to available observations, in the first stage (E) of each iteration, the unknown values of the variable or of latent variables are estimated, and consequently, mixture distribution weights are estimated. The unknown values of latent variables are estimated on the basis of the estimates produced by algorithm for posterior probabilities of each sample observation. The process of obtaining estimates for posterior probabilities of observations has the nature of a *conditional average (expected)* process, where *available observations* and *approximated values of parameters* at each iteration are used. The *randomly generated values* are used as initial values for the weights of mixture probability distributions and for the parameters of the distributions conditioned by probability. In the second (M) stage of each iteration, the estimates of probability distribution parameters of subpopulations (expected values, variances and covariances) are calculated, according to the maximization of likelihood function associated to sample observations, under the assumption that the values of latent variables and of mixture weights are those ones estimated in previous (E) stage.

The value resulted from the likelihood function estimation, with respect to the estimated values of parameters at any iteration, always represents an *increment*, in comparison with the value of this function at previous iteration, such that, the algorithm convergence is provided. The EM type estimating procedure goes on with the execution of new and new iterations, each iteration assuming the alternative execution of the two stages E and M, till the increment of likelihood function value becomes insignificant.

Bearing in mind the fact that the specificity of its functionality, the EM algorithm can be considered as being an *unsupervised statistical learning algorithm*, with a nature similar to that of *partitioning algorithm*, for which, the number of classes is a prior known. The EM algorithm can be used in order to determine the forms classification, on the basis of approximating the posterior probabilities resulted from the execution of each iteration of this algorithm.

II. Statistical – mathematical Fundamentals of EM Algorithm

Generally speaking, the problem of estimating parameters by using the maximum likelihood method can be defined as follows: if $\mathcal{P}$ is a statistical n–dimensional population, whose (intrinsic) attributes or characteristics are represented by means of the

n–dimensional random vector $\mathbf{X}_{\mathcal{P}}$ and whose probability law is represented by means of *n*–dimensional probability density:

$$f_{\mathcal{P}}(\mathbf{x};\boldsymbol{\theta}), \quad \mathbf{x}\in\mathcal{P},\ \boldsymbol{\theta}\in\boldsymbol{\Theta},$$

where $\boldsymbol{\theta}$ is the parameters vector of probability distribution, and $\boldsymbol{\Theta}$ is the set of parameter values, then the *maximum likelihood estimator* for parameters vector $\boldsymbol{\theta}$ is obtained as a solution of a maximization problem of likelihood function associated to the available observations. Hence, if we have an observations sample of T volume, extracted from the *n*–dimensional statistical population $\mathcal{P}$ and made of *n* – dimensional observations $\left\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},...,\mathbf{x}^{(T)}\right\}$ assumed to be *independent*, then, the likelihood function associated to this sample is defined under the form[1]:

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = \mathcal{L}^{*}_{\mathcal{P}}\left(\boldsymbol{\theta}\mid \mathbf{x}^{(1)},\mathbf{x}^{(2)},...,\mathbf{x}^{(T)}\right) = \prod_{t=1}^{T} f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\boldsymbol{\theta}\right),$$

i.e. it is a function depending on the (unknown) parameters vector $\boldsymbol{\theta}$, whose values are conditioned by the sample observations. The maximum likelihood estimator for parameters vector $\boldsymbol{\theta}$ will be:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}),$$

and the procedure of obtaining the sample is called the *maximum likelihood method.*

In order to simplify and bearing in mind the strictly monotonous increasing property of function $\ln(\cdot)$, the problem of estimating parameters by using the maximum likelihood method is frequently formulated under the name of *natural logarithm* of likelihood function. In these conditions, the function to be maximized is under the form:

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}) = \ln \mathcal{L}^{*}_{\mathcal{P}}\left(\boldsymbol{\theta}\mid \mathbf{x}^{(1)},\mathbf{x}^{(2)},...,\mathbf{x}^{(T)}\right) = \sum_{t=1}^{T} \ln f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\boldsymbol{\theta}\right).$$

We remark that, although the sample observations are assumed to be independent, this hypothesis enabling to express the likelihood function as a product of probability densities, these observations can not be assumed to be also identically distributed, as their origin lies in different statistical subpopulations.

Unlike the classical method of maximum likelihood, the method based on EM algorithm is an *iterative type procedure*, that can be used for certain problems of special type, problems including one or more particular random variables, which are random unobservable variables, called latent variables[2].

We shall further formulate an estimating problem of this type, by assuming that the statistical population $\mathcal{P}$ *is not homogeneous*, but structured under the form of a number

[1] A likelihood function defined in this manner is called *likelihood function with complete observations.*

[2] Even in the case of general estimating problem by means of maximum likelihood method, this problem can be conveniently converted, by including an artificial random variable, of latent type, that to enable to estimate parameters by using EM algorithm. This is a very interesting conversion, as EM algorithm provides a simpler, elegant and efficient estimating procedure, in comparison with the classical method of maximum likelihood, especially in cases when maximum conditions can not be easily ascertained.

of K statistical subpopulations $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$ for which, the probability density of random vector $\mathbf{X}_{\mathcal{P}_k}$ abstracting the characteristics of statistical subpopulation $\mathcal{P}_k$ is:

$$f_{\mathcal{P}_k}(\mathbf{x}; \boldsymbol{\theta}^{(k)}),\ \mathbf{x} \in \mathcal{P}_k,\ \boldsymbol{\theta}^{(k)} \in \boldsymbol{\Theta},\ k = 1,2,...,K.$$

Such a statistical population is under the form $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup ... \cup \mathcal{P}_K$ and the probability distribution associated to this statistical population is called *mixture* probability distribution. In this case, the probability density of a random vector $\mathbf{X}_{\mathcal{P}}$ has the nature of an unconditioned total probability density, and according to total probability formulae, the *(total) unconditioned probability density* of random vector $\mathbf{X}_{\mathcal{P}}$ is of the form:

$$f_{\mathcal{P}}(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_{\mathcal{P}_k}(\mathbf{x}; \boldsymbol{\theta}^k),\quad \mathbf{x} \in \mathcal{P},\ \boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K) \in \boldsymbol{\Pi},\ \boldsymbol{\theta} = \{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(K)}\} \in \boldsymbol{\Theta},$$

where $\boldsymbol{\pi}$ is the vector of mixture weights and $\boldsymbol{\Pi} = [0, 1]^K$ and $\boldsymbol{\Theta}$ are the sets of values for the two categories of parameters. As it can be noticed, the set of parameters of mixture probability distributions include both the weights $\pi_1, \pi_2, ..., \pi_K$ and the parameter vectors $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(K)}$ for the K probability distributions of the respective mixture.

The probability distribution of statistical population $\mathcal{P}$ is a mixture probability distribution of subpopulations and has a *K – modal* nature, that is to say that the probability density graph for this distribution has a number of K local maxima. In the figure below, the probability densities of two statistical subpopulations and the probability density corresponding to a mixture of these probability distributions are graphically illustrate.

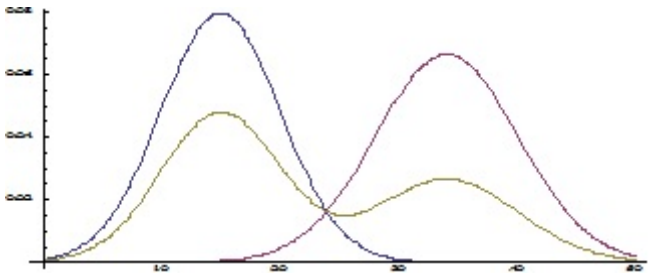

Figure 1: Graphs of probability densities for two statistical subpopulations and their mixture

Being defined the probability density of mixture probability distribution, the probability density of a random vector $\mathbf{x} \in \mathcal{P}_k$ can be considered as being a conditioned probability density, respectively, the density of vector $\mathbf{X}_{\mathcal{P}}$ mentioning the fact that $\mathbf{X}_{\mathcal{P}} = \mathbf{x} \in \mathcal{P}_k$, i.e.:

$$f_{\mathcal{P}}(\mathbf{x} \mid \mathbf{x} \in \mathcal{P}_k; \boldsymbol{\theta}) = f_{\mathcal{P}_k}(\mathbf{x}; \boldsymbol{\theta}^{(k)}),\ \mathbf{x} \in \mathcal{P}_k,\ \boldsymbol{\theta} = \boldsymbol{\theta}^{(k)} \in \boldsymbol{\Theta},\ k = 1,2,...,K.$$

As for the applications of the pattern recognition field, the representation of subpopulations can be made in two manners: by means of values of *discrete* type random variables, whose values usually represent subpopulation indexes, representation known

under the name *of natural coding,* respectively, by means of certain vectors of binary random variables, each vector being associated to a subpopulation, representation known under the name of *complete disjunctive coding.* In what follows, we shall specify the implications of the two types of representation for problems solved by using EM algorithm.

a. Natural coding case

As long as the randomly generated observations of statistical population $\mathcal{P}$ belong to any of subpopulations $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$, by a certain probability, we can mention a new probability distribution, defined besides the K subpopulations and abstracted by means of a new (scalar) random variable of discrete type $W_{\mathcal{P}}$ having a directly *latent unobservable* nature. We shall assume that this random variable has the probability distribution:

$$W_{\mathcal{P}} : \begin{pmatrix} w_1 & w_2 & ... & w_K \\ \pi_1 & \pi_2 & ... & \pi_K \end{pmatrix}_{\sum_{k=1}^{K} \pi_k = 1}$$

In Bayesian language, the probability distribution of random variable $W_{\mathcal{P}}$ is the *prior distribution* of statistical subpopulations $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$. As random variable $W_{\mathcal{P}}$ has a unobservable nature, then, when we refer to its values, we consider an hypothetical context. Nevertheless, EM algorithm enables to approximate the values of this variable.

We shall denote by $R_W = \{w_1, ..., w_K\}$ the set of values of random variable $W_{\mathcal{P}}$ and by $h_{W_{\mathcal{P}}}(w)$ the probability as random variable $W_{\mathcal{P}}$ to take value $w \in R_W$, i.e. $h_{W_{\mathcal{P}}}(w) = P(W_{\mathcal{P}}=w)$, what means that function $h_{W_{\mathcal{P}}}(\cdot)$ defines the *mass* probabilities for the values of random variable $W_{\mathcal{P}}$. For instance, for a particular value w_k of random variable $W_{\mathcal{P}}$, we shall have $h_{W_{\mathcal{P}}}(w_k) = \pi_k$.

As it can be noticed, the set of possible values of random variable $W_{\mathcal{P}}$ is perfectly known, this set representing indexes of classes. Though, random variable $W_{\mathcal{P}}$ has a latent, unobservable nature, as its values are not known (there are hidden) at the level of observations coming from statistical population $\mathcal{P}$, the origin subpopulation being completely unknown for any such observation. Anyway, the EM algorithm enables to estimate the values of this variable for each sample observation.

If we know how its probability distribution is specified, the random variable $W_{\mathcal{P}}$ has the nature of an *identifier* of statistical subpopulation, reason for which, it can be called *subpopulation variable.* Thus, for any observation $\mathbf{x}^{(t)}$ coming from statistical population $\mathcal{P}$, to know the value which random variable $W_{\mathcal{P}}$ could take is equivalent to knowing which is the subpopulation, the respective observation comes from.

If we know both the weights of mixture probability distribution and the probability distributions of statistical subpopulations $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$, the probability distribution of random vector $\mathbf{X}_{\mathcal{P}}$ has the nature of a *marginal distribution* and is *completely* known.

The random variable $W_{\mathcal{P}}$ can be regarded as defining an *additional size* of statistical population $\mathcal{P}$ so that this variable appears as a *(n+1)*–dimensional population. We shall denote by $\mathbf{Z}_{\mathcal{P}}$ the random vector of the *n+1* characteristics of statistical population $\mathcal{P}$, respectively, $\mathbf{Z}_{\mathcal{P}} = (\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}}) = (X_1, X_2, ..., X_n, W_{\mathcal{P}})^t$. As random vector $\mathbf{X}_{\mathcal{P}} = (X_1, X_2, ..., X_n)^t$ contains information about the statistical subpopulations only under an implicit, latent form, by means of the values taken by random variables integrating it, this random vector can be considered as being the generic representation of an *incomplete observation.* Unlike it, the random vector $\mathbf{Z}_{\mathcal{P}}$ containing under an explicit form, also the (estimated) information referring to the origin subpopulation of random vector $\square_{\mathcal{P}}$, it can be considered as being a generic representation of a *complete information*. That is why, data referring to random vector $\mathbf{X}_{\mathcal{P}}$ are called *incomplete data*, while those data referring to vector $\mathbf{Z}_{\mathcal{P}} = (\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}})$ are called *complete data.* We shall assume that the probability distribution of *(n+1)*–dimensional random vector $\mathbf{Z}_{\mathcal{P}}$ is described by the probability density:

$$f_{\mathbf{Z}_{\mathcal{P}}}(\mathbf{z}; \boldsymbol{\pi}, \boldsymbol{\theta}) = f_{(\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}})}(\mathbf{x}, w; \boldsymbol{\pi}, \boldsymbol{\theta}), \quad \mathbf{x} \in \mathcal{P}, \ w \in R_W,$$

where $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ are the parameter vectors of this probability distribution.

In the context of *(n+1)*–dimensional distribution of random vector $\mathbf{Z}_{\mathcal{P}} = (\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}})$ there are of interest, from the perspective of EM algorithm logic, two types of conditioned distributions: probability distribution of random vector $\mathbf{X}_{\mathcal{P}}$ conditioned by the values of random variable $W_{\mathcal{P}}$, respectively, the probability distribution of random variable $W_{\mathcal{P}}$ conditioned by the values of random vector $\mathbf{X}_{\mathcal{P}}$. We shall further deduce the general form of probability densities for these two types of conditioned probability distribution.

For *(n+1)*-dimensional distribution of random vector $\mathbf{Z}_{\mathcal{P}} = (\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}})$, the K possible values of random variable $W_{\mathcal{P}}$ can be considered as representing *conditioning values* or *levels* for random vector $\mathbf{X}_{\mathcal{P}}$. If we use the probabilistic conditioning concepts, the probability density of random vector $\mathbf{Z}_{\mathcal{P}} = (\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}})$ can be written as follows:

$$f_{(\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}})}(\mathbf{x}, w; \boldsymbol{\pi}, \boldsymbol{\theta}) = f_{\mathbf{X}_{\mathcal{P}} | W_{\mathcal{P}} = w}(\mathbf{x} | w; \boldsymbol{\pi}, \boldsymbol{\theta}) \, h_{W_{\mathcal{P}}}(w), \quad \forall \mathbf{x} \in \mathcal{P}, \ \forall w \in R_W.$$

On the other side, if we know the fact that random variable $W_{\mathcal{P}}$ has taken a certain particular value w_k *the probability density of random vector* $\mathbf{X}_{\mathcal{P}}$, being given the value w_k of random variable $W_{\mathcal{P}}$ is exactly the probability density of random vector $\mathbf{X}_{\mathcal{P}}$ at the level of the *k-th* statistical subpopulation $\mathcal{P}_k$, respectively:

$$f_{\mathbf{X}_{\mathcal{P}} | W_{\mathcal{P}} = w_k}(\mathbf{x} | w_k; \boldsymbol{\theta}) = f_{\mathcal{P}_k}(\mathbf{x}; \boldsymbol{\theta}^{(k)}), \quad \mathbf{x} \in \mathcal{P}_k, \ \boldsymbol{\theta}^{(k)} \in \boldsymbol{\Theta},$$

as via value w_k of random variable $W_{\mathcal{P}}$, the classification of random vector $\mathbf{X}_{\mathcal{P}}$ to subpopulation $\mathcal{P}_k$ is univocally identified. As can be noticed, the conditioned probability density of random vector $x_{\mathcal{P}}$, being given the fact that $W_{\mathcal{P}} = w_k$, does not depend only on the

parameters of probability distribution of subpopulation $\mathcal{P}_k$. Moreover, the density of *marginal* (total, unconditioned) probability of random vector $\mathbf{X}_{\mathcal{P}}$ can be obtained by summing, with respect to the possible values of probability distribution of random variable $W_{\mathcal{P}}$, i.e.:

$$f_{\mathcal{P}}(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{k=1}^{K} f_{(\mathbf{X}_{\mathcal{P}},W_{\mathcal{P}})}(\mathbf{x},w_k;\boldsymbol{\pi},\boldsymbol{\theta}); \quad \forall \mathbf{x}\in\mathcal{P},$$

whence, considering the fact that $f_{(\mathbf{X}_{\mathcal{P}},W_{\mathcal{P}})}(\mathbf{x},w;\boldsymbol{\pi},\boldsymbol{\theta}) = f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}=w}(\mathbf{x}\,|\,w;\boldsymbol{\pi},\boldsymbol{\theta})\,h_{W_{\mathcal{P}}}(w)$, it follows that the (total, unconditioned) marginal probability density of random vector $\mathbf{X}_{\mathcal{P}}$ is:

$$f_{\mathcal{P}}(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{k=1}^{K} f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}=w_k}(\mathbf{x}\,|\,w_k;\boldsymbol{\theta}^{(k)})h_{W_{\mathcal{P}}}(w_k) = \sum_{k=1}^{K} f_{\mathcal{P}_k}(\mathbf{x};\boldsymbol{\theta}^{(k)})h_{W_{\mathcal{P}}}(w_k).$$

This result outlines the fact that, really, the (unconditioned) probability density of random vector $x_{\mathcal{P}}$, is a convex combination of probability densities of the K probability distributions of statistical subpopulations $\mathcal{P}_1,\mathcal{P}_2,...,\mathcal{P}_K$, combination with coefficients represented by probabilities associated to the values of latent random variable $W_{\mathcal{P}}$, i.e.:

$$f_{\mathcal{P}}(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_{\mathcal{P}_k}(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta}^{(k)}).$$

b. Complet disjunctive coding case

In order to identify the K subpopulations, a binary random vector is used, so that, instead of latent random variable $W_{\mathcal{P}}$, used in case of natural coding, we have a latent random vector, respectively, the random vector $\mathbf{W}_{\mathcal{P}} = (W_{\mathcal{P}_1}, W_{\mathcal{P}_2}, ... W_{\mathcal{P}_K})^t$, whose (*hypothetical*) realization will be denoted by $\mathbf{w} = (w_1, w_2, ... w_K)^t$. The random variables of random vector $\mathbf{W}_{\mathcal{P}}$ are *binary* variables which verify the following conditions: $W_{\mathcal{P}_k} \in \{0,1\}; \quad \sum_{k=1}^{K} W_{\mathcal{P}_k} = 1$.

Unlike random vector $X_{\mathcal{P}}$, that is the vector of *characteristics (attributes, sizes)* of total population $\mathcal{P}$ with an observable nature, vector $\mathbf{W}_{\mathcal{P}}$ is of an *unobservable* nature, for this reason, being called, *hidden* or *latent* vector.

According to the way of its definition, random vector $\mathbf{W}_{\mathcal{P}}$ has also the nature of an *identifier of subpopulations,* constructed on the basis of a *complete disjunctive coding of classes*, and its hypothetical realizations are represented by the finite set of K–dimensional unit vectors $\mathbf{B} = \{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ..., \mathbf{w}^{(K)}\}$, with the following form:

$$\mathbf{w}^{(1)} = \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \\ \dots \\ 0 \end{pmatrix}_{K\times 1}, \quad \mathbf{w}^{(2)} = \begin{pmatrix} 0 \\ 1 \\ \dots \\ 0 \\ \dots \\ 0 \end{pmatrix}_{K\times 1}, \dots, \quad \mathbf{w}^{(k)} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 1 \\ \dots \\ 0 \end{pmatrix}_{K\times 1}, \dots, \quad \mathbf{w}^{(K)} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ \dots \\ 1 \end{pmatrix}_{K\times 1}$$

i.e. there is a total number of K such distinct vectors of realizations. We shall denote by π_k the probability of random vector $\mathbf{W}_{\mathcal{P}}$ to take value $\mathbf{w}^{(k)}$, namely, $\pi_k = P\left(W_{\mathcal{P}}=\mathbf{w}^{(k)}\right)$. This means that the *k-th* random variable of random vector $\mathbf{X}_{\mathcal{P}}$, i.e. random variable $W_{\mathcal{P}_k}$, takes value 1, respectively $\pi_k = P\left(W_{\mathcal{P}_k}=1\right)$. As the random variables of random vector $\mathbf{W}_{\mathcal{P}}$ are binary, and their values are conversely exclusive, i.e., if one of these variables takes value 1, all the other variables take value 0, probability as random vector $\mathbf{W}_{\mathcal{P}}$ to take a particular value **w** is:

$$P\left(\mathbf{W}_{\mathcal{P}}=\mathbf{w}\right) = h_{\mathbf{W}_{\mathcal{P}}}\left(\mathbf{w}\right) = h_{\mathbf{W}_{\mathcal{P}}}\left(w_1, w_2, \dots, w_2\right) = \prod_{k=1}^{K} \pi_k^{w_{\mathcal{P}_k}}.$$

In pattern recognition theory, to determine the classification of a realizations vector **x,** that is, in fact, a pattern to a certain class $\mathcal{P}_k$, assume to define a membership function under the form:

$$d : \mathcal{P} \rightarrow \mathbf{B} = \{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \dots, \mathbf{w}^{(K)}\}; \quad d(\mathbf{x}) = \mathbf{w}.$$

We shall now consider the *(n+K)*–dimensional random vector $\mathbf{Z}_{\mathcal{P}} = (\mathbf{X}_{\mathcal{P}}, \mathbf{W}_{\mathcal{P}})^t$ and we shall denote by $f_{(\mathbf{X}_{\mathcal{P}}, \mathbf{W}_{\mathcal{P}})}(\cdot,\cdot)$ its probability density and by $f_{\mathbf{X}_{\mathcal{P}}|\mathbf{W}_{\mathcal{P}}}(\cdot,\cdot)$ the conditioned probability density of (sub)vector $\mathbf{X}_{\mathcal{P}}$, being given that $\mathbf{W}_{\mathcal{P}} = \mathbf{w}$. We shall yield:

$$f_{(\mathbf{X}_{\mathcal{P}}, \mathbf{W}_{\mathcal{P}})}(\mathbf{x}, \mathbf{w}) = f_{\mathbf{X}_{\mathcal{P}}|\mathbf{W}_{\mathcal{P}}}(\mathbf{x}|\mathbf{w}) h_{\mathbf{W}_{\mathcal{P}}}(\mathbf{w}); \quad \forall \mathbf{x} \in \mathcal{P}, \ \forall \mathbf{w} \in \mathbf{B}.$$

For instance, in case when $\mathbf{W}_{\mathcal{P}} = \mathbf{w}^{(k)}$, case equivalent to the situation where random variable $W_{\mathcal{P}_k}$ takes value 1, the conditioned density $f_{\mathbf{X}_{\mathcal{P}}|\mathbf{W}_{\mathcal{P}}}\left(\mathbf{x}|\mathbf{w}^{(k)}\right)$ represents exactly the probability density of subpopulation $\mathcal{P}_k$, respectively:

$$f_{\mathbf{X}_{\mathcal{P}}|\mathbf{W}_{\mathcal{P}}}\left(\mathbf{x}|\mathbf{w}^{(k)}\right) = f_{\mathcal{P}_k}(\mathbf{x}),$$

case where the conditioned probability density $h_{\mathbf{X}_{\mathcal{P}}|\mathbf{W}_{\mathcal{P}}}(\cdot,\cdot)$ can be written under the particular form:

$$f_{\mathbf{X}_{\mathcal{P}}|\mathbf{W}_{\mathcal{P}}}\left(\mathbf{x}|\mathbf{w}^{(k)}\right) = \prod_{k=1}^{K} f_{\mathcal{P}_k}(\mathbf{x})^{w_{\mathcal{P}_k}},$$

and the (*marginal) unconditioned* probability density of random (sub)vector $\mathbf{X}_{\mathcal{P}}$, can be obtained via *summing* the probability density $f_{(\mathbf{X}_{\mathcal{P}}, \mathbf{W}_{\mathcal{P}})}(\mathbf{x}, \mathbf{w})$ according to the possible values of random vector $\mathbf{W}_{\mathcal{P}}$, respectively:

$$f_{\mathcal{P}}(\mathbf{x}) = \sum_{k=1}^{K} f_{(\mathbf{X}_{\mathcal{P}}, \mathbf{W}_{\mathcal{P}})}\left(\mathbf{x}, \mathbf{w}^{(k)}\right) = \sum_{k=1}^{K} f_{\mathbf{X}_{\mathcal{P}}|\mathbf{W}_{\mathcal{P}}}\left(\mathbf{x} \mid \mathbf{w}^{(k)}\right) h_{\mathbf{W}_{\mathcal{P}}}\left(\mathbf{w}^{(k)}\right).$$

It follows that the *(unconditioned) marginal* density probability of random vector $\mathbf{X}_{\mathcal{P}}$ has the nature of a mixture of K probability distribution, which are the probability distributions of subpopulations $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$:

$$f_{\mathcal{P}}(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_{\mathcal{P}_k}(\mathbf{x}).$$

As can be noticed, we have found a form to express the *marginal (unconditioned)* density probability of random vector $X_{\mathcal{P}}$, identical to that one obtained in case of natural coding. Further on, we shall use the first method of coding the subpopulations, based on the use of latent random variable $W_{\mathcal{P}}$.

Also with respect to the *(n+1)*–dimensional distribution of vector $\mathbf{Z}_{\mathcal{P}} = (\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}})$, we can define the *conditioned probability distribution of random variable* $W_{\mathcal{P}}$, being given $\mathbf{X}_{\mathcal{P}}=\mathbf{x} \in \mathcal{P}$. We shall denote by $f_{W|\mathbf{X}}(\cdot;\cdot)$ the probability density of this distribution, i.e.:

$$P\left(W = w \mid \mathbf{X}_{\mathcal{P}} = \mathbf{x}\right) = f_{W|\mathbf{X}}(w \mid \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}), \quad \mathbf{x} \in \mathcal{P}.$$

On the basis of this conditioned density, the probability density of random vector $\mathbf{Z} = (\mathbf{X}_{\mathcal{P}}, W_{\mathcal{P}})$ can be written:

$$f_{(\mathbf{X}_{\mathcal{P}}, W)}\left(\mathbf{x}, w; \boldsymbol{\pi}, \boldsymbol{\theta}\right) = f_{W|\mathbf{X}_{\mathcal{P}}}\left(w \mid \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}\right) f_{\mathcal{P}}\left(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}\right), \quad \forall \mathbf{x} \in \mathcal{P}, \ \forall w \in R_W,$$

whence, it follow that:

$$f_{W|\mathbf{X}_{\mathcal{P}}=\mathbf{x}}\left(w \mid \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}\right) = \frac{f_{(\mathbf{X},W)}\left(\mathbf{x}, w; \boldsymbol{\pi}, \boldsymbol{\theta}\right)}{f_{\mathcal{P}}\left(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}\right)}, \quad \forall \mathbf{x} \in \mathcal{P}, \ \forall w \in R_W,$$

relation that is a consequence of Bayes formulae. In fact, density $f_{W|\mathbf{X}_{\mathcal{P}}}\left(w \mid \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}\right)$ has the nature of a posterior probability density. This conditioned probability density can be also used to express the (total, unconditioned) marginal probability density of random vector $\mathbf{X}_{\mathcal{P}}$, respectively:

$$f_{\mathcal{P}}\left(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}\right) = \sum_{k=1}^{K} f_{(\mathbf{X},W)}\left(\mathbf{x}, w_k; \boldsymbol{\pi}, \boldsymbol{\theta}\right) = \sum_{k=1}^{K} f_{W|\mathbf{X}=\mathbf{x}}\left(w_k \mid \mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}^{(k)}\right) f_{\mathcal{P}_k}\left(\mathbf{x}; \boldsymbol{\theta}^{(k)}\right).$$

As we can notice, the unconditioned probability density of random vector $\mathbf{X}_{\mathcal{P}}$ is a function depending both on the parameters vector $\boldsymbol{\pi}$, and on the parameters vector of the K probability distributions of subpopulations $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(K)}$, with the condition that all these parameters to be estimated.

In order to construct the likelihood function, we shall assume that we have the *n*–dimensional sample of incomplete observations $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}\right\}$, extracted from statistical population $\mathcal{P}$, for which the selection model is represented by the list of random (independents) vectors $\left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(T)}\right\}$. The likelihood function associated to this sample is defined as:

$$\mathcal{L}_{,\mathcal{P}}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \ln \mathcal{L}^{*}_{,\mathcal{P}}(\boldsymbol{\pi}, \boldsymbol{\theta} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}) = \sum_{t=1}^{T} \ln f_{,\mathcal{P}}(\mathbf{x}^{(t)}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{t=1}^{T} \ln \left[\sum_{k=1}^{K} \pi_k f_{,\mathcal{P}_k}(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}^{(k)}) \right],$$

and the maximum likelihood estimators for parameter vectors $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ will be:

$$\{\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}\} = \arg \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}_{,\mathcal{P}}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \arg \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{t=1}^{T} \ln \left[\sum_{k=1}^{K} \pi_k f_{,\mathcal{P}_k}(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}^{(k)}) \right].$$

As a result of the fact that the likelihood function is expressed in terms of *logarithm of a sum*, to analytically obtain a solution of this problem, is rather difficult, but not impossible, so, the single possibility of solving is the numerical one, by using an iterative procedure. Moreover, the maximization of likelihood function in its previous form, rises also other difficulties. A first difficulty is related to the possibility of the existence of certain singularities for certain probability distributions, as *centered* distributions. Thus, if the estimation for expected values vector of a probability distribution of subpopulations comes *very close* to one of the T observations, i.e. if $\mathbf{x}^{(t)} \approx \boldsymbol{\mu}^{(k)}$ then, the variance of the respective probability distribution is very close to 0, and the value of subpopulation probability density comes close to infinite, which means that the value of likelihood function comes close to infinite. Another difficulty is related to the fact that for a mixture of K probability distributions, the problem of maximizing the likelihood function has a number of $K!$ solutions. Fortunately, these $K!$ solutions are perfectly *equivalent*, they being different only from the point of view of the order in which the estimated values on subpopulations are shown.

The most elegant modality of numerically solving the previous maximization problem is provided by EM algorithm, which is a special iterative numerical procedure with a high efficiency. Further on, we shall describe in detail the logic of this algorithm.

III. Inducing EM algorithm

The main idea of EM algorithm is that of maximizing the likelihood function $\mathcal{L}_{,\mathcal{P}}(\boldsymbol{\pi}, \boldsymbol{\theta})$, in a simplified and indirect manner, based on constructing a special auxiliary function, easily to be maximized. In fact, this manner consists in expressing the likelihood function in terms of posterior probabilities of available observations and in determining the weights of mixture probability distributions on the basis of approximations calculated for posterior probabilities. The next sentence synthesizes the fundamental idea of EM algorithm and can be considered as proving the entire essence of EM algorithm.

Propozition: Estimator vectors $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\theta}}$ which are solutions of maximum problem $\max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \tilde{Q}_{,\mathcal{P}}(\boldsymbol{\pi}, \boldsymbol{\theta})$ are also solutions of maximum problem $\max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}_{,\mathcal{P}}(\boldsymbol{\pi}, \boldsymbol{\theta})$, where:

$$\tilde{Q}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{k=1}^{K} f_{W|\mathbf{X},\mathcal{P}}(w_k \mid \mathbf{x}^{(t)}; \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q) \ln \left[h_{W,\mathcal{P}}(w_k; \boldsymbol{\pi}, \boldsymbol{\theta}) \right].$$

Proof. We shall assume that after the *q-th* iteration of EM algorithm, the values of parameter estimations are represented by vectors $\hat{\boldsymbol{\pi}}_q$ and $\hat{\boldsymbol{\theta}}_q$, following to find, at next iteration, the estimates $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ so that, these ones to determine a *maximum increment* of the likelihood function logarithm, with respect to iteration *q*. Under this assumption, $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ have the nature of certain random variables whose values depending on known values $\hat{\boldsymbol{\pi}}_q$ and $\hat{\boldsymbol{\theta}}_q$, so that we can consider them as random variables conditioned by form $(\boldsymbol{\pi},\boldsymbol{\theta})\,|\,(\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q)$. If we shall denote increase the likelihood function by:

$$\Delta\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) = \mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) - \mathcal{L}_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right),$$

we shall yield:

$$\begin{aligned}\Delta\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) &= \ln\mathcal{L}^*_{\mathcal{P}}\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(T)}\right) - \ln\mathcal{L}^*_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q,\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(T)}\right)\\ &= \sum_{t=1}^{T}\ln f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\boldsymbol{\pi},\boldsymbol{\theta}\right) - \sum_{t=1}^{T}\ln f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\\ &= \sum_{t=1}^{T}\left[\ln f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\boldsymbol{\pi},\boldsymbol{\theta}\right) - \ln f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\right] = \sum_{t=1}^{T}\ln\frac{f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\boldsymbol{\pi},\boldsymbol{\theta}\right)}{f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)}.\end{aligned}$$

Now expressing the marginal density $f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\boldsymbol{\pi},\boldsymbol{\theta}\right)$ as the density of a mixture probability distribution, we shall get:

$$\Delta\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) = \sum_{t=1}^{T}\ln\left(\frac{\sum_{k=1}^{K} f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)}{f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)}\right),$$

whence, on the basis of a simple trick, we have:

$$\Delta\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) = \sum_{t=1}^{T}\ln\left(\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\frac{f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\,h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)}{f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)}\right).$$

According to Bayes formulae:

$$f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) = \frac{f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)}{f_{\mathcal{P}}\left(\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)},$$

we shall have:

$$\Delta_q\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) = \sum_{t=1}^{T}\ln\left(\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\frac{f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\,h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)}{f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)}\right).$$

On the other side, as:

$$f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\geq 0,\ \forall k=1,2,\ldots,K \quad \text{and} \quad \sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) = 1,$$

the sum representing the logarithm argument can be considered as a convex combination having these coefficients, so that, according to Jensen inequality, we shall have:

$$\Delta_q\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) \geq \sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\frac{f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_h;\boldsymbol{\pi},\boldsymbol{\theta}\right)h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)}{f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)}$$

$$\geq \sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]$$

$$-\sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\right].$$

As $f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\cdot\right)$ is the probability density of subpopulation $\mathcal{P}_k$, it does not depend on parameters vector $\boldsymbol{\pi}$, the previous inequality can be rewritten under the form:

$$\Delta_q\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) \geq \sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\theta\right)h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]$$

$$-\sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\hat{\boldsymbol{\theta}}_q\right)h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\right]$$

$$\geq \sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]$$

$$-\sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\right].$$

We shall denote:

$$Q\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) = \sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]$$

$$-\sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\right].$$

The problem is to determine vectors $\hat{\boldsymbol{\pi}}_{q+1}$ and $\hat{\boldsymbol{\theta}}_{q+1}$ which provide *the largest increment* of the value of likelihood function logarithm at next iteration. As:

$$\sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\right],$$

does not depend on $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\pi}}_{q+1}$ and $\hat{\boldsymbol{\theta}}_{q+1}$ are solutions of the next maximization problem:

$$\left(\hat{\boldsymbol{\pi}}_{q+1},\hat{\boldsymbol{\theta}}_{q+1}\right) = \arg\max_{\boldsymbol{\pi},\boldsymbol{\theta}}\left\{\tilde{Q}(\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]\right\}. \blacksquare$$

If we shall consider the results of previous proof, then the log likelihood function verifies the following condition:

$$\mathcal{L}_{\mathcal{P}}(\boldsymbol{\pi},\boldsymbol{\theta}) \geq \mathcal{L}_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) + Q\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right).$$

If we shall denote:

$$L_{\mathcal{P}}\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) = \mathcal{L}_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right) + Q\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right),$$

we shall have $\mathcal{L}_{\mathcal{P}}(\boldsymbol{\pi},\boldsymbol{\theta}) \geq L_{\mathcal{P}}\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)$. For $(\boldsymbol{\pi},\boldsymbol{\theta}) = \left(\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_q\right)$, we have:

$$L_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q \mid \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) = \mathcal{L}_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) + Q\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q \mid \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right)$$

$$= \mathcal{L}_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) + \sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k \mid \mathbf{x}^{(t)}; \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) \ln\left[\frac{h_{W_{\mathcal{P}}}\left(w_k; \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right)}{h_{W_{\mathcal{P}}}\left(w_k; \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right)}\right]$$

$$= \mathcal{L}_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right),$$

i.e. at q iteration, the value of log likelihood function is identical to the value of function $Q(\cdot)$. At the same time, the lower bound of the original value of likelihood function can be calculated as follows:

$$\mathcal{L}^{*}_{\mathcal{P}}\left(\boldsymbol{\pi}, \boldsymbol{\theta} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}\right) = e^{\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\pi}, \boldsymbol{\theta} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}\right)} \geq e^{\mathcal{L}_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}\right) + Q\left(\boldsymbol{\pi}, \boldsymbol{\theta} \mid \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right)}.$$

It follows that at $q+1$ iteration, the estimators of maximum likelihood for parameters of mixture probability distributions can be determined as solutions of the next extreme problem:

$$\left(\hat{\boldsymbol{\pi}}_{q+1}, \hat{\boldsymbol{\theta}}_{q+1}\right) = \arg \max_{\boldsymbol{\pi}, \boldsymbol{\theta}} \left\{\sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k \mid \mathbf{x}^{(t)}; \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) \ln\left[h_{W_{\mathcal{P}}}\left(w_k; \boldsymbol{\pi}, \boldsymbol{\theta}\right)\right]\right\}.$$

As can be noticed, the second sum of the previous relation defines the *expected value* of $\ln[\cdot]$, this expectation being made according to posterior probability distribution of random variable *W*, such as this distribution is approximated for each observation. This means that:

$$\sum_{t=1}^{T}\sum_{k=1}^{K} f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k \mid \mathbf{x}^{(t)}; \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) \ln\left[h_{W_{\mathcal{P}}}\left(w_k; \boldsymbol{\pi}, \boldsymbol{\theta}\right)\right] = E_{f_{W|\mathbf{X}_{\mathcal{P}}}}\left[\ln\left(h_{W_{\mathcal{P}}}\left(w_k; \boldsymbol{\pi}, \boldsymbol{\theta}\right)\right)\right].$$

The previous relation suggests the following computational *strategy,* specific to each iteration of EM algorithm: (a) in **E** stage of each q iteration, the **expected values** of prior probability $h_{W_{\mathcal{P}}}\left(w_k; \boldsymbol{\pi}, \boldsymbol{\theta}\right)$ for all observations $\mathbf{x}^{(t)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}$ are calculated; (b) these expected values are used to estimate weights vector $\hat{\boldsymbol{\pi}}_q$ of mixture probability distributions; (c) weights vector $\hat{\boldsymbol{\pi}}_q$ is used in **M** stage for maximizing ***the sum of expected values*** of all observations, with a view to obtaining approximation $\hat{\boldsymbol{\theta}}_q$ for parameters vector $\boldsymbol{\theta}$.

On the other hand, for each observation $\mathbf{x}^{(t)}$, the second sum of the previous relation defines the *(negative) prior distribution cross entropy* of random variable W or the (negative) prior distribution cross entropy of subpopulations $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K$, entropy evaluated with respect to the posterior distribution probabilities of variable W.

Consequently, the problem of maximizing the likelihood function is equivalent to the problem of *minimizing* the sum of prior distribution cross entropies of latent variable W for all available observations, such as these entropies are evaluated at the level of each iteration, respectively:

$$\left(\hat{\boldsymbol{\pi}}_{q+1}, \hat{\boldsymbol{\theta}}_{q+1}\right) = \arg \max_{\boldsymbol{\pi}, \boldsymbol{\theta}} \left\{-\sum_{t=1}^{T} H_{f_{W|\mathbf{X}_{\mathcal{P}}=\mathbf{x}^{(t)}}}\left[h_{W_{\mathcal{P}}}\left(w_k; \boldsymbol{\pi}, \boldsymbol{\theta}\right)\right]\right\} = \arg \min_{\boldsymbol{\pi}, \boldsymbol{\theta}} \left\{\sum_{t=1}^{T} H_{f_{W|\mathbf{X}_{\mathcal{P}}=\mathbf{x}^{(t)}}}\left[h_{W_{\mathcal{P}}}\left(w_k; \boldsymbol{\pi}, \boldsymbol{\theta}\right)\right]\right\}.$$

Within each *q* iteration, EM algorithm firstly estimates approximation $\hat{\boldsymbol{\pi}}_q$ of weight vector $\boldsymbol{\pi}$, using approximation $\hat{\boldsymbol{\theta}}_{q-1}$ from the previous *q-1* iteration, where after it uses vector $\hat{\boldsymbol{\pi}}_q$ in order to calculate the current approximation $\hat{\boldsymbol{\theta}}_q$. Under these conditions, when vector $\hat{\boldsymbol{\pi}}_q$ is calculated, the conditioned probability density $f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}(\cdot;\cdot)$ does not depend anymore on $\boldsymbol{\theta}$, it being exactly the probability density of subpopulation $\mathcal{P}_k$, with parameters vector $\hat{\boldsymbol{\theta}}_{q-1}$, namely:

$$f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}\left(\mathbf{x}^{(t)}\,|\,w_k;\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}}_q\right)=f_{\mathcal{P}_k}\left(\mathbf{x}^{(t)},\hat{\boldsymbol{\theta}}_{q-1}\right),$$

and the above inequality will become:

$$\begin{aligned}\Delta_q\left(\boldsymbol{\pi},\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)&\geq\sum_{t=1}^{T}\sum_{k=1}^{K}f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]\\&\quad+\sum_{t=1}^{T}\sum_{k=1}^{K}f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)\ln\left[\frac{f_{\mathcal{P}_k}\left(\mathbf{x}^{(t)},\hat{\boldsymbol{\theta}}_{q-1}\right)}{h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)f_{\mathcal{P}_k}\left(\mathbf{x}^{(t)},\hat{\boldsymbol{\theta}}_{q-1}\right)}\right]\\&\geq\sum_{t=1}^{T}\sum_{k=1}^{K}f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]\\&\quad+\sum_{t=1}^{T}\sum_{k=1}^{K}f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)\ln\left[\frac{1}{h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)}\right]\\&\geq\sum_{t=1}^{T}\sum_{k=1}^{K}f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]\\&\quad-\sum_{t=1}^{T}\sum_{k=1}^{K}f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)\right].\end{aligned}$$

In this way, the maximization problem of log likelihood function is more simplified, as follows:

$$\left(\hat{\boldsymbol{\pi}}_{q+1},\hat{\boldsymbol{\theta}}_{q+1}\right)=\arg\max_{\boldsymbol{\pi},\boldsymbol{\theta}}\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}f_{W|\mathbf{X}_{\mathcal{P}}}\left(w_k\,|\,\mathbf{x}^{(t)};\hat{\boldsymbol{\pi}}_q,\hat{\boldsymbol{\theta}}_{q-1}\right)\ln\left[h_{W_{\mathcal{P}}}\left(w_k;\boldsymbol{\pi},\boldsymbol{\theta}\right)\right]\right\}.$$

In the box below, stages and iterations of EM algorithm, corresponding to a certain general case, where probability distributions of subpopulations are certain distributions, are described.

Expected Maximization Algorithm (EM)

I. Initialization: set $q = 1$ and randomly generates start values for parameters vectors, respectively $\boldsymbol{\pi}_0$ and $\boldsymbol{\theta}_0$;

II. q iteration:

• **E stage**: • based on estimations $\hat{\boldsymbol{\pi}}_{q-1}$ and $\hat{\boldsymbol{\theta}}_{q-1}$ from previous iteration, compute (posterior probabilities and mixture weights):

$$\bullet\, w_{tk;q} = \frac{f_{W|\mathbf{X}}\left(w_k \mid \mathbf{x}^{(t)}; \hat{\boldsymbol{\pi}}_{q-1}, \hat{\boldsymbol{\theta}}^{(k)}_{q-1}\right) f_{\mathcal{P}_k}\left(\mathbf{x}^{(t)}; \boldsymbol{\theta}^{(k)}_{q-1}\right)}{\sum_{j=1}^{K} f_{W|\mathbf{X}}\left(w_j \mid \mathbf{x}^{(t)}; \hat{\boldsymbol{\pi}}_{q-1}, \hat{\boldsymbol{\theta}}^{(j)}_{q-1}\right) f_{\mathcal{P}_k}\left(\mathbf{x}^{(t)}; \boldsymbol{\theta}^{(j)}_{q-1}\right)} = \frac{f_{W|\mathbf{X}}\left(w_k \mid \mathbf{x}^{(t)}; \hat{\boldsymbol{\pi}}_{q-1}, \hat{\boldsymbol{\theta}}^{(k)}_{q-1}\right)}{\sum_{j=1}^{K} f_{W|\mathbf{X}}\left(w_j \mid \mathbf{x}^{(t)}; \hat{\boldsymbol{\pi}}_{q-1}, \hat{\boldsymbol{\theta}}^{(j)}_{q-1}\right)};\quad t=1,2,...,T;\ k=1,2,...,K;$$

$$\bullet\, \hat{\pi}_{k;q} = \sum_{t=1}^{T} w_{tk;q} \Big/ \sum_{t=1}^{T}\sum_{j=1}^{K} w_{tj;q},\quad k=1,2,...,K;$$

• **M stage**: • using the weights vector $\hat{\boldsymbol{\pi}}_{k;q}$, calculated in **E** stage, compute the maximum likelihood estimate for parameters vector $\boldsymbol{\theta}$, as solution of maxim problem:

$$\bullet\, \hat{\boldsymbol{\theta}}_q = \arg\max_{\boldsymbol{\theta}} \left\{ \ln \mathcal{L}^*_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q, \boldsymbol{\theta} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}\right) = \sum_{t=1}^{T} \ln\left[\sum_{k=1}^{K} \hat{\pi}_{k;q} f_{\mathcal{P}_k}\left(\mathbf{x}; \hat{\boldsymbol{\pi}}_q, \boldsymbol{\theta}^{(k)}\right)\right]\right\};$$

III. Stop criterion:

• evaluation likelihood function:

• $\mathcal{L}^*_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}\right)$;

• if:

• $\left|\mathcal{L}^*_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}\right) - \mathcal{L}^*_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_{q-1}, \hat{\boldsymbol{\theta}}_{q-1} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}\right)\right| \geq \varepsilon$,

then: set $q = q + 1$ and ***goto II***;

otherwise: ***Stop***.

From the view point of EM algorithm convergence, numerical applications and tests have proved a relatively high convergence speed of this algorithm. Really, from the perspective of a certain q iteration, within the EM algorithm, approximations $\hat{\boldsymbol{\pi}}_{q+1}$ and $\hat{\boldsymbol{\theta}}_{q+1}$ for parameter vectors $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are determined at $q+1$ iteration, such as to lead to a maximum increment of likelihood function value, respectively:

$$\Delta_q\left(\boldsymbol{\pi}, \boldsymbol{\theta} \mid \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) = \mathcal{L}_{\mathcal{P}}(\boldsymbol{\pi}, \boldsymbol{\theta}) - \mathcal{L}_{\mathcal{P}}\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right),$$

whence, as $\Delta_q\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q \mid \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) = 0$, it follows that:

$$\Delta_{q+1}\left(\hat{\boldsymbol{\pi}}_{q+1}, \hat{\boldsymbol{\theta}}_{q+1} \mid \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) \geq \Delta_q\left(\hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q \mid \hat{\boldsymbol{\pi}}_q, \hat{\boldsymbol{\theta}}_q\right) = 0,$$

i.e. the likelihood function is non-decreasing along iterations.

Unfortunately, as happens in case of most nonlinear optimization algorithms, the final optimizations for parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ can have the nature of some local minima or even worse, the nature of some saddle type points. However, in most applications, EM algorithm proved a very high robustness.

IV. Using EM Algorithm to Unsupervised Pattern Recognition

Assuming that the number of classes (clusters) is known, the logic according to which, the EM algorithm iterates, can be successfully " borrowed" , in order to develop an *unsupervised statistical learning algorithm* or a *cluster analysis*. Such an algorithm has the nature of *partitioning algorithms*, as *K – means algorithm* is and it is based on iterative approximation of posterior probabilities of available forms for unsupervised learning process. Approximations of posterior probabilities finally resulted, after the execution of algorithm, are used to assigning forms on classes, according to rule: a form is assigned to that cluster where the largest posterior probability is noticed.

The unsupervised statistical learning algorithm developed on the basis of EM algorithm supposes the iterative construction of clusters, corresponding to iterations of EM algorithm, on the basis of using approximations of posterior probabilities obtained at each iteration of EM algorithm. Such an algorithm is distinguished from the partitioning algorithms, from *three* points of view: setting certain *seeds* of clusters is not required; an effective re-assignation of forms at each iteration is not required, but only one final assignation , after the algorithm stops. We shall use , instead of the distances among the forms which have to assigned to clusters and cluster centroids the approximations of posterior probabilities of these forms, namely, a form will be assigned to that cluster with the largest posterior probability. In fact, the reciprocal posterior probabilities of classes, calculated for each form, can be considered as representing the *distances* between the respective form and clusters.

A great advantage of EM statistical learning algorithm is the fact that the obtained results, after the algorithm execution, can be also used to *make predictions*, so that, such a program can have also valences, which are specific only to unsupervised learning algorithm. Although the unsupervised learning algorithm developed on the basis EM algorithm is very performing, it has the disadvantage that, in comparison with other partitioning algorithms, supposes to know the probability laws of classes, i.e. the conditioned probability distributions.

In order to illustrate the modality of using EM algorithm as an unsupervised statistical learning algorithm, we shall consider that the latent random variable $W_{,P}$ via the value taken for a certain form, respectively, one of the values $w_1, w_2, ..., w_K$, the class (cluster) , the respective form belongs to, is identified. At the same time, we shall assume that we have a set T of forms with unknown classification , and we have to determine the classification of these forms to classes. In this context, the forms are supposed to represent a number of T realizations of random vector $X_{,P}$, represented by *n*-dimensional vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}$, organized under the so called *observation matrix:*

$$\mathbf{X}_{T\times n} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \cdots & & & \\ x_{T1} & x_{T2} & \cdots & x_{Tn} \end{pmatrix}.$$

If $f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}(\cdot|\cdot)$ is the conditioned probability density of random vector $X_{\mathcal{P}}$, being given the value of random variable $W_{\mathcal{P}}$, then, the marginal distribution of random vector $\mathbf{X}_{\mathcal{P}}$ is described by density:

$$f_{\mathcal{P}}(\mathbf{x}) = \sum_{k=1}^{K} f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}(\mathbf{x}|w_k)p(w_k), \quad \forall \mathbf{x} \in \mathcal{P},$$

where the *k-th* term of the previous sum can be rendered as an *absolute contribution* of class $\mathcal{P}_k$ to the value of probability density in the point represented by vector **x**. This absolute contribution is, in fact, a non-normalized posterior distribution. Each observation vector $\mathbf{x}^{(t)}$, $t=1,2,...,T$ can be associated to one of the values $w_1, w_2, ..., w_K$ of random variable $W_{\mathcal{P}}$, depending on the highest value of this absolute contribution. So, if:

$$\tilde{k} = \arg\max_k \left\{ f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}(\mathbf{x}|w_k)p(w_k) = f_{\mathcal{P}_k}(\mathbf{x})\pi_k \right\},$$

then, vector $\mathbf{x}^{(t)}$ can be associated to class $\mathcal{P}_k$. The previous associating rule can be considered as a *decision rule regarding* the "***origin"*** of form $\mathbf{x}^{(t)}$, respectively:

- if: $\tilde{k} = \arg\max_k \left\{ \pi_k f_{\mathcal{P}_k}(\mathbf{x}^{(t)}) \right\}$, then: $\mathbf{x}^{(t)} \in \mathcal{P}_{\tilde{k}}$.

As can be noticed, term $\pi_k f_{\mathcal{P}_k}(\mathbf{x}^{(t)})$ is exactly the posterior probability numerator for subpopulation $\mathcal{P}_k$ estimated for the observed vector $\mathbf{x}^{(t)}$. The probabilities $\pi_1, \pi_2, ..., \pi_K$ have the nature of *prior probabilities* (historical information) and the products of the form $\pi_1 f_{\mathcal{P}_1}(\mathbf{x}^{(t)}), \pi_2 f_{\mathcal{P}_2}(\mathbf{x}^{(t)}), ..., \pi_K f_{\mathcal{P}_K}(\mathbf{x}^{(t)})$ have the nature of *non-normalized posterior probabilities*, obtained after observing realization $\mathbf{x}^{(t)}$. The posterior probabilities for the observation $\mathbf{x}^{(t)}$ are:

$$h_{W_{\mathcal{P}}|\mathbf{X}_{\mathcal{P}}}(w_k|\mathbf{x}^{(t)}) = \frac{f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}(\mathbf{x}^{(t)}|w_k)p(w_k)}{\sum_{j=1}^{K} f_{\mathbf{X}_{\mathcal{P}}|W_{\mathcal{P}}}(\mathbf{x}^{(t)}|w_j)p(w_j)} = \frac{f_{\mathcal{P}_k}(\mathbf{x}^{(t)})\pi_k}{\sum_{j=1}^{K} f_{\mathcal{P}_j}(\mathbf{x}^{(t)})\pi_j}, \quad k=1,2,...,K.$$

In terms of posterior probabilities, the rule of setting the "*origin*" of generic observation $\mathbf{x}^{(t)}$ can be formulated in the following terms:

- if: $\tilde{k} = \max_k \left\{ h_{W_{\mathcal{P}}|\mathbf{X}_{\mathcal{P}}}(w_k|\mathbf{x}^{(t)}) = \frac{\pi_k f_{\mathcal{P}_k}(\mathbf{x}^{(t)})}{\sum_{j=1}^{K} \pi_j f_{\mathcal{P}_j}(\mathbf{x}^{(t)})} \right\}$, then: $\mathbf{x}^{(t)} \in \mathcal{P}_{\tilde{k}}$.

In statistical learning theory, posterior probabilities $h_{W_{\mathcal{P}}|\mathbf{X}_{\mathcal{P}}}(w_k|\mathbf{x}^{(t)})$ are known also under the generic name of *responsibilities*.

In the next box, the stages of using EM algorithm as unsupervised statistical learning algorithm are concisely described.

EM Algorithm for Unsupervised Statistical Learning

Stage I: Executes EM algorithm using the existing data, referring to the values taken by the characteristics of forms and which are the object of classes assignment and retain the approximations resulted for posterior probabilities, i.e.:

$$\tilde{w}_{tk};\quad k=1,2,...,K;\ t=1,2,...,T;$$

Stage II: Uses the approximations $\tilde{w}_{tk}$ of posterior probabilities in order to assign all the T forms, according to rule:

$$\mathbf{x}^{(t)} \in \mathcal{P}_{\tilde{k}},\quad \text{where}\quad \tilde{k} = \max_{k} \left\{w_{t1}, w_{t2}, ..., w_{tk}, ..., w_{tK}\right\};\quad t=1,2,...T.$$

V. Numerical Exemplification for Gaussian Distributions

The most important problem related to mixture probability distributions is that of estimating the parameters of these mixtures, respectively, the problem of estimating coefficients $\pi_1, \pi_2, ..., \pi_K$ of mixture and estimating parameters of probability densities $f_{\mathcal{P}_1}(\mathbf{x}^{(t)}), f_{\mathcal{P}_2}(\mathbf{x}^{(t)}), ..., f_{\mathcal{P}_K}(\mathbf{x}^{(t)})$, respectively, their expected values, variances and covariances. Generically, the problem is *to model data* represented by observations $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}$ coming from a mixture probability distribution, respectively, *to identify the probabilistic model* for the random mechanism of mixture probability distribution type, which generates observations, without knowing the coefficients value of the respective mixture. Considering the nature of mixture probability distribution, the most appropriate modality to estimating their parameters is that one based on **EM** algorithm (**E**xpected **M**aximization).

We shall assume that we have a sample of *n*-dimensional (independent) observations $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(T)}\right\}$, that the observations are generated from normally distributed subpopulations and are independent, i.e. the likelihood functions has the following form:

$$\mathcal{L}^{*}\left(\pi_1,...,\pi_K, \boldsymbol{\mu}^{(1)},...,\boldsymbol{\mu}^{(K)}, \boldsymbol{\Sigma}^{(1)},...,\boldsymbol{\Sigma}^{(K)} \mid \mathbf{x}^{(1)},...,\mathbf{x}^{(T)}\right) = \prod_{t=1}^{T} f_{\mathcal{P}}\left(\mathbf{x}^{(t)}\right) = \prod_{t=1}^{T}\left[\sum_{k=1}^{K} \pi_k f_{\mathcal{P}_k}\left(\mathbf{x}^{(t)}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right)\right],$$

where:

$$f_{\mathcal{P}_k}\left(\mathbf{x}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right) = \frac{1}{(2\pi)^{n/2}\,|\boldsymbol{\Sigma}|^{1/2}} e^{\frac{1}{2}\left(x-\boldsymbol{\mu}^{(k)}\right)\left(\boldsymbol{\Sigma}^{(k)}\right)^{-1}\left(x-\boldsymbol{\mu}^{(k)}\right)},\quad k=1,2,...,K,$$

are the probability distributions of subpopulations, and $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ is the vector of expected values, respectively, the covariance matrix corresponding to subpopulation $\mathcal{P}_k$. We shall denote:

$$\begin{aligned}\mathcal{L}\left(\pi_1,...,\pi_K, \boldsymbol{\mu}^{(1)},...,\boldsymbol{\mu}^{(K)}, \boldsymbol{\Sigma}^{(1)},...,\boldsymbol{\Sigma}^{(K)}\right) &= \ln \mathcal{L}^{*}\left(\pi_1,...,\pi_K, \boldsymbol{\mu}^{(1)},...,\boldsymbol{\mu}^{(K)}, \boldsymbol{\Sigma}^{(1)},...,\boldsymbol{\Sigma}^{(K)} \mid \mathbf{x}^{(1)},...,\mathbf{x}^{(T)}\right) \\ &= \sum_{t=1}^{T} \ln\left[\sum_{k=1}^{K} \pi_k f_{\mathcal{P}_k}\left(\mathbf{x}^{(t)}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right)\right].\end{aligned}$$

In this context, the problem is to obtain the estimates for mixture probability parameters, respectively $\pi_1,...,\pi_K, \boldsymbol{\mu}^{(1)},...,\boldsymbol{\mu}^{(K)}, \boldsymbol{\Sigma}^{(1)},...,\boldsymbol{\Sigma}^{(K)}$, so that the T observations sample to be as representative as possible (higher likelihood), what is equivalent to maximizing the likelihood function, with respect to the respective parameters.

Assuming that values of parameters $\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)},...,\boldsymbol{\mu}^{(K)},\boldsymbol{\Sigma}^{(1)},\boldsymbol{\Sigma}^{(2)},...,\boldsymbol{\Sigma}^{(K)}$ are known, the maximum likelihood estimates for parameters $\pi_1,...,\pi_K$ are obtained as solutions of the next conditioned extreme problem:

$$\begin{cases} \max\limits_{\pi_1,...,\pi_K} \mathcal{L}\left(\pi_1,,...,\pi_K \mid \boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)},...,\boldsymbol{\mu}^{(K)},\boldsymbol{\Sigma}^{(1)},\boldsymbol{\Sigma}^{(2)},...,\boldsymbol{\Sigma}^{(K)}\right) \\ \sum\limits_{k=1}^{K}\pi_k = 1, \end{cases}$$

wherefore, the Lagrangean function is of the form:

$$G\left(\pi_1,,....,\pi_K;\lambda\right) = \sum_{t=1}^{T}\ln\left[\sum_{k=1}^{K}\pi_k f_{;\mathcal{P}_k}\left(\mathbf{x}^{(t)};\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}\right)\right] - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).$$

From the maximum conditions of Lagrangean function, according to parameters $\pi_1,\pi_2,...,\pi_K$, it follows that estimations for parameters $\pi_1,\pi_2,...,\pi_K$ are:

$$\hat{\pi}_k = \sum_{t=1}^{T}\frac{w_{tk}}{\sum\limits_{i=1}^{T}\sum\limits_{j=1}^{K}w_{ij}},\ \text{where:}\ w_{tk} = h_{W,\mathcal{P}\mid\mathbf{X},\mathcal{P}}\left(w_k\mid\mathbf{x}^{(t)}\right) = \frac{\pi_k f_{;\mathcal{P}_k}\left(\mathbf{x}^{(t)};\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}\right)}{\sum\limits_{j=1}^{K}\pi_j f_{;\mathcal{P}_j}\left(\mathbf{x}^{(t)};\boldsymbol{\mu}^{(j)},\boldsymbol{\Sigma}^{(j)}\right)},\quad k=1,2,...,K.$$

Supposing as known the weights $\pi_1,\pi_2,...,\pi_K$, by annulling the partial derivatives of log likelihood function $\mathcal{L}\left(\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)},...,\boldsymbol{\mu}^{(K)},\boldsymbol{\Sigma}^{(1)},\boldsymbol{\Sigma}^{(2)},...,\boldsymbol{\Sigma}^{(K)}\mid\pi_1,\pi_2,...,\pi_K\right)$, according to vectors of avarages and covariance matrices of subpopulations, it follows that maximum likelihood estimates for them, are:

$$\overline{\mathbf{x}}^{(k)} = \sum_{t=1}^{T}\frac{w_{tk}}{\sum\limits_{j=1}^{T}w_{jk}}\mathbf{x}^{(t)};\qquad \mathbf{S}^{(k)} = \sum_{t=1}^{T}\frac{w_{tk}}{\sum\limits_{j=1}^{T}w_{jk}}\left(\mathbf{x}^{(t)}-\overline{\mathbf{x}}^{(k)}\right)\left(\mathbf{x}^{(t)}-\overline{\mathbf{x}}^{(k)}\right)^t,$$

where:

$$w_{tk} = h_{W,\mathcal{P}\mid\mathbf{X},\mathcal{P}}\left(w_k\mid\mathbf{x}^{(t)}\right) = \frac{\pi_k f_{;\mathcal{P}_k}\left(\mathbf{x}^{(t)};\boldsymbol{\mu}^{(k)},\boldsymbol{\Sigma}^{(k)}\right)}{\sum\limits_{j=1}^{K}\pi_j f_{;\mathcal{P}_j}\left(\mathbf{x}^{(t)};\boldsymbol{\mu}^{(j)},\boldsymbol{\Sigma}^{(j)}\right)}.$$

As we notice, there is a certain "vicious circle" (circularity) given by the fact that the estimates for mixture probability distribution parameters, inclusively, the estimates for mixture probability distribution coefficients, depend on parameters $\pi_1,\pi_2,...,\pi_K$, i.e. they can be never estimated. For this reason, in order to estimate the mixture probability density parameters, we use the iterative procedure called ***EM algorithm***, that eliminates the mentioned vicious circle.

Hereinafter, we shall present a modality of software implementation of EM algorithm and we shall numerically exemplify how this algorithm is running, for the case where the probability distributions of subpopulations are of Gaussian type.

We shall consider the case of a statistical population $\mathcal{P}$, structured on three statistical subpopulations $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$, with weights represented by vector $\boldsymbol{\pi} = (0.40, 0.35, 0.25)$, whose probability laws are of Gaussian type, with vectors of expected values:

$$\boldsymbol{\mu}^{(1)} = \begin{pmatrix} 5.0 \\ 4.0 \end{pmatrix}, \quad \boldsymbol{\mu}^{(2)} = \begin{pmatrix} 11.0 \\ 13.0 \end{pmatrix}, \quad \boldsymbol{\mu}^{(3)} = \begin{pmatrix} 16.0 \\ 7.0 \end{pmatrix},$$

respectively, with covariance matrices:

$$\boldsymbol{\Sigma}^{(1)} = \begin{pmatrix} 6.5000 & -3.3727 \\ -3.3727 & 7.0000 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{(2)} = \begin{pmatrix} 5.5000 & 3.5875 \\ 3.5875 & 6.5000 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{(3)} = \begin{pmatrix} 6.0000 & 2.7386 \\ 2.7386 & 5.0000 \end{pmatrix}.$$

As it is known, the vector of expected values of mixture probability distribution (of statistical population $\mathcal{P}$) is defined as a convex combination of statistical subpopulation expected values, respectively:

$$\boldsymbol{\mu}^{(\mathcal{P})} = \pi_1 \boldsymbol{\mu}^{(1)} + \pi_2 \boldsymbol{\mu}^{(2)} + \pi_3 \boldsymbol{\mu}^{(3)} = \begin{pmatrix} 9.85 \\ 7.90 \end{pmatrix},$$

and the *covariance matrix of mixture probability*, defined as the sum between the *common covariance matrix*, calculated as a convex combination of subpopulation covariance matrices, namely:

$$\boldsymbol{\Sigma}^{(\mathrm{w})} = \pi_1 \boldsymbol{\Sigma}^{(\mathcal{P}_1)} + \pi_2 \boldsymbol{\Sigma}^{(\mathcal{P}_2)} + \pi_3 \boldsymbol{\Sigma}^{(\mathcal{P}_3)} = \begin{pmatrix} 6.0250 & 0.5912 \\ 0.5912 & 6.3250 \end{pmatrix},$$

and the *covariance matrix of conditioned expected values*, i.e.:

$$\boldsymbol{\Sigma}^{(\mathcal{P})} = \boldsymbol{\Sigma}^{(\mathrm{w})} + \boldsymbol{\Sigma}^{(\mathrm{b})} = \begin{pmatrix} 6.0250 & 0.5912 \\ 0.5912 & 6.3250 \end{pmatrix} + \begin{pmatrix} 19.3275 & 8.2350 \\ 8.2350 & 15.3900 \end{pmatrix} = \begin{pmatrix} 25.3525 & 8.8262 \\ 8.8262 & 21.7150 \end{pmatrix}.$$

In the figures bellow, the probability density of the mixture of the three probability distributions and the isoprobanility curves for this density are graphically represented.

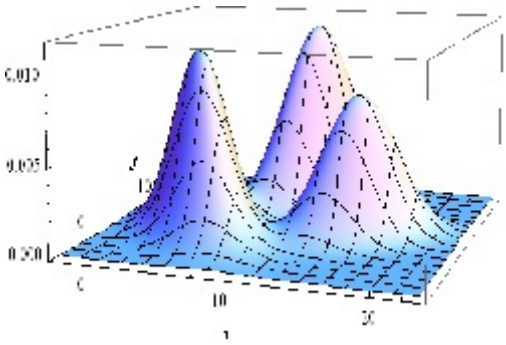

Figure 2: Graph of the probability density function for Gaussien mixture distribution

Figure 3: Contour plot (iso-probabilities) for Gaussien mixture distribution

In order to use the EM algorithm for estimating the mixture parameters of the three probability distributions, we have randomly generated a number of $T = 1000$ of statistical

population $\mathcal{P}$, observations which we shall use for algorithm running. The next figure illustrates the graphic representation of the 1000 observations.

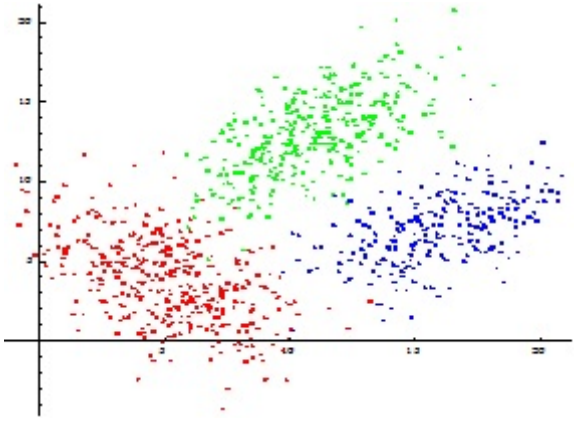

Figure 4: Graph of the sample from Gaussien mixture distribution (1000 obs.)

The mixture distribution parameters are represented by the weights vector and by the parameters vector of the three mixture distribution, respectively, the vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)^t$ and the parameters vector $\theta = (\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, ..., \boldsymbol{\mu}^{(3)}, \boldsymbol{\Sigma}^{(1)}, \boldsymbol{\Sigma}^{(2)}, \boldsymbol{\Sigma}^{(3)})^t$. As start values for parameter vectors $\boldsymbol{\pi}$ and θ, randomly generated values of uniform distribution have been used. The solution representing the estimates of the two parameter vectors has been obtained in a number of 32 iterations, and the values obtained at each iteration for maximized likelihood function, for its increments and for mixture weights, are given in the next table:

Table 1

Iteration	$\ln \mathcal{L}_{\mathcal{P}}^{*}(\hat{\boldsymbol{\pi}}, \hat{\theta})$	$\Delta \ln \mathcal{L}_{\mathcal{P}}^{*}(\hat{\boldsymbol{\pi}}, \hat{\theta})$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$
1	-5949.88	239.24800	0.370612	0.316577	0.312811
2	-5853.49	96.38680	0.361588	0.326097	0.312315
3	-5819.25	34.24240	0.346887	0.344603	0.308510
4	-5788.86	30.38620	0.331885	0.362969	0.305146
5	-5769.34	19.52180	0.320630	0.374674	0.304696
...					
14	-5703.48	4.21714	0.275952	0.386679	0.337369
...					
21	-5530.47	63.11210	0.332168	0.370092	0.29774
...					
25	-5469.60	5.37763	0.379999	0.359323	0.260677
...					
30	-5463.93	0.07270	0.391535	0.357548	0.250918
31	-5463.90	0.02344	0.391907	0.357442	0.250651
32	-5463.89	0.00755	0.392123	0.357372	0.250505

The values of optimized likelihood function corresponding to each of the 32 iterations are illustrated in Figure 5.

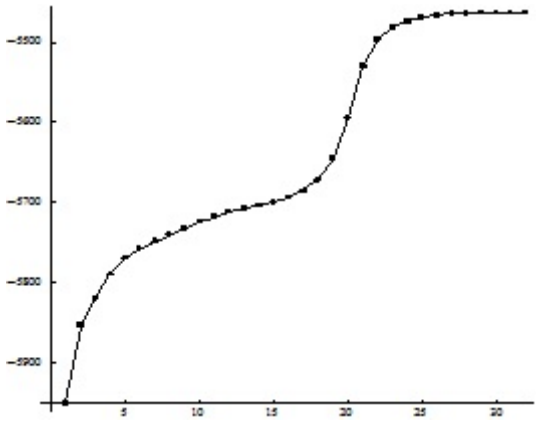

Figure 5: Graph of the log likelihood function

The limit error used in order to formulate the stop criterion was $\varepsilon = 0.01$ and the estimates for mixture parameters, respectively, weights vector, subpopulation expected valuesvectors and subpopulation covariance matrices are:

$$\hat{\boldsymbol{\pi}} = \begin{pmatrix} 0.392123 \\ 0.357372 \\ 0.250505 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}^{(1)} = \begin{pmatrix} 5.16678 \\ 3.71517 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}^{(2)} = \begin{pmatrix} 10.9015 \\ 12.8536 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}^{(3)} = \begin{pmatrix} 15.9406 \\ 6.8707 \end{pmatrix};$$

$$\boldsymbol{\Sigma}^{(1)} = \begin{pmatrix} 5.84161 & -3.05546 \\ -3.05546 & 6.44357 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{(2)} = \begin{pmatrix} 6.09675 & 3.94700 \\ 3.94700 & 6.53935 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{(3)} = \begin{pmatrix} 5.91361 & 2.85728 \\ 2.85728 & 4.66311 \end{pmatrix}.$$

According to these estimates, the estimates of expected values vector and of covariance matrix at population level $\mathcal{P}$ are the following:

$$\hat{\boldsymbol{\mu}}^{(\mathcal{P})} = \begin{pmatrix} 9.91509 \\ 7.77147 \end{pmatrix};$$

$$\hat{\boldsymbol{\Sigma}}^{(\mathcal{P})} = \hat{\boldsymbol{\Sigma}}^{(w)} + \hat{\boldsymbol{\Sigma}}^{(b)} = \begin{pmatrix} 5.95083 & 0.92819 \\ 0.92819 & 6.03178 \end{pmatrix} + \begin{pmatrix} 18.28370 & 7.98435 \\ 7.98435 & 15.88530 \end{pmatrix} = \begin{pmatrix} 24.23450 & 8.91254 \\ 8.91254 & 21.91710 \end{pmatrix}.$$

As can be noticed, the estimates obtained by using the EM algorithm for mixture parameters of the three Gaussian probability distributions come very close to the true values of parameters.

The EM algorithm, via the provided results, can be also used as an unsupervised learning algorithm, respectively, as a cluster analysis algorithm. Thus, according to the approximations obtained after the last iteration for posterior probabilities, the 1000 observations can be assigned on the three classes (subpopulations), depending on the class wherefore the highest posterior probability is obtained. As the 1000 observations have been generated on subpopulations, we know the real classification of observations, so that we can compare the real classification with the assignations based on posterior probabilities. The matrix of classification and the corresponding matrix of relative frequencies are:

$$\mathbf{M}_{\text{Clasif}} = \begin{pmatrix} 393 & 4 & 2 \\ 4 & 344 & 2 \\ 3 & 2 & 246 \end{pmatrix}; \quad \mathbf{M}_{\text{Freq Clasif}} = \begin{pmatrix} 0.9825 & 0.0114 & 0.0080 \\ 0.0100 & 0.9829 & 0.0080 \\ 0.0075 & 0.00571 & 0.9840 \end{pmatrix}.$$

We can notice that, from the 1000 observations, the classification of the 983 observations has been *correctly predicted* , i.e. a very high percentage of classification correctness, respectively 98.30%.

The software implementation of the EM algorithm described above was made in Mathematica language. The code corresponding to this implementation, used to achieve the previously described application, is found in the next box.

Source Code of the EM Algorithm Implementation

```
{On[General::stop], maxiter=100, delt=10,eps=10^(-Round[8.0/Log[Sqrt[T]]]), iter=1};
{w1=Array[0,T], w2=Array[0,T], w3=Array[0,T], w=Array[0,T], wn1=Array[0,T], wn2=Array[0,T], wn3=Array[0,T], wn=Array[0,T]};
{fverosim=Array[0,maxiter], med=Mean[Esant], cov=Covariance[Esant]};
{repe=MultinormalDistribution[med,cov]; v[x_]=PDF[repe,x]; fverosimold=1.05 func {sum from {t=1} to T}Log[v[Esant[[t]]]];};
{lix=0.65 med[[1]], lsx=1.30 med[[1]], liy=0.65 med[[2]], lsy=1.30 med[[2]]};
{li11=0.65 cov[[1,1]], ls11=1.30 cov[[1,1]], li12=0.65 cov[[1,2]], ls12=1.30 cov[[1,2]], li22=0.65 cov[[2,2]], ls22=1.30 cov[[2,2]]};
{m1={RandomReal[{lix,lsx}], RandomReal[{liy,lsy}]}, s1={{RandomReal[{li11,ls11}], a=RandomReal[{li12,ls12}]},
{a,RandomReal[{li22,ls22}]}}};
{m2={RandomReal[{lix,lsx}], RandomReal[{liy,lsy}]}, s2={{RandomReal[{li11,ls11}], a=RandomReal[{li12,ls12}]},
{a,RandomReal[{li22,ls22}]}}};
{m3={RandomReal[{lix,lsx}], RandomReal[{liy,lsy}]}, s3={{RandomReal[{li11,ls11}], a=RandomReal[{li12,ls12}]},{a,RandomReal[{li22,ls22}]}}};
{q1=RandomReal[{0.25,0.50}], q2=RandomReal[{0.15,0.50}], q3=1-(q1+q2)};
While[(delt>eps && iter <= maxiter),
 {repe1=MultinormalDistribution[m1,s1]; repe2=MultinormalDistribution[m2,s2]; repe3=MultinormalDistribution[m3,s3];};
 {v1[x_]=PDF[repe1,x]; v2[x_]=PDF[repe2,x]; v3[x_]=PDF[repe3,x];};
 For[i=1,i<=T,
  {obs=Esant[[i]]; wn[[i]]=q1 v1[obs]+q2 v2[obs]+q3 v3[obs];};
  {wn1[[i]]=q1 v1[obs]/wn[[i]]; wn2[[i]]=q2 v2[obs]/wn[[i]]; wn3[[i]]=q3 v3[obs]/wn[[i]];};
  i++;
 ];
 {sumwn1= Sum[wn1[[t]], {t,1,T}]; sumwn2= Sum[wn2[[t]], {t,1,T}]; sumwn3= Sum[wn3[[t]], {t,1,T}]; sumwn=sumwn1+sumwn2+sumwn3;};
 {qn1=sumwn1/sumwn; qn2=sumwn2/sumwn; qn3=sumwn3/sumwn; fverosimnew= Sum[Log[wn[[t]]], {t,1,T}]; delt=fverosimnew-fverosimold;};
 If[delt>0,
  {m1=(1.0/sumwn1) Sum[(wn1[[t]] Esant[[t]]), {t,1,T}]; s1=1.0/(sumwn1) Sum[(wn1[[t]] Transpose[{Esant[[t]]-m1}].{Esant[[t]]-m1}), {t,1,T}];};
  {m2=(1.0/sumwn2) Sum[(wn2[[t]] Esant[[t]]), {t,1,T}]; s2=1.0/(sumwn2) Sum[(wn2[[t]] Transpose[{Esant[[t]]-m2}].{Esant[[t]]-m2}), {t,1,T}];};
  {m3=(1.0/sumwn3) Sum[(wn3[[t]] Esant[[t]]), {t,1,T}]; s3=1.0/(sumwn3) Sum[(wn3[[t]] Transpose[{Esant[[t]]-m3}].{Esant[[t]]-m3}), {t,1,T}];};
  {w1=wn1; w2=wn2; w3=wn3; w=wn; fverosimold=fverosimnew; fverosim[[iter]]=fverosimnew;};
  {sumw1=sumwn1; sumw2=sumwn2; sumw3=sumwn3; sumw=sumwn; q1=qn1; q2=qn2; q3=qn3; iter++};
 ];
 Clear[repe1,repe2,repe3,v1,v2,v3];
];
{iter--; delt=fverosim[[iter]]-fverosim[[iter-1]]; m=q1 m1+q2 m2+q3 m3, SW=q1 s1+q2 s2+q3 s3};
```

REFERENCES

[1] **Banerjee A., Merugu S., Dhillon I. S., Ghosh J.** (2005), ***Clustering with Bregman divergences***. Journal of Machine Learning Research, 6:1705–1749, 2005

[2] **Bishop C.** (1995), ***Neural Networks for Pattern Recognition***, Clarendon Press, Oxford, 1995

[3] **Cappe O., Moulines E.** (2009), ***On-line Expectation-Maximization algorithm for latent data models***, Journal Royal Statistics Society, B, 71(3) : 593-613, 2009

[4] **Crane, R., Sornette, D.** (2008), ***Robust dynamic classes revealed by measuring the response function of a social system***, Proceedings of the National Academy of Sciences, 105, 15649–15653, 2008

[5] **Dempster A. P., Laird N. M., Rubin D. B.** (1977), ***Maximum likelihood from incomplete data via the em algorithm***, Journal of the Royal Statistical Society: Series B, 39(1):1–38, November 1977

[6] **Ghahramami Z., Jordan M.** (1995), ***Learning from incomplete data***, Technical Report AI Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, August 1995

[7] **Ghahramani Z., Jordan M.** (1994), ***Supervised learning from incomplete data via an EM approach***, in Advances in Neural Information Processing Systems, editors Jack D. Cowan, Gerald Tesauro, Joshua Alspector, volume 6, Morgan Kaufmann, 1994

[8] **Jordan M., Jacobs R.** (1994), ***Hierarchical mixtures of experts and the em algorithm***, Neural Computation, 6:181–214, 1994

[9] **Lange K.** (2002), ***Mathematical and Statistical Methods for Genetic Analysis***, 2nd ed. Springer, New York, 2002

[10] **Liang P., Klein D.** (2009), ***Online EM for Unsupervised Models***, In Proc NAACL Conference, 2009

[11] **McLachlan G., Thriyambakam K.** (1996), ***The EM Algorithm and Extensions***, John Wiley & Sons, New York, 1996

[12] **McLachlan G. J., Krishnan T.** (1997), ***The EM Algorithm and Extensions***, Wiley, New York, 1997

[13] **Neal R. M., Hinton G. E.** (1999), ***A view of the EM algorithm that justifies incremental, sparse and other variants***, in Learning in graphical models, editor Jordan M. L., pages 355-368, MIT Press, Cambridge, MA, USA, 1999

[14] **Qin J., Zhang B., Leung D.** (2009), ***Empirical likelihood in missing data problems***, Journal of the American Statistical Association, 104: 1492-1503, 2009

[15] **Rabiner L., Juang B. H.** (1993), ***Fundamentals of Speech Recognition***, Prentice Hall, Signal Processing Series, 1993

[16] **Redner R., Walker H.** (1984), ***Mixture densities, maximum likelihood and the EM algorithm***, SIAM Review, 26(2), 1984

[17] **Ruxanda Gh.** (2010), ***Learning Neural Network Perceptron with Backpropagation Algorithm***, Economic Computation and Economic Cybernetics Studies and Research, no. 4/2010, Academy of Economic Studies, 2010

[18] **Ruxanda Gh.** (2009), ***Supervised Pattern Recognition with Potential Functions Methods***, Economic Computation and Economic Cybernetics Studies and Research, no. 2/2009, Academy of Economic Studies, 2009

[19] **Sato M., Ishii S.** (2000), ***On-line EM algorithm for the normalized Gaussian network***, Neural Computation, 12: 407-432, 2000

[20] **Sato M.** (2000). ***Convergence of on-line EM algorithm***, in Proc. International Conference on Neural Information Processing, 1:476-481, 2000

[21] **Titterington D. M.** (1984), ***Recursive parameter estimation using incomplete data***, J. Roy. Statist. Soc. B, 46(2): 257-267, 2000

[22] **Xu L., Jordan M.** (1996), ***On convergence properties of the EM algorithm for gaussian mixtures***, Neural Computation, 8:129–151, 1996