



**ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS**

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

**A COMPARATIVE STUDY OF
DIVERGENCE MEASURES USED
IN IMAGE SEGMENTATION**

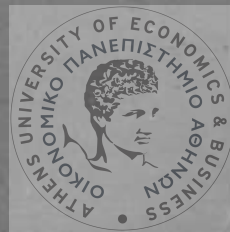
By

Chrissanthi S. Seizi

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
2005

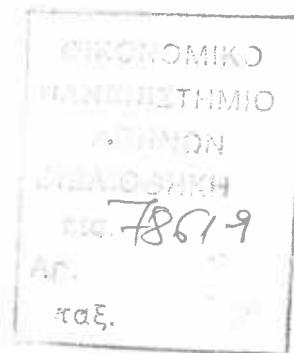


01000000554459



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ





**ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS**

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

**A COMPARATIVE STUDY OF
DIVERGENCE MEASURES USED
IN IMAGE SEGMENTATION**

By

Chrissanthi S. Seizi

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
June 2005





ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΣΥΓΚΡΙΤΙΚΗ ΜΕΛΕΤΗ ΜΕΤΡΩΝ ΑΠΟΚΛΙΣΗΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΝΤΑΙ ΣΤΗΝ ΚΑΤΑΤΜΗΣΗ ΕΙΚΟΝΑΣ

Χρυσάνθη Σταύρου Σεΐζη

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

**Αθήνα
Ιούνιος 2005**





**ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS
DEPARTMENT OF STATISTICS**

A Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

**A COMPARATIVE STUDY OF
DIVERGENCE MEASURES USED
IN IMAGE SEGMENTATION**

Chrissanthi S. Seizi



Approved by the Graduate Committee

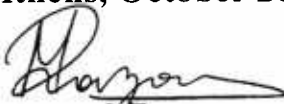
P. Tsiamyrtzis
Lecturer
Thesis Supervisor

E. Ioannidis
Lecturer

D. Karlis
Assistant Professor

Members of the Committee

Athens, October 2005


**Michael Zazanis, Professor
Director of the Graduate Program**



ACKNOWLEDGEMENTS

I would like to thank my supervisor, Lecturer Panagiotis Tsiamyrtzis, for his support during the development of this thesis. I would also like to thank Professor Ioannis Pavlidis of Computational Physiology Lab, Dept. of Computer Science, University of Houston for providing me with the data set used in this thesis.





VITA

I was born in Veria in 1979. On 1997 I graduated from Lyceum and I started my studies at the Department of Mathematics in the Aristotle University of Thessaloniki. Four years later I graduated and after a year, I was accepted as a postgraduate student at the Department of Statistics in the Athens University of Economics and Business.





ABSTRACT

Chrissanthi Seizi

“A Comparative Study of Divergence Measures Used in Image Segmentation”

June 2005

The problem described in this thesis is a general problem of image segmentation when images come from a thermal video sequence taping human subject(s). Our goal is (in real time) to isolate, within each frame, the pixels indicating human body.

Using the temperature value of each pixel (recorded by the thermal camera), we will describe two methods which will allow us to segment the first frame of the video. Then, a dynamic process will take action in order to categorize the pixels of each of the subsequent frames to the appropriate class.

A series of different distance and discrepancy measures can be used in the dynamic approach. Theoretic properties regarding several measures will be provided, along with their specific forms under special distributional assumptions of the study. Apart from that, the issue on how to update the parameters in a mixture after collapsing some of their components will be covered as well.

This work will conclude by applying and judging upon their performance, several of the aforementioned techniques in a short thermal video clip.





ΠΕΡΙΛΗΨΗ

Χρυσάνθη Σεΐζη

“Συγκριτική Μελέτη Μέτρων Απόκλισης που Χρησιμοποιούνται στην Κατάτμηση Εικόνας”

Ιούνιος 2005

Το πρόβλημα που περιγράφεται σε αυτή τη διατριβή είναι ένα γενικό πρόβλημα κατάτμησης εικόνας, όταν η εικόνα προέρχεται από μία θερμική βίντεο εγγραφή ανθρώπου(ων). Σκοπός μας είναι (σε πραγματικό χρόνο) να απομονώσουμε, σε κάθε εικόνα, τις ψηφίδες (pixels) που αναλογούν στο ανθρώπινο σώμα.

Χρησιμοποιώντας τη θερμοκρασία της κάθε ψηφίδας (η οποία καταγράφεται από την θερμική κάμερα), θα περιγράψουμε δύο μεθόδους οι οποίες θα επιτρέψουν την κατάτμηση της πρώτης εικόνας του βίντεο. Εν συνεχεία με τη βοήθεια μιας δυναμικής διαδικασίας θα κατηγοριοποιούνται οι ψηφίδες καθεμιάς από τις επερχόμενες εικόνες του βίντεο στην κατάλληλη κλάση.

Μια σειρά από διαφορετικές μετρήσεις αποστάσεων και αποκλίσεων μπορούν να χρησιμοποιηθούν στη δυναμική διαδικασία. Θεωρητικές ιδιότητες που αφορούν διάφορες μετρήσεις θα αναφερθούν, μαζί με τις ειδικές μορφές που λαμβάνουν κάτω από συγκεκριμένες υποθέσεις κατανομών. Πέραν τούτου, θα αναφερθεί και το θέμα του πως ανανεώνουμε τις παραμέτρους μιας μίξης στην περίπτωση που συγχωνεύουμε κάποια μέλη της.

Η παρούσα εργασία θα περατωθεί με την εφαρμογή και τη συγκριτική μελέτη, απόδοσης διαφόρων τεχνικών που έχουν αναφερθεί παραπάνω σε ένα σύντομο θερμικό βίντεο.





TABLE OF CONTENTS

	Page
Chapter 1: Introduction	1
Chapter 2: Initialization	
2.1 Introduction	5
2.2 K-means	5
2.2.1 The method	6
2.2.2 The distance in k-means.....	6
2.2.3 The initial centroids	6
2.2.4 The algorithm's convergence.....	7
2.2.5 The algorithm's rate of convergence.....	7
2.2.6 Other properties	7
2.3 EM algorithm	8
2.3.1 The method	8
2.3.2 Criteria of convergence	10
2.3.3 Characteristics	11
2.4 Example	12
2.5 Spatial aware classification scheme	17
Chapter 3: The matching operation	
3.1 Introduction	19
3.2 Measures of distance	20
3.2.1 A review of distance measures	21
3.2.2 Inequalities among distance measures	38
3.3 The example	48
Chapter 4: Updating the parameters	
4.1 Introduction	51
4.2 The method of moments	52



4.3 Method of moments in our example	53
4.4 Simulation Study	56
 Chapter 5: Experimental results	
5.1 Introduction	61
5.2 The results	61
 Chapter 6: Conclusions – Further research	
6.1 Conclusions	69
6.2 Further research	70



LIST OF TABLES

<i>Table</i>	<i>Page</i>
5.2.1: The number of pixels of the last frame in the area of interest which where assigned to human body, background and grey zone for some divergences, with the corresponding percentages.....	63
5.2.2: The number of pixels of the last frame in the area of interest which where assigned to human body, background and grey zone for some formulas for ρ , with the corresponding percentages.....	66





LIST OF FIGURES

<i>Figure</i>	<i>Page</i>
2.4.1: Thermal image of the first frame's temperatures	12
2.4.2: A histogram of the pixel temperature values	13
2.4.3: The classification of first frame's pixels through k-means method	14
2.4.4: The classification of first frame's pixels through EM algorithm ...	15
2.4.5: The differences between the classification of first frame's pixels through k-means and EM algorithm	16
3.3.1: The assignment of second frame's pixels according to Jeffrey's divergence	48
3.3.2: The assignment of second frame's pixels according to L_2 distance	49
4.4.1: A qq plot.....	57
5.2.1: The area of pixels in the first frame in which our results were checked	62
5.2.2: The area of pixels in the last frame in which our results were checked	62
5.2.3: The classification of last frame's pixels using L_2 , H_2 , Kullback Leibler divergence, Jeffrey's divergence, Bhattacharrya and Chi-Squared divergence.....	64



5.2.4: The classification of last frame’s pixels using formulas 1, 2,
3, 4, 5 and 6 for ρ 67



CHAPTER 1

INTRODUCTION

The problem described in this thesis is a general problem of image segmentation when images come from a video sequence. In order to describe the problem a little bit more extensively, we can say that a video sequence consists of N frames where in each frame/image we have $r \times c$ pixels, with r and c corresponding to the number of rows and columns respectively. The data values available to us, are the temperatures of each pixel and based on these we are interested in classifying each pixel of every frame into one of the two broad categories that we will call skin and background.

In practice it turned out that the class of background pixels was consisting of two subclasses: the actual background pixels (referring to the environment around the subject) and the skin pixels covered by clothes, whose temperatures were higher than the actual background pixels but lower than the subject's (uncovered) skin pixels. Thus, we decided to work with three categories, calling the new one, grey zone.

Our method is similar to Grimson's et al (1998) method in the sense that we also use a multi-Normal representation at the pixel level. However, this is where the similarity ends. We, just like Morellas et al (2003), use an expectation-maximization (EM) algorithm to initialize our models. However, in contrast to Morellas et al (2003) who then use just the Jeffreys divergence, we then use several divergence measures as the matching criterion between Normals of incoming pixels and existing model Normals.

So, initially, in chapter 2, we will try to partition the first frame in three (non-overlapping) areas corresponding to skin, background and grey zone through model based clustering. This will be attempted by two different cluster analysis techniques. The first one is k-means which manages to split



the image in three areas committing each pixel to one particular group. However, we would rather assign each pixel with probability to all of the groups and that is attempted with the second technique: the EM algorithm. After this, each pixel x_j of the first frame is considered to be a mixture of the three Normal distributions:

$$x_j \sim w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + w_3 N(\mu_3, \sigma_3^2)$$

So, from the application of EM algorithm on the first frame's pixels we take estimators $\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2$ for the parameters $\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2$ of the mixture of distributions. Those estimators are the same for all of the first frame's pixels.

Of course, in what follows, each pixel x_j of the following frames is also considered to be a mixture of three Normal distributions, but as we will mention in what follows, the parameters of the mixture are different for each one of the pixels of the following frames.

We then, in chapter 3, start the process of awareness of next frame's pixels. We need to classify the pixels of incoming frames to one of the three Normal distributions $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), N(\mu_3, \sigma_3^2)$. We assume that every new pixel comes from a Normal distribution $N(\mu_p, \sigma_p^2)$, where μ_p is the actual value of the pixel's temperature and σ_p^2 is a (known) value related to the camera's accuracy. We then measure how close is the distribution of the incoming pixel $N(\mu_p, \sigma_p^2)$ to the three existing distributions $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), N(\mu_3, \sigma_3^2)$. Several distance and divergence measures are presented and applied in this chapter. The incoming pixel is then ascribed to that Normal from which it desist less.

When we have founded the one of the three Normal distributions $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), N(\mu_3, \sigma_3^2)$, say $N(\mu_i, \sigma_i^2)$ which is the 'closest' one to the distribution $N(\mu_p, \sigma_p^2)$ of the incoming pixel, the incoming pixel contributes to the process of awareness of the parameters of $N(\mu_i, \sigma_i^2)$. $N(\mu_i, \sigma_i^2)$ and $N(\mu_p, \sigma_p^2)$ are approximated by a single Normal component. The next issue to be clarified is what will be the parameters (weight, mean

and variance) of this new updated distribution. We present, in chapter 4, the method of moments in updating those parameters and we propose some other formulas as well. When the procedure is applied to all $r \times c$ pixels of second frame, the three distributions accounting for each pixel are different from the distributions accounting for another pixel.

The process of awareness is applied to all subsequent frames of a short video clip. The parameters of the mixture of the three Normal distributions accounting for a pixel in a frame are now different from the parameters of the mixture accounting for the corresponding pixel of another frame. Then, the performance of different methods is judged in chapter 5. Finally, the Matlab code used along with a short description is presented in an appendix.

We should make a discussion here about the parameters of the three Normal distributions accounting for a pixel. Those parameters are initially the same for all of the first frame's pixels but when the algorithm continues the things change. It would not be correct to have the same value for all the skins temperatures because in fact there are certain areas in the face (for example the areas around the veins) which have larger temperatures than other areas of the face. So, except for the first frame, in all the other frames the mixtures of Normal distributions accounting for each pixel are different.

Moreover, because of the fact that the subject is not static through the video, it is moving; some pixels can change from background to skin or grey zone and reverse. So, the parameters are reset in every frame and the algorithm is award of the new areas to which the pixels move.

We could then say that one of the important advantages of the algorithm is that everything is being examined pixel wise and it thus makes global fit versus local fit.

In what follows the algorithm which has been applied is given with all its steps.

Step 1: We apply k-means on first frame's pixels.

Step 2: We use the k-means estimates as initial values in EM which is applied on first frame's pixels. EM provides estimators $\hat{w}_1, \hat{w}_2, \hat{w}_3, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2$ about the weights, means and variances of the three distributions

which stand for skin, background and grey zone. Those estimators are the same for all of the pixels of the first frame.

Step 3: We take the second frame's pixels and assume that every new pixel comes from a Normal distribution $N(\mu_p, \sigma_p^2)$, where μ_p is the actual value of the pixel's temperature and σ_p^2 is a (known) value related to the camera's accuracy. For each pixel we calculate the distance of distribution $N(\mu_p, \sigma_p^2)$ from each one of the Normal distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$. Each pixel is then assigned to that distribution from which $N(\mu_p, \sigma_p^2)$ desists less.

Step 4: For each pixel of second frame we allow $N(\mu_p, \sigma_p^2)$ to contribute to the process of awareness of the parameters of $N(\mu_i, \sigma_i^2)$, which is the distribution from which $N(\mu_p, \sigma_p^2)$ desists less. This updating of the parameters is being happened according to the method of moments but we also present some other formulas too. After this step, the parameters of the three distributions accounting for each pixel are different from the parameters of the distributions accounting for another pixel.

Steps 3 and 4 succeed each other for all of the rest frames. Every time, for every frame, the assumed distributions of the incoming pixel $N(\mu_p, \sigma_p^2)$ contributes to the awareness of the parameters of $N(\mu_i, \sigma_i^2)$ from which $N(\mu_p, \sigma_p^2)$ desists less. After this, the parameters of the three Normal distributions accounting for a pixel in a frame are different from the parameters of the mixture accounting for the corresponding pixel of another frame.

Step 5: The obtained results are checked. We choose an area which in all frames includes skin pixels and count the number of pixels of the last frame which were assigned to each of Normal distributions accounting for human body, background and grey zone.



CHAPTER 2

INITIALIZATION

2.1 Introduction

In the first frame of the sequence we have a $r \times c$ matrix of continuous data (temperatures) and we need to derive a method which will split this image in three (non-overlapping and possibly non continuous) areas corresponding to skin, grey zone and background. In other words we need to cluster the 'similar' pixels in three groups, or to put it differently, we need to classify each pixel in one of the three clusters. This can be done in several ways. In what follows we will present two different cluster analysis techniques that can be employed here: the k-means and the EM method. We will provide the theory for these methods and then we will apply the two algorithms in the short thermal video data. The comparison of the two methods on these data will conclude this chapter.

2.2 K-means

K-means method assumes that the number (k) of clusters is known a priori. Although this is generally quite restrictive, in our problem we have already decided to use three clusters, so this is not an issue here.



2.2.1 The method.

The method operates iteratively. It starts using k observations as centroids of the clusters and then calculates the distance of each observation from the k centroids. Each observation will be assigned to the centroid, from which the distance is minimum. When all the observations have been assigned to one of the k centroids, i.e. the k initial clusters have been created; the new centroid of each cluster is calculated as the mean value of cluster's observations. The same procedure is repeated until there is no difference in the centroids between two iterations.

2.2.2 The distance in k-means method.

In order to calculate the distance between two observations here we will just mention, that there are several measures of distance that can be used. The most widely used is the Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2} \quad (2.2.2.1)$$

where, $x = (x_1, x_2, \dots, x_p)$, $y = (y_1, y_2, \dots, y_p)$ the two observations.

2.2.3 The initial centroids.

There are several algorithms that can be used to calculate the initial centroids. The choice of one of them depends on the nature of the problem and the data. Software packages have automated procedures for selecting the initial centroids.

The free choice of initial centroids is a major disadvantage of the method. Different choices can lead to totally different clusters. This means that choosing different points as initial centroids, one can find completely different solutions at the end of the algorithm. The best thing one can do is to test several observations as initial centroids and choose those ones that finally



give the “best” cluster representation of the data. The general philosophy in the literature is that the initial centroids have to be chosen to be as far from each other as possible.

2.2.4 Algorithm’s termination.

Another point of concern is the algorithm’s terminating condition. Or in other words when it can be said that there is no much difference between two iterations? Again, the answer depends on each case’s needs. Usually a “small” tolerance number will be defined and we terminate the iterations when all successive centroid differences are smaller than this tolerance number.

2.2.5 The algorithm’s rate of convergence.

K-means method is a fast algorithm and usually only a few iterations are needed. When appropriate initial centroids have been selected, the clusters created from the first iterations are very similar to the final solution; the differences exist only to those few observations which are among two clusters. So, there is no need for a large number of iterations. This property makes k-means to be appealing when we have large data sets.

2.2.6 Other properties.

Another property of the algorithm according to Karlis (2004) is that the clusters created contain approximately the same number of observations. Moreover, it should be noted that the clusters are always convex, non overlapping sets.



2.3 EM algorithm

In our problem, the results provided by k-means method were good enough, but not exciting. That is because with this method each pixel is committed to one particular of the three groups. We would prefer to assign the observations with probability to a group and then define some threshold which will determine the status of a pixel. If we account three Normal distributions, $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$ for the three groups of pixels (human body, background and grey zone), each data point should ideally be partially committed to all of the existing distributions. The level of each distribution's commitment should be described by appropriate weighting factors. In other words, each pixel x_j of the first frame will be considered as a mixture of the three Normal distributions:

$$x_j \sim w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + w_3 N(\mu_3, \sigma_3^2)$$

where, $w_i \geq 0$ and $\sum w_i = 1$.

In order to calculate the distributions' parameters, which will be represented from now on by the vector $\phi = (w_1, w_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T$, we will use the EM algorithm. EM algorithm for this particular application has been formulated in the past by Hasseblad (1966, 1969). A short description of this method is as follows.

2.3.1 The method.

EM algorithm is an iterative algorithm, which was originally suggested by Dempster, Laird and Rubin (1977) and it was proposed for the computation of maximum likelihood estimates when missing cases exist. Many additional details and alternatives are discussed by McLachlan and Krishnahn (1997). It proceeds iteratively in two steps; the Expectation or E-step and the Maximization or M-step (it is due to those two steps that the algorithm is named EM algorithm). The EM method uses the observed data to obtain a value of the estimate that maximizes the log likelihood



$L_C(\phi) = \sum_{i=1}^n \log \left(\sum_{j=1}^k w_j f(x_i / \phi_j) \right)$, where $f(x_i / \phi_j)$ is the probability density of the data and with $\phi = (w_1, \dots, w_k, \phi_1, \dots, \phi_k)$ we symbolize the parameters we have to estimate.

Although it could be claimed that it is just a numerical maximization method, the EM algorithm has many applications mainly because of its amazing statistical interpretation and the simplification of problems it provides.

The EM algorithm is applied in problems of missing data or when we can express the problem as if missing data exist. In our case, which is a problem of mixtures of distributions, we treat as missing variables the indicator variables z_{ij} , ($i=1,2,3$ and $j=1, \dots, rc$) which are the probabilities the j -th pixel to belong to the i -th group. (We use subscript $i=1$ for the background, $i=2$ for the grey zone, and $i=3$ for the skin distribution).

Using some initial values of ϕ , say $\phi^{(0)}$, the algorithm's first step (E-step) requires the calculation of the expectation of the complete data log likelihood, $L_C(\phi)$ which has been given above, conditional on the observed data and the initial values. McLachlan and Basford (1988), point out that this step is affected simply by replacing each indicator variable z_{ij} by its expectation conditional on x_j . That is, z_{ij} is replaced by the initial estimate of the posterior probability that j -th pixel belongs to i -th group. In our case, we provide the k-means estimates of the parameters $w_i^{(0)}$, $\mu_i^{(0)}$, $\sigma_i^{(0)}$, $i=1,2,3$ as starting points of the algorithm and the E-step of algorithm in the k -th iteration calculates

$$z_{ij}^{(k)} = \frac{w_i^{(k)} (\sigma_i^{(k)})^{-1} \exp \left\{ -\frac{1}{2(\sigma_i^{(k)})^2} (x_j - \mu_i^{(k)})^2 \right\}}{\sum_{i=1}^3 w_i^{(k)} (\sigma_i^{(k)})^{-1} \exp \left\{ -\frac{1}{2(\sigma_i^{(k)})^2} (x_j - \mu_i^{(k)})^2 \right\}}$$

i.e. the probability that the j -th pixel belongs to the i -th group where,

$$0 \leq z_{ij}^{(k)} \leq 1 \text{ and } \sum_{i=1}^3 z_{ij}^{(k)} = 1$$

On the M-step, the intention is to choose the value of ϕ , that maximizes the expectation of the complete data log likelihood $L_c(\phi) = \sum_{i=1}^n \log \left(\sum_{j=1}^k w_j f(x_i / \phi_j) \right)$. One nice feature of EM algorithm in the problem we examine is that the solution to the M step exists in closed forms. In our case, M-step calculates the new estimators. More precisely:

$$w_i^{(k+1)} = \frac{\sum_{j=1}^{rc} z_{ij}^{(k)}}{rc}$$

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^{rc} z_{ij}^{(k)} x_j}{rc w_i^{(k+1)}}$$

$$(\sigma_i^{(k+1)})^2 = \frac{\sum_{j=1}^{rc} z_{ij}^{(k)} (x_j - \mu_i^{(k+1)})^2}{rc w_i^{(k+1)}}.$$

E and M steps succeed each other until some convergence criterion is satisfied.

2.3.2 Lack of progress.

Two kinds of criteria are usually used. The first one stops the iterations when the relative increase of log likelihood between two successive iterations is smaller than a crucial value of tolerance (tol_1). This criterion has the form:

$$\left| \frac{L(\phi^{(r+1)}) - L(\phi^{(r)})}{L(\phi^{(r+1)})} \right| \leq tol_1$$

Where $L(\phi^{(r)})$ is the log likelihood after r iterations. The other type criterion stops the algorithm when the parameters do not change much between two successive iterations. These criteria have the form:

$$\max \left(\left| \phi^{(r+1)} - \phi^{(r)} \right| \right) \leq tol_2,$$

where the $\max(\)$ function returns the maximum coordinate of a vector.



In both cases the algorithm stops when there is not much change from one iteration to the next. In our case we use partly the second type of criterion, where the terminating condition was set to be:

$$|w_i^{(k+1)} - w_i^{(k)}| \leq 10^{-10}$$

for $i=1, 2, 3$.

2.3.3 Characteristics.

It should be noted at this point the importance of the fact that z_{ij} presents the probability that the j -th pixel belongs to the i -th group. When the algorithm converges, those probabilities are available and are in fact those ones that can be used to classify each observation to that cluster in which it belongs with the higher probability. One of the advantages of EM over k-means method (probably the most important one), is that each observation belongs to all clusters with some probability. That is why the clusters created by EM may overlap, a fact that is impossible in k-means algorithm.

McLachlan, (1992) claims that another nice feature of EM algorithm is that the log likelihood for the incomplete data specification is non-decreasing from one iteration to the next. This means that it can never be decreased after an EM iteration. i.e. $L(\phi^{(k+1)}) \geq L(\phi^{(k)})$. However, we have to be careful because in some cases there is the danger of trapping in a local maximum rather than a global one.

Although the programming of this algorithm is quite easy and mainly, can be done in any statistical software, however, EM algorithm has the disadvantage to be quite slow as a method. That is because convergence is quite slow. McLachlan and Krishnahn (1997) discuss about speeding up the convergence of the algorithm. We will just mention that one simple way to achieve faster convergence is to provide the algorithm with the most suitable initial values possible. So the speed of convergence is affected, among others, and by the choice of initial values. However, we should mention here that there still remains the danger of trapping in a local maximum. In our example, we have provided the algorithm with the k-means estimates of the parameters of interest.



2.4 Example

In our example we have one subject (more precisely a woman) monitored with an infrared camera. The analysis of each frame of the video is 254×318 pixels. In figure 2.4.1 we provide a thermal image of a frame with temperatures.



Figure 2.4.1. Thermal image of the first frame's temperatures.

We can see in figure 2.4.1 that the background pixels have, as expected, lower temperatures than the skin pixels. One interesting thing to note is that the pixels denoting the nose of the subject have relatively low temperatures, similar to those of the background pixels. This happens because the breathing function causes our nose to have similar to the environment temperatures.

Next, in figure 2.4.2 we present a histogram of the pixel temperature values referring to the frame of figure 2.4.1.

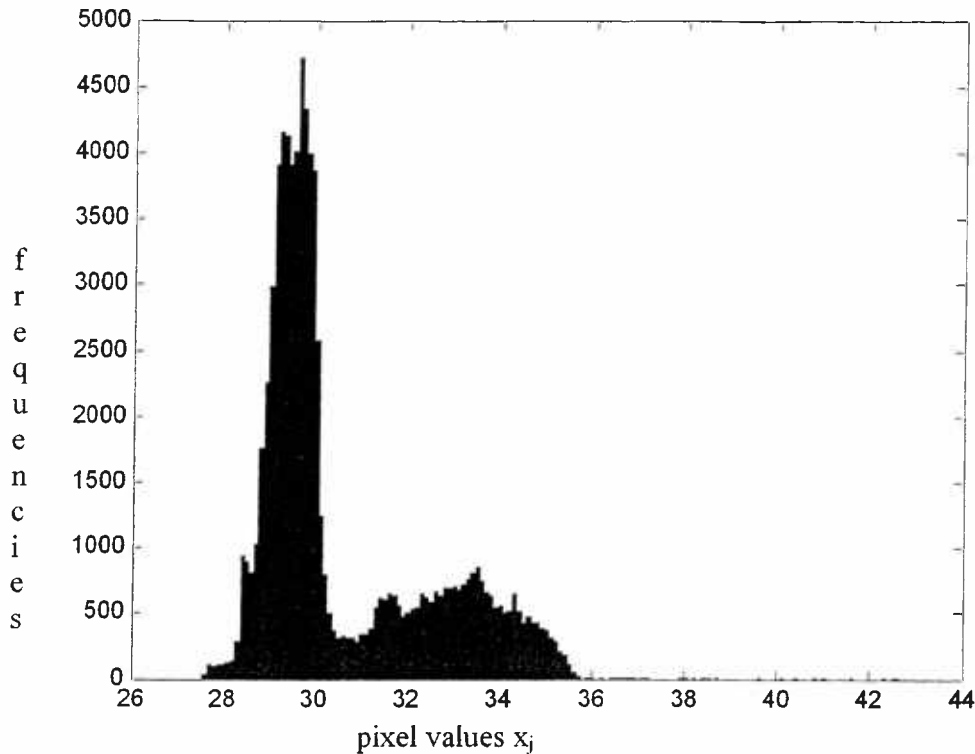


Figure 2.4.2. A histogram of the pixel temperature values.

From several histograms studied, it appears to have three modes. In figure 2.4.2 one mode is in the temperature of approximately 29, one a bit lower of 32 and one close to 34. The three modes of this histogram resemble the form of a mixture of three Normal distributions. This fact led us to account for three Normal distributions; one on the lower band of temperatures which accounts for background pixels, one in the upper temperature values to represent the human body pixels and one for the 'grey zone'. The last group contains pixels that have higher temperatures than background but lower than human body pixels.

The resulting figure of the initial frame segmentation to skin, background and grey zone pixels can be seen in figure 2.4.3.



Figure 2.4.3. The classification of first frame's pixels through k-means method. Blue colour denotes the pixels ascribed to background, green colour denotes the pixels ascribed to skin and red colour denotes the pixels ascribed to grey zone.

Clearly, most of the subject's body pixels have been correctly committed to the 'body group'. However, there are some pixels on the nose that have been committed to the background group and some pixels around the nose, on the right eye, on cheeks and on the chin that have been committed to the grey zone group. This misclassification is easily explained by the fact that in those areas of body, the temperature is quite low; thus those pixels could not be ascribed to the group of body pixels that had higher temperatures. Another misclassification has happened to those pixels of skin on the shoulders, which are in fact pixels of skin covered by clothes. However those pixels have not been ascribed to grey-zone group but to body group. The explanation of this lies on the fact that being spring time the subject is dressed with light clothes and in this area the cloth touches the skin causing the high temperature value of covered skin pixels.

Although the classification of pixels obtained by k-means is satisfactory enough, in our problem, as already mentioned, we prefer each

pixel to be committed not to one particular group, but partially committed to all of the existing groups.

That was the reason why we then decided to apply EM method. We first estimated the weights, mean values and variances based on the groups created by k-means method and then provided EM algorithm with those values, i.e. $w_1^{(0)} = 0.6726$, $w_2^{(0)} = 0.1572$, $w_3^{(0)} = 0.1702$, $\mu_1^{(0)} = 29.4754$, $\mu_2^{(0)} = 31.9259$, $\mu_3^{(0)} = 33.7167$, $\sigma_1^{(0)} = 0.6173$, $\sigma_2^{(0)} = 1.1079$, $\sigma_3^{(0)} = 1.1243$.

The new values of the parameters of interest obtained by the EM algorithm are: $w_1^{(k)} = 0.6815$, $w_2^{(k)} = 0.0605$, $w_3^{(k)} = 0.2580$, $\mu_1^{(k)} = 29.4366$, $\mu_2^{(k)} = 31.5141$, $\mu_3^{(k)} = 33.3902$, $\sigma_1^{(k)} = 0.4918$, $\sigma_2^{(k)} = 0.4337$, $\sigma_3^{(k)} = 1.0539$ and the classification of pixels to that group for which the probability z_{ij} was maximum is shown in figure 2.4.4.



Figure 2.4.4. The classification of first frame's pixels through EM algorithm. Blue colour denotes the pixels ascribed to background, green colour denotes the pixels ascribed to skin and red colour denotes the pixels ascribed to grey zone.

We can see in figure 2.4.4 that the misclassification happens in less pixels comparing with the misclassifications in k-means method.

The differences between the results of the above methods are given in figure 2.4.5:



Figure 2.4.5. The differences between the classification of first frame's pixels through k-means and EM algorithm. Blue colour denotes the pixels that have been ascribed to the same group by both methods, green colour denotes the pixels ascribed to grey zone by k-means but to skin by EM, and red colour denotes the pixels ascribed to grey zone by k-means but to background by EM.

It is clear from the last picture that the differences between the two methods exist mainly to those pixels that have been committed to the 'grey zone' cluster by k-means methods, whereas EM placed them to the 'skin' group. We can observe that some of those pixels are clearly skin, but some others are skin covered by clothes. The disagreement of the two methods on classification of those pixels is justified because of their nature to lie between the two major components characterized by background and skin.

The goal of the initialization phase is to provide statistically valid values for the temperatures of first frame's pixels corresponding to the scene.

These values will then be used as starting points for the process of awareness of next frames' pixels.

2.5 Spatial aware classification scheme

In the previous section we classified each pixel to the group with the highest probability. Alternatively, we could have defined a threshold T (like $T = 0.6, 0.7, 0.8$) and we would have classified only if $\max_{i=1,2,3} \{z_{ij}\} > T$. Otherwise we would define a specific neighborhood of the pixel and study what happened there. The pixel would be finally ascribed to that group to which most of the pixels in the neighborhood were ascribed.

However, as we are interested in updating on real time we only do rough classification and we don't take under consideration what happens in the neighborhood of pixels. We will simply classify each pixel to that group in which it belongs with the higher probability z_{ij} .





CHAPTER 3

THE MATCHING OPERATION

3.1 Introduction

We have ascribed first frame's pixels to three Normal distributions. In other words, each pixel x_j of the first frame is considered to be a mixture of the three Normal distributions:

$$x_j \sim w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + w_3 N(\mu_3, \sigma_3^2)$$

Next phase is concerned with the classification of next incoming frames' pixels. In this chapter we will describe a method to ascribe the pixel values of an incoming camera frame to one of three Normal distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ or $N(\mu_3, \sigma_3^2)$.

We assume that every new pixel comes from a Normal distribution $N(\mu_p, \sigma_p^2)$, where μ_p is the actual value of the pixel's temperature and σ_p^2 is a value related to the camera's accuracy.

Thus, the issue is to find a way to measure how close is the distribution of the incoming pixel $N(\mu_p, \sigma_p^2)$ to the three existing distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$. Once we have measured the 'distances' between those distributions, the pixel will be ascribed to that Normal from which it desist less.



3.2 Measures of distance

Distance measures are being used to measure how close two probability distributions are to each other, or in other words how easy is to distinguish between the two distributions. According to Ali and Silvey (1966), these measures have been called in many different ways in literature; measures of distance (Adhikari and Joshi, 1956), measures of separation (Rao, 1952), measures of discriminatory information (Chernoff, 1952; Lehmann, 1959), measures of variation-distance (Kolmogorov, 1963), coefficients of divergence (Kullback, 1959).

It should be noted at this point that in topology in order to characterize a measure $d(\cdot, \cdot)$ between two distributions f_1, f_2 as distance (or metric) it has to satisfy the following three properties:

- (1) $d(f_1, f_2) = 0$ if and only if $f_1 \equiv f_2$
- (2) $d(f_1, f_2) = d(f_2, f_1)$ (the symmetric property of distance)
- (3) $d(f_1, f_3) \leq d(f_1, f_2) + d(f_2, f_3)$ (the triangular inequality)

Corollary

If a measure between two distributions $d(f_1, f_2)$ is a distance (or metric) it satisfies the Shannon's inequality: $d(f_1, f_2) \geq 0$

Proof

Subsequently from the third property and by setting $f_3 \equiv f_1$ we take:

$$d(f_1, f_1) \leq d(f_1, f_2) + d(f_2, f_1)$$

But, by the first property we have $d(f_1, f_1) = 0$ and by the second we have $d(f_1, f_2) = d(f_2, f_1)$. So, we will have that $d(f_1, f_2) \geq 0$.

Not all of the coefficients used in literature (mentioned below) to measure the discrepancy between two distributions satisfy the three conditions. Some of them satisfy only the first two and in this case are called divergences. There are some other measures, however, that do not satisfy the



symmetry property either. Here, all of them will be mentioned as distances, but not in the strict sense the word has in metric spaces.

The applications of these measures can be found in statistical inference, in analysis of contingency tables, in approximation of probability distributions, in pattern recognition, in signal processing etc.

Many known distance measures between probability distributions will be given, starting from some general classes of divergence coefficients and going to particular cases. Most of them have been found in literature and collected by Basseville (1988). Moreover, for some of the divergences which will be mentioned, we will provide these measures in closed forms when the distributions are 1-dimensional Gaussian distributions.

3.2.1 A review of distance measures.

- **Csiszar f divergence**

$$d(f_1, f_2) = g \left[E_1 \left[f \left(\frac{df_1}{df_2} \right) \right] \right] \quad (3.2.1.1)$$

where $\frac{df_1}{df_2} \triangleq \phi(x) = \frac{f_2(x)}{f_1(x)}$, i.e. the likelihood ratio, f : a continuous convex real function on R_+ , g an increasing function on R and E_1 the expectation with respect to f_1 .

This class of distance measures has been introduced independently by Ali and Silvey (1966) and by Csiszar (1967a and 1967b) and its many properties were studied by Vajda (1972). For different functions f and g it gives a number of very widely used distance measures.

It should be noted, however, that this class of distance measures satisfy the first property of a metric, i.e. $d(f_1, f_1) = 0$ only if $g[E_1[f(1)]] = 0$. For this

reason (3.2.1.1) can be written as: $d(f_1, f_2) = g \left[E_1 \left[f \left(\frac{df_1}{df_2} \right) \right] \right] - g[E_1[f(1)]]$.



- **Alpha divergence measure (of fractional order $\alpha \in [0,1]$)**

$$D_\alpha(f_1, f_2) = \frac{1}{\alpha-1} \ln \int f_2 \left(\frac{f_1}{f_2} \right)^\alpha dz = \frac{1}{\alpha-1} \ln \int f_1^\alpha(z) f_2^{1-\alpha}(z) dz \quad (3.2.1.2)$$

The class of alpha-divergences is also known as Renyi divergence and leads to different measures by selection of fractional order. Hero et al. (2001) claims that when the distributions are very similar, i.e. when it is difficult to discriminate between them, the optimal choice of α is $\alpha = \frac{1}{2}$ which corresponds to Hellinger affinity $D_{\frac{1}{2}}(f_1, f_2) = 2 \ln \int \sqrt{f_1(x) f_2(x)} dx$. The performance of this member of the class is, in this case, even better than that of Kullback Leibler divergence, which is $D_{\alpha \rightarrow 1}(f_1, f_2)$.

Lemma: The alpha divergence measure when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:

$$D_\alpha(f_1, f_2) = \frac{1/2}{\alpha-1} \ln \frac{(\sigma_1^2)^{1-\alpha} (\sigma_2^2)^\alpha}{\alpha \sigma_1^2 + (1-\alpha) \sigma_2^2} + \frac{\alpha}{2} \frac{(\mu_1 - \mu_2)^2}{(1-\alpha) \sigma_1^2 + \alpha \sigma_2^2}$$

Proof

$$D_\alpha(f_1, f_2) = \frac{1}{\alpha-1} \ln \int f_2 \left(\frac{f_1}{f_2} \right)^\alpha dz = \frac{1}{\alpha-1} \ln \int f_1^\alpha(z) f_2^{1-\alpha}(z) dz$$

If f_1, f_2 Gaussians $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$, then:

$$\begin{aligned} \int f_1^\alpha(x) f_2^{1-\alpha}(x) dx &= \\ &= \int \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right)^\alpha \exp \left\{ -\frac{\alpha(x-\mu_1)^2}{2\sigma_1^2} \right\} \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \right)^{1-\alpha} \exp \left\{ -\frac{(1-\alpha)(x-\mu_2)^2}{2\sigma_2^2} \right\} dx = \\ &= \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right)^\alpha \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \right)^{1-\alpha} \int \exp \left\{ -\frac{\alpha(x-\mu_1)^2 \sigma_2^2 + (1-\alpha)(x-\mu_2)^2 \sigma_1^2}{2\sigma_1^2 \sigma_2^2} \right\} dx = \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi(\sigma_1^2)^\alpha(\sigma_2^2)^{1-\alpha}}} \int \exp \left\{ -\frac{1}{2\sigma_1^2\sigma_2^2} \left[ax^2\sigma_2^2 - 2a\mu_1\sigma_2^2x + a\mu_1^2\sigma_2^2 + x^2\sigma_1^2 - 2\mu_2\sigma_1^2x + \right. \right. \\
&\quad \left. \left. + \mu_2^2\sigma_1^2 - a\sigma_1^2x^2 + 2a\mu_2\sigma_1^2x - a\mu_2^2\sigma_1^2 \right] \right\} dx \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2)^\alpha(\sigma_2^2)^{1-\alpha}}} \int \exp \left\{ -\frac{1}{2\sigma_1^2\sigma_2^2} \left[(a\sigma_2^2 + \sigma_1^2 - a\sigma_1^2)x^2 - (2a\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2 - 2a\mu_2\sigma_1^2)x + \right. \right. \\
&\quad \left. \left. + a\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 - a\mu_2^2\sigma_1^2 \right] \right\} dx \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2)^\alpha(\sigma_2^2)^{1-\alpha}}} \int \exp \left\{ -\frac{a\sigma_2^2 + (1-a)\sigma_1^2}{2\sigma_1^2\sigma_2^2} \left[x^2 - 2\frac{a\mu_1\sigma_2^2 + (1-a)\mu_2\sigma_1^2}{a\sigma_2^2 + (1-a)\sigma_1^2}x + \left(\frac{a\mu_1\sigma_2^2 + (1-a)\mu_2\sigma_1^2}{a\sigma_2^2 + (1-a)\sigma_1^2} \right)^2 \right] \right. \\
&\quad \left. * \exp \left\{ \frac{(a\mu_1\sigma_2^2 + (1-a)\mu_2\sigma_1^2)^2}{2\sigma_1^2\sigma_2^2(a\sigma_2^2 + (1-a)\sigma_1^2)} - \frac{a\mu_1^2\sigma_2^2 + (1-a)\mu_2^2\sigma_1^2}{2\sigma_1^2\sigma_2^2} \right\} \right\} dx = \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2)^\alpha(\sigma_2^2)^{1-\alpha}}} \int \exp \left\{ -\frac{\left(x - \frac{a\mu_1\sigma_2^2 + (1-a)\mu_2\sigma_1^2}{a\sigma_2^2 + (1-a)\sigma_1^2} \right)^2}{\frac{2\sigma_1^2\sigma_2^2}{a\sigma_2^2 + (1-a)\sigma_1^2}} \right\} * \\
&\quad * \exp \left\{ -\frac{a(1-a)\sigma_1^2\sigma_2^2(\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2)}{2\sigma_1^2\sigma_2^2(a\sigma_2^2 + (1-a)\sigma_1^2)} \right\} dx = \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2)^\alpha(\sigma_2^2)^{1-\alpha}}} \sqrt{2\pi \frac{\sigma_1^2\sigma_2^2}{a\sigma_2^2 + (1-a)\sigma_1^2}} \cdot \exp \left\{ -\frac{a(1-a)\sigma_1^2\sigma_2^2(\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2)}{2\sigma_1^2\sigma_2^2(a\sigma_2^2 + (1-a)\sigma_1^2)} \right\} = \\
&\quad * \exp \left\{ -\frac{a(1-a)\sigma_1^2\sigma_2^2(\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2)}{2\sigma_1^2\sigma_2^2(a\sigma_2^2 + (1-a)\sigma_1^2)} \right\} dx = \\
&= \sqrt{\frac{(\sigma_1^2)^{1-\alpha}(\sigma_2^2)^\alpha}{a\sigma_2^2 + (1-a)\sigma_1^2}} \exp \left\{ -\frac{a(1-a)}{2} \cdot \frac{(\mu_1 - \mu_2)^2}{a\sigma_2^2 + (1-a)\sigma_1^2} \right\}
\end{aligned}$$

$$\text{So: } D_\alpha(f_1, f_2) = \frac{1/2}{\alpha - 1} \ln \frac{(\sigma_1^2)^{1-\alpha} (\sigma_2^2)^\alpha}{\alpha \sigma_1^2 + (1-\alpha) \sigma_2^2} + \frac{\alpha}{2} \frac{(\mu_1 - \mu_2)^2}{(1-\alpha) \sigma_1^2 + \alpha \sigma_2^2}$$

- **Un-normalized α -divergence measure**

The un-normalized α -divergence is defined according to Hero et al. (2001) as a multiple of the alpha divergence measure.

$$D_\alpha^u(f_1, f_2) = -\ln \int f_1^\alpha(z) f_2^{1-\alpha}(z) dz = (1-\alpha) D_\alpha(f_1, f_2) \quad (3.2.1.3)$$

This class of divergences is found in literature as Chernoff distances, while the integral $\int f_1^\alpha(z) f_2^{1-\alpha}(z) dz$ is found to be named Chernoff coefficient.

Un-normalized α -divergence measures are obtained by (3.2.1.1) by setting $f(x) = -x^{1-\alpha}$, $0 \leq \alpha \leq 1$ and $g(x) = -\log x$ in the Csiszar's divergence.

Lemma: The un-normalized alpha divergence measure when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:

$$D_\alpha^u(f_1, f_2) = (1-\alpha) D_\alpha(f_1, f_2) = -\frac{1}{2} \ln \frac{(\sigma_1^2)^{1-\alpha} (\sigma_2^2)^\alpha}{\alpha \sigma_1^2 + (1-\alpha) \sigma_2^2} + \frac{\alpha(1-\alpha)}{2} \frac{(\mu_1 - \mu_2)^2}{(1-\alpha) \sigma_1^2 + \alpha \sigma_2^2}$$

- **Amari α -divergence**

$$A^\alpha(f_1, f_2) = \frac{4}{1-\alpha^2} \int \frac{1-\alpha}{2} f_1(x) + \frac{1+\alpha}{2} f_2(x) - [f_1(x)]^{\frac{1-\alpha}{2}} [f_2(x)]^{\frac{1+\alpha}{2}} dx, \quad \alpha \in R \quad (3.2.1.4)$$

This family of parametric divergence functions was introduced, and investigated by Amari (1982, 1985). It is sometimes found in literature as α -divergences, as well (Zhang (2004)).

Lemma: The Amari α -divergence when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:



$$A^\alpha(f_1, f_2) = \frac{4}{1-\alpha^2} \left[1 - \sqrt{\frac{2(\sigma_1^2)^{\frac{1+\alpha}{2}} (\sigma_2^2)^{\frac{1-\alpha}{2}}}{(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2}} \exp \left\{ -\frac{1-\alpha^2}{4} \frac{(\mu_1 - \mu_2)^2}{(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2} \right\} \right]$$

Proof

$$A^\alpha(f_1, f_2) = \frac{4}{1-\alpha^2} \cdot \frac{1-\alpha}{2} \int f_1(x) dx + \frac{4}{1-\alpha^2} \cdot \frac{1+\alpha}{2} \int f_2(x) dx - \frac{4}{1-\alpha^2} \int f_1^{\frac{1-\alpha}{2}}(x) f_2^{\frac{1+\alpha}{2}}(x) dx$$

or:

$$A^\alpha(f_1, f_2) = \frac{2}{1+\alpha} + \frac{2}{1-\alpha} - \frac{4}{1-\alpha^2} \int f_1^{\frac{1-\alpha}{2}}(x) f_2^{\frac{1+\alpha}{2}}(x) dx$$

or:

$$A^\alpha(f_1, f_2) = \frac{4}{1-\alpha^2} \left(1 - \int f_1^{\frac{1-\alpha}{2}}(x) f_2^{\frac{1+\alpha}{2}}(x) dx \right)$$

But:

$$\begin{aligned} \int f_1^{\frac{1-\alpha}{2}}(x) f_2^{\frac{1+\alpha}{2}}(x) dx &= \\ &= \int \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right)^{\frac{1-\alpha}{2}} \exp \left\{ -\frac{(1-\alpha)(x-\mu_1)^2}{4\sigma_1^2} \right\} \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \right)^{\frac{1+\alpha}{2}} \exp \left\{ -\frac{(1+\alpha)(x-\mu_2)^2}{4\sigma_2^2} \right\} dx = \\ &= \frac{1}{\sqrt{2\pi} \left(\sqrt{\sigma_1^2} \right)^{\frac{1-\alpha}{2}} \left(\sqrt{\sigma_2^2} \right)^{\frac{1+\alpha}{2}}} \int \exp \left\{ -\frac{(1-\alpha)(x-\mu_1)^2 \sigma_2^2 + (1+\alpha)(x-\mu_2)^2 \sigma_1^2}{4\sigma_1^2 \sigma_2^2} \right\} dx = \\ &= \frac{1}{\sqrt{2\pi} \left(\sqrt{\sigma_1^2} \right)^{\frac{1-\alpha}{2}} \left(\sqrt{\sigma_2^2} \right)^{\frac{1+\alpha}{2}}} \int \exp \left\{ -\frac{(1-\alpha)(x^2 - 2\mu_1 x + \mu_1^2) \sigma_2^2 + (1+\alpha)(x^2 - 2\mu_2 x + \mu_2^2) \sigma_1^2}{4\sigma_1^2 \sigma_2^2} \right\} dx = \\ &= \frac{1}{\sqrt{2\pi} \left(\sqrt{\sigma_1^2} \right)^{\frac{1-\alpha}{2}} \left(\sqrt{\sigma_2^2} \right)^{\frac{1+\alpha}{2}}} \int \exp \left\{ -\frac{[(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2]x^2 - 2[(1-\alpha)\mu_1\sigma_2^2 + (1+\alpha)\mu_2\sigma_1^2]x + [(1-\alpha)\mu_1^2\sigma_2^2 + (1+\alpha)\mu_2^2\sigma_1^2]}{4\sigma_1^2 \sigma_2^2} \right\} dx \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi} \left(\sqrt{\sigma_1^2}\right)^{\frac{1-a}{2}} \left(\sqrt{\sigma_2^2}\right)^{\frac{1+a}{2}}} \\
&\int \exp \left\{ -\frac{(1-a)\sigma_2^2 + (1+a)\sigma_1^2}{4\sigma_1^2\sigma_2^2} \left[x^2 - 2\frac{(1-a)\mu_1\sigma_2^2 + (1+a)\mu_2\sigma_1^2}{(1-a)\sigma_2^2 + (1+a)\sigma_1^2} x + \left(\frac{(1-a)\mu_1\sigma_2^2 + (1+a)\mu_2\sigma_1^2}{(1-a)\sigma_2^2 + (1+a)\sigma_1^2} \right)^2 \right] \right. \\
&\quad \cdot \exp \left\{ \frac{\left[(1-a)\mu_1\sigma_2^2 + (1+a)\mu_2\sigma_1^2 \right]^2}{\left[(1-a)\sigma_2^2 + (1+a)\sigma_1^2 \right] 4\sigma_1^2\sigma_2^2} - \frac{\left[(1-a)\mu_1^2\sigma_2^2 + (1+a)\mu_2^2\sigma_1^2 \right]}{4\sigma_1^2\sigma_2^2} \right\} dx = \\
&= \frac{1}{\sqrt{2\pi} \left(\sqrt{\sigma_1^2}\right)^{\frac{1-a}{2}} \left(\sqrt{\sigma_2^2}\right)^{\frac{1+a}{2}}} \cdot \int \exp \left\{ -\frac{\left(x - \frac{(1-a)\mu_1\sigma_2^2 + (1+a)\mu_2\sigma_1^2}{(1-a)\sigma_2^2 + (1+a)\sigma_1^2} \right)^2}{2\frac{\sigma_1^2\sigma_2^2}{(1-a)\sigma_2^2 + (1+a)\sigma_1^2}} \right\} * \\
&* \exp \left\{ -\frac{(1-\alpha^2)\sigma_1^2\sigma_2^2(\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2)}{4\sigma_1^2\sigma_2^2[(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2]} \right\} dx = \\
&= \frac{1}{\sqrt{2\pi} \left(\sqrt{\sigma_1^2}\right)^{\frac{1-a}{2}} \left(\sqrt{\sigma_2^2}\right)^{\frac{1+a}{2}}} \cdot \sqrt{2\pi \frac{2\sigma_1^2\sigma_2^2}{(1-a)\sigma_2^2 + (1+a)\sigma_1^2}} \cdot \exp \left\{ -\frac{(1-\alpha^2)\sigma_1^2\sigma_2^2(\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2)}{4\sigma_1^2\sigma_2^2[(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2]} \right\} = \\
&= \sqrt{\frac{2(\sigma_1^2)^{\frac{1+\alpha}{2}} (\sigma_2^2)^{\frac{1-\alpha}{2}}}{(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2}} \exp \left\{ -\frac{1-\alpha^2}{4} \frac{(\mu_1 - \mu_2)^2}{(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2} \right\}
\end{aligned}$$

So:

$$A^\alpha(f_1, f_2) = \frac{4}{1-\alpha^2} \left[1 - \sqrt{\frac{2(\sigma_1^2)^{\frac{1+\alpha}{2}} (\sigma_2^2)^{\frac{1-\alpha}{2}}}{(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2}} \exp \left\{ -\frac{1-\alpha^2}{4} \frac{(\mu_1 - \mu_2)^2}{(1-\alpha)\sigma_2^2 + (1+\alpha)\sigma_1^2} \right\} \right]$$

- **Hellinger distance**

$$H_p(f_1, f_2) = \sqrt[p]{\int (f_1^{1/p} - f_2^{1/p})^p} \quad (3.2.1.5)$$

The family of Hellinger distances satisfies all of the conditions required for a measure to be characterized as distance. It is a Csiszar f divergence with $f(x) = |1 - x^{1/p}|^p$ and $g(x) = x^{1/p}$, $p \geq 1$.

This family is sometimes called “Generalized Matusita distance” and denoted by $M_p(f_1, f_2)$.

It gives many different distance measures for various values of p . The most widely used of them are:

for $p=1$: $H_1(f_1, f_2) = \int |f_1 - f_2|$ and

for $p=2$: $H_2(f_1, f_2) = \left(\int (\sqrt{f_1} - \sqrt{f_2})^2 \right)^{\frac{1}{2}}$

In literature the term “Hellinger distance” is usually referred to H_2^2 , which is, according to Basseville (1988), a Csiszar f divergence; (set in (3.2.1.1) $f(x) = (\sqrt{x} - 1)^2$ and $g(x) = x$).

Furthermore, Hero et al. (2001) related this Hellinger distance (H_2^2) to Hellinger affinity $D_{\frac{1}{2}}(f_1, f_2) = 2 \ln \int \sqrt{f_1(x)f_2(x)} dx$, which is an Alpha divergence measure. He first proved that:

$$H_2^2(f_1, f_2) = \int (\sqrt{f_1(x)} - \sqrt{f_2(x)})^2 dx = 2 \left(1 - \exp \left(\frac{1}{2} D_{\frac{1}{2}}(f_1, f_2) \right) \right).$$

H_2 is sometimes called Matusita distance and being symbolized by: $M_2(f_1, f_2)$.

Lemma: The Hellinger distance when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:

$$H_2(f_1, f_2) = \sqrt{2 - 2\sqrt{2} \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2} \exp \left\{ -\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)} \right\}}$$

Proof

$$H_2(f_1, f_2) = \left(\int (\sqrt{f_1(x)} - \sqrt{f_2(x)})^2 \right)^{\frac{1}{2}} = \sqrt{\int f_1(x) dx + \int f_2(x) dx - 2 \int \sqrt{f_1(x)f_2(x)} dx =}$$

$$= \sqrt{2 - 2 \int \sqrt{f_1(x)f_2(x)} dx}$$

$$\int \sqrt{f_1(x)f_2(x)} dx = \int \sqrt{\frac{1}{\sqrt{2\pi\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \cdot \exp\left\{-\frac{(x-\mu_1^2)}{2\sigma_1^2} - \frac{(x-\mu_2^2)}{2\sigma_2^2}\right\}} dx =$$

$$= \int \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \left[\exp\left\{-\frac{x^2 - 2\mu_1x + \mu_1^2}{2\sigma_1^2} - \frac{x^2 - 2\mu_2x + \mu_2^2}{2\sigma_2^2}\right\} \right] dx =$$

$$= \int \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \exp\left\{-\frac{(\sigma_1^2 + \sigma_2^2)x^2 - 2(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)x + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}{4\sigma_1^2\sigma_2^2}\right\} dx =$$

$$= \exp\left\{\frac{(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)^2}{4\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)} - \frac{\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}{4\sigma_1^2\sigma_2^2}\right\} \int \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \exp\left\{-\frac{\left(\sqrt{\sigma_1^2 + \sigma_2^2}x - \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)^2}{4\sigma_1^2\sigma_2^2}\right\} dx =$$

$$= \exp\left\{\frac{2\mu_1\mu_2\sigma_1^2\sigma_2^2 - \mu_1^2\sigma_1^2\sigma_2^2 - \mu_2^2\sigma_1^2\sigma_2^2}{4\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)}\right\} \int \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \exp\left\{-\frac{(\sigma_1^2 + \sigma_2^2)\left(x - \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2}{4\sigma_1^2\sigma_2^2}\right\} dx =$$

$$= \exp\left\{-\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right\} \int \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \exp\left\{-\frac{\left(x - \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2}{\frac{4\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}}\right\} dx =$$

$$= \exp\left\{-\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right\} \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \int \frac{1}{\sqrt{2\pi \frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}}} \exp\left\{-\frac{\left(x - \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2}{2 \frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}}\right\} dx =$$

$$= \sqrt{2} \sqrt{\frac{\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left\{-\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right\}$$

So:
$$H_2(f_1, f_2) = \sqrt{2 - 2\sqrt{2} \sqrt{\frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left\{-\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right\}}$$

- **L_p distance**

$$L_p(f_1, f_2) = \left(\int |f_1 - f_2|^p \right)^{\frac{1}{p}}, \quad 0 < p < \infty \quad (3.2.1.6)$$

The family of L_p distances satisfies the first two of the distance properties, so it is a class of divergences. It gives different measures when one changes p . For example:

for $p=1$: $L_1(f_1, f_2) = \int |f_1 - f_2|$, which is a metric.

for $p=2$: $L_2(f_1, f_2) = \left(\int |f_1 - f_2|^2 \right)^{\frac{1}{2}}$, also found in literature as Patrick and Fisher distance.

Lemma: The L_2 distance when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:

$$L_2(f_1, f_2) = \sqrt{\frac{1}{\sqrt{4\pi\sigma_1^2}} + \frac{1}{\sqrt{4\pi\sigma_2^2}} - 2 \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \cdot \exp\left\{-\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right\}}$$

Proof

$$\begin{aligned} L_2(f_1, f_2) &= \left[\int |f_1(x) - f_2(x)|^2 dx \right]^{1/2} = \\ &= \sqrt{\int [f_1(x)]^2 dx + \int [f_2(x)]^2 dx - 2 \int f_1(x) f_2(x) dx} \end{aligned}$$

$$\begin{aligned} \int [f_1(x)]^2 dx &= \int \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right)^2 \left[\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \right]^2 dx = \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \int \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right) \exp\left\{-\frac{(x-\mu_1)^2}{\sigma_1^2}\right\} dx = \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \sqrt{\frac{1}{2}} \int \left(\frac{1}{\sqrt{2\pi\frac{1}{2}\sigma_1^2}} \right) \exp \left\{ -\frac{(x-\mu_1)^2}{2\frac{1}{2}\sigma_1^2} \right\} dx = \\
&= \frac{1}{\sqrt{4\pi\sigma_1^2}}
\end{aligned}$$

and similarly: $\int [f_2(x)]^2 dx = \frac{1}{\sqrt{4\pi\sigma_2^2}}$

$$\begin{aligned}
\int f_1(x)f_2(x)dx &= \int \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp \left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_2)^2}{2\sigma_2^2} \right\} dx = \\
&= \frac{1}{\sqrt{2\pi}} \int \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2}} \exp \left\{ -\frac{(\sigma_1^2 + \sigma_2^2)x^2 - 2(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)x + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}{2\sigma_1^2\sigma_2^2} \right\} dx = \\
&= \frac{1}{\sqrt{2\pi}} \exp \left\{ \frac{(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)} - \frac{\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}{2\sigma_1^2\sigma_2^2} \right\} \int \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2}} \exp \left\{ -\frac{\left(\sqrt{\sigma_1^2 + \sigma_2^2}x - \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \right)^2}{2\sigma_1^2\sigma_2^2} \right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \exp \left\{ \frac{2\mu_1\mu_2\sigma_1^2\sigma_2^2 - \mu_1^2\sigma_2^2\sigma_1^2 - \mu_2^2\sigma_1^2\sigma_2^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)} \right\} \int \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2}} \exp \left\{ -\frac{(\sigma_1^2 + \sigma_2^2) \left(x - \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \right)^2}{2\sigma_1^2\sigma_2^2} \right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \right\} \frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \int \frac{1}{\sqrt{2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}}} \exp \left\{ -\frac{\left(x - \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \right)^2}{\frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}} \right\} dx = \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp \left\{ -\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}
\end{aligned}$$

$$\text{So: } L_2(f_1, f_2) = \sqrt{\frac{1}{\sqrt{4\pi\sigma_1^2}} + \frac{1}{\sqrt{4\pi\sigma_2^2}} - 2\frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \cdot \exp\left\{-\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right\}}$$

- **Kolmogorov variational distance**

$$V(f_1, f_2) = \int |f_2(x) - f_1(x)| dx \quad (3.2.1.7)$$

It is a Csiszar f divergence with $f(x) = |1-x|$ and $g(x) = x$ and is distance as it satisfies all of the conditions required. $V(f_1, f_2)$ is also a Hellinger and L_p distance obtained by (3.2.1.5) and (3.2.1.6) setting $p=1$.

- **Kullback Leibler information number**

$$K(f_1, f_2) = \begin{cases} \int f_1 \log \frac{f_1}{f_2} & \text{if } f_1 \text{ is absolutely continuous with respect to } f_2 \\ \infty, & \text{otherwise} \end{cases} \quad (3.2.1.8)$$

This measure was introduced by Kullback and Leibler (1951). It is, according to Onishi and Imai (1997), the most fundamental divergence in information theory.

Kullback information number is non-negative, additive, but not symmetric, which means that it satisfies only the first out of the three conditions of a metric, so it is neither a divergence nor a distance.

It should also be noted that $K(f_1, f_2)$ is undefined if $f_2(x) = 0$ and $f_1(x) \neq 0$ for any x . This means that, according to Kullback (1967), distribution $f_1(x)$ has to be absolutely continuous with respect to $f_2(x)$ for $K(f_1, f_2)$ to be defined.

$K(f_1, f_2)$ has been characterized as a Csiszar f divergence; it can be obtained by (3.2.1.1) by setting $f(x) = -\log x$ and $g(x) = x$. According to Hero et al. (2001) it is also obtained by (3.2.1.2) when $\alpha \rightarrow 1$; i.e. is an alpha divergence measure.

Lemma: The Kullback-Leibler information number when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:

$$K(f_1, f_2) = \log \frac{\sigma_1}{\sigma_2} - \frac{1}{2} + \frac{\sigma_2^2}{2\sigma_1^2} + \frac{(\mu_1 - \mu_2)^2}{2\sigma_1^2}$$

Proof

$$K(f_1, f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)}$$

and

$$\begin{aligned} \log \frac{f_1(x)}{f_2(x)} &= \log \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}} = \\ &= \log \frac{\sigma_2}{\sigma_1} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2} \end{aligned}$$

So:

$$\begin{aligned} K(f_1, f_2) &= \int f_1(x) \log \frac{f_1(x)}{f_2(x)} = E_{f_1} \left[\log \frac{f_1(x)}{f_2(x)} \right] = \\ &= E_{f_1} \left[\log \frac{\sigma_2}{\sigma_1} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2} \right] = \\ &= \log \frac{\sigma_2}{\sigma_1} - \frac{E_{f_1}[(x-\mu_1)^2]}{2\sigma_1^2} + \frac{E_{f_1}[(x-\mu_2)^2]}{2\sigma_2^2} = \\ &= \log \frac{\sigma_2}{\sigma_1} - \frac{\sigma_1^2}{2\sigma_1^2} + \frac{\sigma_1^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2}{2\sigma_2^2} = \\ &= \log \frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}. \end{aligned}$$

$$\text{Thus: } K(f_1, f_2) = \log \frac{\sigma_1}{\sigma_2} - \frac{1}{2} + \frac{\sigma_2^2}{2\sigma_1^2} + \frac{(\mu_1 - \mu_2)^2}{2\sigma_1^2}$$

- **Jeffrey's divergence number:**

$$J(f_1, f_2) = \int [f_1 - f_2] \log \frac{f_1}{f_2} \quad (3.2.1.9)$$

Jeffrey's divergence (Jeffrey's (1948)) is closely related to the Kullback Leibler divergence number in the sense that $J(f_1, f_2)$ is the symmetric version of $K(f_1, f_2)$ since:

$$J(f_1, f_2) = K(f_1, f_2) + K(f_2, f_1)$$

Clearly, $K(f_1, f_2)$ and $J(f_1, f_2)$ share most of their properties. Similarly to $K(f_1, f_2)$, $J(f_1, f_2)$ requires that $f_1(x)$ and $f_2(x)$ be absolutely continuous with respect to each other. This is one of the problems that, just like $K(f_1, f_2)$, $J(f_1, f_2)$ has.

$J(f_1, f_2)$ is a Csiszar f divergence. It is obtained by (3.2.1.1) by setting $f(x) = (x-1)\log x$ and $g(x) = x$.

Lemma: The Jeffrey's divergence number when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:

$$J(f_1, f_2) = \frac{1}{2} \left(\frac{\sigma_1}{\sigma_2} - \frac{\sigma_2}{\sigma_1} \right)^2 + \frac{(\mu_1 - \mu_2)^2}{2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)$$

Proof

$$\begin{aligned} J(f_1, f_2) &= K(f_1, f_2) + K(f_2, f_1) = \\ &= -1 + \frac{\sigma_2^2}{2\sigma_1^2} + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{2\sigma_1^2} + \frac{(\mu_2 - \mu_1)^2}{2\sigma_2^2} \\ &= \frac{\sigma_1^4 + \sigma_2^4}{2\sigma_1^2\sigma_2^2} - 1 + \frac{(\mu_1 - \mu_2)^2}{2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \\ &= \frac{1}{2} \left(\frac{\sigma_1}{\sigma_2} - \frac{\sigma_2}{\sigma_1} \right)^2 + \frac{(\mu_1 - \mu_2)^2}{2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \end{aligned}$$

- **Lin's information number**

$$I(f_1, f_2) = \int f_1(x) \log \frac{f_1(x)}{\frac{1}{2}f_1(x) + \frac{1}{2}f_2(x)} dx \quad (3.2.1.10)$$

Although the $K(f_1, f_2)$ and $J(f_1, f_2)$ measures have many useful properties, they require that the probability distributions involved satisfy the condition of absolute continuity. Also, there are certain bounds that neither K nor J can provide, which is an issue that will be studied in another paragraph. $I(f_1, f_2)$ was introduced by Lin, (1991) to overcome those difficulties. It is a measure that preserves most of the desirable properties of K ; it is in fact closely related to K and can be described in terms of $K(f_1, f_2)$:

$$I(f_1, f_2) = K\left(f_1, \frac{1}{2}f_1 + \frac{1}{2}f_2\right).$$

$I(f_1, f_2)$, just like $K(f_1, f_2)$, satisfies the first one out of the three conditions of a metric, so it is neither a divergence nor a distance. It is clear that $I(f_1, f_2)$ is well defined and independent of the values of $f_1(x)$ and $f_2(x)$.

$I(f_1, f_2)$ coincides with the Csiszar f divergence if $f(x) = x \log \frac{2x}{1+x}$ and $g(x) = x$.

- **Symmetric version of Lin's information number**

$$L(f_1, f_2) = \int f_1(x) \log \frac{f_1(x)}{\frac{1}{2}f_1(x) + \frac{1}{2}f_2(x)} + f_2(x) \log \frac{f_2(x)}{\frac{1}{2}f_1(x) + \frac{1}{2}f_2(x)} dx \quad (3.2.1.11)$$

$I(f_1, f_2)$ is obviously not a symmetric measure. Lin, (1991) defined L , which is a symmetric divergence based on I as:

$$L(f_1, f_2) = I(f_1, f_2) + I(f_2, f_1).$$

Obviously, $L(f_1, f_2)$ is related to $I(f_1, f_2)$, in the same way that $J(f_1, f_2)$ is related to $K(f_1, f_2)$.

- **Jensen-Shannon Divergence Measure**



Let $\pi_1, \pi_2 \geq 0$, $\pi_1 + \pi_2 = 1$, be the weights of the two probability distributions f_1, f_2 respectively and H , a concave function. The Jensen-Shannon divergence is defined as

$$JS_{\pi}(f_1, f_2) = H(\pi_1 f_1 + \pi_2 f_2) - \pi_1 H(f_1) - \pi_2 H(f_2) \quad (3.2.1.12)$$

Since H is a concave function, according to Jensen's inequality, $JS_{\pi}(f_1, f_2)$ is nonnegative and equal to zero only when $f_1 = f_2$.

$JS_{\pi}(f_1, f_2)$ has been derived by Lin, (1991) as a generalization of $L(f_1, f_2)$ and one of its major features is that we can assign different weights to the distributions involved according to their importance. This is particularly useful in the study of decision problems where the weights could be the prior probabilities.

- **Bhattacharrya distance**

$$B(f_1, f_2) = -\log \rho(f_1, f_2) \quad (3.2.1.13)$$

where $\rho(f_1, f_2) = \int \sqrt{f_1 f_2} dx = 1 - \frac{1}{2} H_2^2(f_1, f_2)$, as Bhattacharrya coefficient. It is obtained by (3.2.1.1) by setting $f(x) = -\sqrt{x}$ and $g(x) = -\log(-x)$ and thus belongs to the family of Csiszar f divergences. It is also an alpha-divergences obtained by (3.2.1.3) for $\alpha = \frac{1}{2}$.

Lemma: The Bhattacharrya distance when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:

$$B(f_1, f_2) = \frac{1}{2} \log \frac{\sigma_1^2 + \sigma_2^2}{2\sqrt{\sigma_1^2 \sigma_2^2}} + \frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}$$

Proof

$$B(f_1, f_2) = -\log \int \sqrt{f_1 f_2} dx$$

and

$$\int \sqrt{f_1 f_2} dx = 1 - \frac{1}{2} H_2^2(f_1, f_2)$$

But:

$$H_2^2(f_1, f_2) = 2 - 2\sqrt{2} \sqrt{\frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp \left\{ -\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)} \right\}$$

So:

$$\begin{aligned} \int \sqrt{f_1 f_2} dx &= 1 - \frac{1}{2} H_2^2(f_1, f_2) = 1 - 1 + \sqrt{2} \sqrt{\frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp \left\{ -\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)} \right\} = \\ &= \sqrt{2} \sqrt{\frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp \left\{ -\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)} \right\} \end{aligned}$$

and:

$$B(f_1, f_2) = -\log \sqrt{2} \sqrt{\frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}} + \frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)} = \frac{1}{2} \log \frac{\sigma_1^2 + \sigma_2^2}{2\sqrt{\sigma_1^2 \sigma_2^2}} + \frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}$$

- **$1 - P_e$, where P_e : Error probability in classification of the optimal Bayes rule**

$$d(f_1, f_2) = 1 - P_e = 1 - \int \min[\pi f_1, (1 - \pi) f_2] dx \quad (3.2.1.14)$$

where $\pi, 1 - \pi$: a priori probabilities.

It is known that the error probability P_e of the optimal Bayes rule for the classification into two classes with a priori probabilities π and $1 - \pi$ and with corresponding densities of the observations f_1 and f_2 , is $P_e = \int \min[\pi f_1, (1 - \pi) f_2] dx$. It results that $1 - P_e$, which is a way to measure the distance between f_1 and f_2 , is obtained by (3.2.1.1) by setting $f(x) = -\min(x, 1 - x)$ and $g(x) = x + 1$.

- **Lissack and Fu distance**

$$d(f_1, f_2) = \int |f_1 - f_2|^\alpha dx, \alpha > 0 \quad (3.2.1.15)$$

It is a Csiszar f divergence only if $\alpha = 1$. In this case it is identical to L_1 and H_1 .

- **Chi-Squared divergence**

$$d(f_1, f_2) = \int \frac{[f_1(x) - f_2(x)]^2}{f_2(x)} dx = -1 + \int \frac{[f_1(x)]^2}{f_2(x)} dx \quad (3.2.1.16)$$

It is, according to Rigau et al. (2003) a Csiszar f divergence obtained by (3.2.1.1) by setting $f(x) = (x-1)^2$ and $g(x) = x$.

Lemma: The chi-squared divergence when f_1, f_2 are Gaussians, 1-dimensional distributions is given by:

$$d(f_1, f_2) = \frac{\sigma_2^2}{\sqrt{\sigma_1^2} \sqrt{2\sigma_2^2 - \sigma_1^2}} \exp \left\{ \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2 - \sigma_1^2} \right\} - 1$$

Proof

$$d(f_1, f_2) = \int \frac{[f_1(x) - f_2(x)]^2}{f_2(x)} dx = -1 + \int \frac{[f_1(x)]^2}{f_2(x)} dx$$

and

$$\int \frac{[f_1(x)]^2}{f_2(x)} dx = \int \frac{\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right)^2 \exp \left\{ -\frac{2(x-\mu_1)^2}{2\sigma_1^2} \right\}}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp \left\{ -\frac{2(x-\mu_2)^2}{2\sigma_2^2} \right\}} dx =$$

$$= \frac{\sqrt{2\pi\sigma_2^2}}{(\sqrt{2\pi\sigma_1^2})^2} \int \exp \left\{ -\frac{2(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2} \right\} dx =$$

$$= \frac{\sqrt{\sigma_2^2}}{\sqrt{2\pi\sigma_1^2}} \int \exp \left\{ -\frac{2(x-\mu_1)^2 \sigma_2^2 - (x-\mu_2)^2 \sigma_1^2}{2\sigma_1^2 \sigma_2^2} \right\} dx =$$

$$= \frac{\sqrt{\sigma_2^2}}{\sqrt{2\pi\sigma_1^2}} \int \exp \left\{ -\frac{(2\sigma_2^2 - \sigma_1^2)x^2 - 2(2\mu_1\sigma_2^2 - \mu_2\sigma_1^2)x + 2\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2}{2\sigma_1^2 \sigma_2^2} \right\} dx =$$

$$\begin{aligned}
&= \frac{\sqrt{\sigma_2^2}}{\sqrt{2\pi}\sigma_1^2} \int \exp \left\{ -\frac{2\sigma_2^2 - \sigma_1^2}{2\sigma_1^2\sigma_2^2} \left[x^2 - 2 \frac{2\mu_1\sigma_2^2 - \mu_2\sigma_1^2}{2\sigma_2^2 - \sigma_1^2} x + \left(\frac{2\mu_1\sigma_2^2 - \mu_2\sigma_1^2}{2\sigma_2^2 - \sigma_1^2} \right)^2 \right] \right\} dx \\
&\quad * \exp \left\{ \frac{\left(2\mu_1\sigma_2^2 - \mu_2\sigma_1^2 \right)^2}{2\sigma_1^2\sigma_2^2(2\sigma_2^2 - \sigma_1^2)} - \frac{2\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2}{2\sigma_1^2\sigma_2^2} \right\} = \\
&= \frac{\sqrt{\sigma_2^2}}{\sqrt{2\pi}\sigma_1^2} \int \exp \left\{ -\frac{\left(x - \frac{2\mu_1\sigma_2^2 - \mu_2\sigma_1^2}{2\sigma_2^2 - \sigma_1^2} \right)^2}{2 \frac{\sigma_1^2\sigma_2^2}{2\sigma_2^2 - \sigma_1^2}} \right\} dx \cdot \exp \left\{ \frac{2\sigma_1^2\sigma_2^2(\mu_1 - \mu_2)^2}{2\sigma_1^2\sigma_2^2(2\sigma_2^2 - \sigma_1^2)} \right\} = \\
&= \frac{\sqrt{\sigma_2^2}}{\sqrt{2\pi}\sigma_1^2} \sqrt{2\pi \frac{\sigma_1^2\sigma_2^2}{2\sigma_2^2 - \sigma_1^2}} \cdot \exp \left\{ \frac{2\sigma_1^2\sigma_2^2(\mu_1 - \mu_2)^2}{2\sigma_1^2\sigma_2^2(2\sigma_2^2 - \sigma_1^2)} \right\} = \\
&= \frac{\sigma_2^2}{\sqrt{\sigma_1^2} \sqrt{2\sigma_2^2 - \sigma_1^2}} \cdot \exp \left\{ \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2 - \sigma_1^2} \right\}.
\end{aligned}$$

All of the above measures of divergences which have been given in closed forms when the distributions are 1-dimensional Gaussians take into account the first and second moments.

3.2.2 Inequalities among distance measures

In this paragraph, some inequalities among distance measures are given. Most of them have been found in literature; some of their proofs are presented as well.

At first, several inequalities between the Error probability in classification P_e and many of the above mentioned distance measures are given. Those inequalities were derived in an attempt to find bounds of the

classification error probability. They have been collected by Basseville, (1988):

- $\frac{1}{2} \left[1 - \sqrt{1 - 4\pi(1-\pi)\rho^2(f_1, f_2)} \right] \leq P_e(f_1, f_2) \leq \sqrt{\pi(1-\pi)\rho(f_1, f_2)}, \quad (3.2.2.1)$

where $\rho(f_1, f_2) = \int \sqrt{f_1 f_2} d\lambda$

- $\frac{1}{2} \min(\pi, 1-\pi) e^{-J(f_1, f_2)} \leq P_e(f_1, f_2) \leq \sqrt{\pi(1-\pi)} \left[\frac{J(f_1, f_2)}{4} \right]^{-1/4} \quad (3.2.2.2)$

- $P_e(f_1, f_2) \leq \frac{1}{2} - \frac{1}{2} V(f_1, f_2) \quad (3.2.2.3)$

- $P_e(f_1, f_2) \leq \frac{1}{2} - \frac{1}{2} H_p^p(f_1, f_2) \quad (3.2.2.4)$

Some other known inequalities among the most widely used distance measures, are the following:

- $H_2^2(f_1, f_2)(2 - H_2^2(f_1, f_2)) = 1 - \rho^2(f_1, f_2) \quad (3.2.2.5)$

- $e^{-\frac{1}{2}K(f_1, f_2)} \leq \rho(f_1, f_2) \quad (3.2.2.6)$

- $H^2(f_1, f_2) \leq V(f_1, f_2) \leq H(f_1, f_2) \sqrt{2 - H^2(f_1, f_2)} \quad (3.2.2.7)$

- $\frac{1}{4} e^{K(f_1, f_2)} \leq 1 - V(f_1, f_2) \leq \rho(f_1, f_2) \quad (3.2.2.8)$

Between Kullback Leibler number and L_1 distance Bretagnolle and Huber, (1979) proved the following:

Lemma: $L_1(f_1, f_2) \leq 2\sqrt{1 - e^{-K(f_1, f_2)}} \leq 2 - e^{-K(f_1, f_2)}$ (3.2.2.9)

Proof

$$\begin{aligned}
 -K(f_1, f_2) &= -\int f_1 \log \frac{f_1}{f_2} = \int f_1 \log \frac{f_2}{f_1} = \int f_1 \log \frac{f_2}{f_1} + \int f_1 \log 1 = \\
 &= \int f_1 \log \left(\min \left\{ \frac{f_2}{f_1}, 1 \right\} \right) + \int f_1 \log \left(\max \left\{ \frac{f_2}{f_1}, 1 \right\} \right) \leq \\
 &\leq \log \int f_1 \min \left\{ \frac{f_2}{f_1}, 1 \right\} + \log \int f_1 \max \left\{ \frac{f_2}{f_1}, 1 \right\} = \\
 &= \log \left[1 - \frac{1}{2} \int |f_1 - f_2| \right] + \log \left[1 + \frac{1}{2} \int |f_1 - f_2| \right] = \\
 &= \log \left[\left(1 - \frac{1}{2} \int |f_1 - f_2| \right) \left(1 + \frac{1}{2} \int |f_1 - f_2| \right) \right] = \\
 &= \log \left[1 - \frac{1}{4} \left(\int |f_1 - f_2| \right)^2 \right] = \\
 &= \log \left[1 - \frac{1}{4} L_1^2(f_1, f_2) \right]
 \end{aligned}$$

so:

$$\begin{aligned}
 -K(f_1, f_2) &\leq \log \left[1 - \frac{1}{4} L_1^2(f_1, f_2) \right] \Leftrightarrow e^{-K} \leq 1 - \frac{1}{4} L_1^2(f_1, f_2) \Leftrightarrow \\
 L_1^2(f_1, f_2) &\leq 4 - 4e^{-K(f_1, f_2)}
 \end{aligned}$$

or: $L_1(f_1, f_2) \leq 2\sqrt{1 - e^{-K(f_1, f_2)}} \leq 2 - e^{-K(f_1, f_2)}$

This inequality can be restated as follows:

$$\int \min(f_1, f_2) \geq \frac{1}{2} e^{-K(f_1, f_2)}$$

Another inequality between Kullback Leibler number and L_1 distance was the following derived by Kullback, (1967), Csizar, (1967) and Kemperman (1969):

Lemma: $L_1(f_1, f_2) \leq \sqrt{2K(f_1, f_2)}$ (3.2.2.10)

Proof

If $A = \{f_1 \geq f_2\}$, $f_3 = gI_A / q$, $q = \int_A f_2$, $p = \int_A f_1$

$$\begin{aligned} \int_A f_1 \log \frac{f_1}{f_2} &= \int f_3 \frac{f_1}{f_2} \log \frac{f_1}{f_2} \int_A f_2 \geq \int f_3 \frac{f_1}{f_2} \log \left(\int f_3 \frac{f_1}{f_2} \right) \int_A f_2 = \\ &= \int_A f_1 \log \left(\int_A f_1 / \int_A f_2 \right) = p \log \frac{p}{q} \end{aligned}$$

$$\int_{A^c} f_1 \log \frac{f_1}{f_2} \geq (1-p) \log \frac{1-p}{1-q}$$

$$\text{so: } K(f_1, f_2) \geq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} = H(f_1, f_2)$$

if: $p - q = r$:

$$K(f_1, f_2) \geq H(r) = (q+r) \log \frac{q+r}{q} + (1-q-r) \log \frac{1-q-r}{1-q},$$

or:

$$K(f_1, f_2) \geq H(r) = (q+r) \log \left(1 + \frac{r}{q} \right) + (1-q-r) \log \left(1 - \frac{r}{1-q} \right)$$

$$H'(r) = \log \left(1 + \frac{r}{q} \right) - \log \left(1 - \frac{r}{1-q} \right)$$

$$H''(r) = \frac{1}{q+r} + \frac{1}{1-q-r} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)} \geq 4$$

$$\begin{aligned} K(f_1, f_2) &\geq 4 \frac{r^2}{2} = 2(p-q)^2 = 2 \left(\int_A f_1 - \int_A f_2 \right)^2 = 2 \left[\frac{\int |f_1 - f_2|}{2} \right]^2 \\ &= \frac{1}{2} \left[\int |f_1 - f_2| \right]^2 = \frac{1}{2} L_1^2(f_1, f_2) \end{aligned}$$

or:

$$L_1^2(f_1, f_2) \leq 2K(f_1, f_2) \Leftrightarrow L_1(f_1, f_2) \leq \sqrt{2K(f_1, f_2)}$$

Devroye (1987) proved that Hellinger distances $H_1(f_1, f_2)$ and $H_2(f_1, f_2)$ are related with the inequality:

Lemma:

$$H_2^2(f_1, f_2) \leq H_1(f_1, f_2) \leq H_2(f_1, f_2) \sqrt{4 - H_2^2(f_1, f_2)} \leq 2H_2(f_1, f_2) \quad (3.2.2.11)$$

Proof

$$H_1 = \int |f_1 - f_2|$$

$$H_2 = \left[\int (\sqrt{f_1} - \sqrt{f_2})^2 \right]^{1/2}$$

$$H_1 = \int |f_1 - f_2| = \int (\sqrt{f_1} - \sqrt{f_2})(\sqrt{f_1} + \sqrt{f_2}) \geq \int (\sqrt{f_1} - \sqrt{f_2})^2 = H_2^2$$

$$\text{so: } H_2^2(f_1, f_2) \leq H_1(f_1, f_2)$$

$$H_1^2(f_1, f_2) = \left[\int |f_1 - f_2| \right]^2 = \left[\int (\sqrt{f_1} - \sqrt{f_2})(\sqrt{f_1} + \sqrt{f_2}) \right]^2 \leq$$

$$\left(\begin{array}{l} \text{Cauchy-} \\ \text{Schwarz} \\ \text{inequality} \end{array} \right) \leq \int (\sqrt{f_1} - \sqrt{f_2})^2 \int (\sqrt{f_1} + \sqrt{f_2})^2 =$$

$$= H_2^2 (\sqrt{f_1} + \sqrt{f_2})^2 = H_2^2 (2 + 2 \int \sqrt{f_1 f_2}) = H_2^2 (4 - H_2^2) \leq 4H_2^2$$

$$\left(\text{because: } H_2^2(f_1, f_2) = \int (\sqrt{f_1} - \sqrt{f_2})^2 = 2 - 2 \int \sqrt{f_1 f_2} \right)$$

$$\text{So: } H_1(f_1, f_2) \leq H_2(f_1, f_2) \sqrt{4 - H_2^2(f_1, f_2)} \leq 2H_2(f_1, f_2)$$

Devroye (1987) also proved that Hellinger distances $H_1(f_1, f_2)$ and $H_2(f_1, f_2)$ are related by the inequality:

$$\textbf{Lemma: } 2 - H_1(f_1, f_2) \geq \left[1 - \frac{1}{2} H_2^2(f_1, f_2) \right]^2 \quad (3.2.2.12)$$

Proof

We will use Le Cam's inequality:

$$\int \min\{f_1, f_2\} \geq \frac{1}{2} \int (\sqrt{f_1 f_2})^2$$

Proof of Le Cam's inequality:

$$\left(\int_{f_1 < f_2} \sqrt{f_1 f_2} \right)^2 = \left(\int_{f_1 < f_2} f_1 \sqrt{\frac{f_2}{f_1}} \right)^2 \leq \int_{f_1 < f_2} f_1 \frac{f_2}{f_1} \int_{f_1 < f_2} f_1 = \int_{f_1 < f_2} f_2 \int_{f_1 < f_2} f_1 \leq \int_{f_1 < f_2} f_1$$

$$\text{Similarly: } \left(\int_{f_2 < f_1} \sqrt{f_1 f_2} \right)^2 \leq \int_{f_2 < f_1} f_2$$

$$\begin{aligned} \left(\int \sqrt{f_1 f_2} \right)^2 &= \left(\int_{f_1 < f_2} \sqrt{f_1 f_2} + \int_{f_2 < f_1} \sqrt{f_1 f_2} \right)^2 \leq 2 \left(\int_{f_1 < f_2} \sqrt{f_1 f_2} \right)^2 + 2 \left(\int_{f_2 < f_1} \sqrt{f_1 f_2} \right)^2 \leq \\ &\leq 2 \int_{f_1 < f_2} f_1 + 2 \int_{f_2 < f_1} f_2 = 2 \int \min\{f_1, f_2\} \end{aligned}$$

$$\text{So: } 2 \int \min\{f_1, f_2\} \geq \int (\sqrt{f_1 f_2})^2$$

$$\text{But: } \frac{H_1}{2} = \frac{\int |f_1 - f_2|}{2} = 1 - \int \min\{f_1, f_2\} \Leftrightarrow \int \min\{f_1, f_2\} = 1 - \frac{H_1}{2}$$

and:

$$H_2^2(f_1, f_2) = \int (\sqrt{f_1} - \sqrt{f_2})^2 = 2 - 2 \int \sqrt{f_1 f_2} \Leftrightarrow$$

$$\left(\int \sqrt{f_1 f_2} \right)^2 = \left[1 - \frac{1}{2} H_2^2(f_1, f_2) \right]^2$$

$$\text{So: } 2 - H_1(f_1, f_2) \geq \left[1 - \frac{1}{2} H_2^2(f_1, f_2) \right]^2$$

One, rather obvious, inequality between Kullback and Jeffreys number is:

$$\bullet \quad K(f_1, f_2) \leq J(f_1, f_2) \quad (3.2.2.13)$$

Lemma: For Hellinger H_2 and L_2 distances if f_1, f_2 Gaussians:

$$H_2^2(f_1, f_2) \leq 2 \left(1 - \frac{\sigma_1^2}{\sigma_2^2} \right)^2 + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} \quad (3.2.2.14)$$

Moreover if: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $\sigma^2 \geq \frac{1}{4\pi}$:

$$L_2(f_1, f_2) \leq H_2(f_1, f_2) \quad (3.2.2.15)$$

Proof

$$L_2(f_1, f_2) = \sqrt{\frac{1}{\sqrt{4\pi\sigma^2}}} \sqrt{2 - 2 \exp \left\{ -\frac{(\mu_1 - \mu_2)^2}{4\sigma^2} \right\}}$$

$$H_2(f_1, f_2) = \sqrt{2 - 2 \exp \left\{ -\frac{(\mu_1 - \mu_2)^2}{8\sigma^2} \right\}}$$

$$x \geq 0 \Leftrightarrow x \geq \frac{x}{2} \Leftrightarrow -x \leq -\frac{x}{2} \Leftrightarrow \exp\{-x\} \leq \exp\left\{-\frac{x}{2}\right\}$$

$$\Leftrightarrow 2 - 2 \exp\{-x\} \leq 2 - 2 \exp\left\{-\frac{x}{2}\right\}$$

if: $\sigma^2 \geq \frac{1}{4\pi}$:

$$\frac{1}{\sqrt{4\pi\sigma^2}} [2 - 2 \exp\{-x\}] \leq 2 - 2 \exp\{-x\} \leq 2 - 2 \exp\left\{-\frac{x}{2}\right\}$$

So: $L_2^2(f_1, f_2) \leq H_2^2(f_1, f_2)$

Or: $L_2(f_1, f_2) \leq H_2(f_1, f_2)$

Several relationships have been found in literature between Kullback information number or Jeffrey's number and $I(f_1, f_2)$ and $L(f_1, f_2)$ divergences, introduced by Lin, (1991).

First of all, from the definition of $I(f_1, f_2)$, (3.2.1.14), it is obvious that I directed divergence is bounded by the K divergence.

$$\bullet \quad I(f_1, f_2) \leq \frac{1}{2} K(f_1, f_2) \quad (3.2.2.16)$$

The L divergence is related to the J divergence in the same way as I is related to K . From inequality (3.2.2.16) and the definition of $L(f_1, f_2)$, (3.2.1.15), we can easily derive the following:

$$\textbf{Lemma:} \quad L(f_1, f_2) \leq \frac{1}{2} K(f_1, f_2) \quad (3.2.2.17)$$

Proof

$$\frac{f_1 + f_2}{2} \leq \sqrt{f_1 f_2}$$

$$I(f_1, f_2) = \int f_1(x) \log \frac{f_1(x)}{\frac{1}{2} f_1(x) + \frac{1}{2} f_2(x)} \leq \int f_1(x) \log \frac{f_1(x)}{\sqrt{f_1(x) f_2(x)}} =$$

$$\int f_1(x) \log \sqrt{\frac{f_1(x)}{f_2(x)}} = \frac{1}{2} \int f_1(x) \log \frac{f_1(x)}{f_2(x)} = \frac{1}{2} K(f_1, f_2)$$

Moreover, a lot of effort has been devoted to finding the relationship (in terms of bounds) between the $K(f_1, f_2)$ directed divergence and the variational distance. The variational distance between two probability distributions is defined as

$$V(f_1, f_2) = \sum_x |f_2(x) - f_1(x)|$$

Bounds of distance measures concerned to the variational distance are useful in decision-making applications. Several lower bounds for $K(f_1, f_2)$ in terms of $V(f_1, f_2)$ have been found, among which the sharpest is given by:



- $K(f_1, f_2) \geq \max\{L_1(V(f_1, f_2)), L_2(V(f_1, f_2))\}, \quad (3.2.2.18)$

where:

$$L_1(V(f_1, f_2)) = \log \frac{2+V(f_1, f_2)}{2-V(f_1, f_2)} - \frac{2V(f_1, f_2)}{2+V(f_1, f_2)}, \quad 0 \leq V(f_1, f_2) \leq 2 \quad (3.2.2.19)$$

established by Vajda (1970) and

$$L_2(V(f_1, f_2)) = \frac{V^2(f_1, f_2)}{2} + \frac{V^4(f_1, f_2)}{36} + \frac{V^6(f_1, f_2)}{288}, \quad 0 \leq V(f_1, f_2) \leq 2 \quad (3.2.2.20)$$

derived by Toussaint (1975).

However, according to Lin, (1991), no general upper bound exists for either $K(f_1, f_2)$ or $J(f_1, f_2)$ in terms of the variational distance. This is another difficulty in using the $K(f_1, f_2)$ directed divergence as a measure of discrepancy between probability distributions. In contrast to those situations for the K and J divergences, both lower and upper bounds exist for $I(f_1, f_2)$ and $L(f_1, f_2)$ divergences, introduced by Lin, (1991).

Lemma: For the $I(f_1, f_2)$ directed divergence the following lower bound holds:

$$I(f_1, f_2) \geq \max\left\{L_1\left(\frac{V(f_1, f_2)}{2}\right), L_2\left(\frac{V(f_1, f_2)}{2}\right)\right\}, \quad (3.2.2.21)$$

where L_1, L_2 are defined by (3.2.3.19) and (3.2.3.20), respectively.

Proof

$$K(f_1, f_2) \geq \max\{L_1(V(f_1, f_2)), L_2(V(f_1, f_2))\}$$

$$I(f_1, f_2) \leq K\left(f_1, \frac{1}{2}f_1 + \frac{1}{2}f_2\right)$$

$$\text{So: } I(f_1, f_2) \geq \max \left\{ L_1 \left(V \left(f_1, \frac{1}{2} f_1 + \frac{1}{2} f_2 \right) \right), L_2 \left(V \left(f_1, \frac{1}{2} f_1 + \frac{1}{2} f_2 \right) \right) \right\}$$

But:

$$\begin{aligned} V \left(f_1, \frac{1}{2} f_1 + \frac{1}{2} f_2 \right) &= \int \left| f_1(x) - \left(\frac{1}{2} f_1(x) + \frac{1}{2} f_2(x) \right) \right| = \int \left| \frac{1}{2} f_1(x) - \frac{1}{2} f_2(x) \right| \\ &= \frac{1}{2} \int |f_1(x) - f_2(x)| = \frac{1}{2} V(f_1, f_2) \end{aligned}$$

Lin, (1991) proved that the variational distance and the L divergence measure satisfy the inequality:

- $L(f_1, f_2) \leq V(f_1, f_2)$ (3.2.2.22)

and since $I(f_1, f_2)$ is clearly not greater than $L(f_1, f_2)$, i.e.:

- $I(f_1, f_2) \leq L(f_1, f_2)$ (3.2.2.23)

we can derive both lower and upper bounds for $L(f_1, f_2)$:

- $I(f_1, f_2) \leq L(f_1, f_2) \leq V(f_1, f_2)$ (3.2.2.24)

Moreover, from (3.2.2.24) it is clear that the variational distance serves as an upper bound to I divergence as well. Thus, considering (3.2.2.21) and (3.2.2.24) we can provide the lower and upper bounds for $I(f_1, f_2)$:

- $\max \left\{ L_1 \left(\frac{V(f_1, f_2)}{2} \right), L_2 \left(\frac{V(f_1, f_2)}{2} \right) \right\} \leq I(f_1, f_2) \leq V(f_1, f_2)$ (3.2.2.25)

Some other inequalities proved by Lin, (1991) about I and L divergences are those concerning to their boundedness, namely,

- $I(f_1, f_2) \leq 1$ (3.2.2.26)

and

- $L(f_1, f_2) \leq 2$ (3.2.2.27)

3.3 The example

In our example we have already ascribed first frame's pixels to three Normal distributions. We want to classify each one of next frames' pixels to one of the three Normal distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ or $N(\mu_3, \sigma_3^2)$.

We assume that every new incoming pixel comes from a Normal distribution $N(\mu_p, \sigma_p^2)$ and we measure how close is the distribution of the incoming pixel $N(\mu_p, \sigma_p^2)$ to the three existing distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$. The pixel is ascribed to that distribution from which it desist less.

In order to measure those distances and ascribe the pixels of second frame to one of the Normal distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$ we have applied several of the measures mentioned above. More precisely we used the Hellinger H_2 distance, the L_2 distance, the Kullback- Leibler information number, the Jeffrey's divergence number, the Bhattacharrya distance and the Chi- Squared distance.

For illustration reasons we present the assignment of second frame's pixels according to Jeffrey's divergence (figure 3.3.1) and the assignment according to L_2 distance (figure 3.3.2).



Figure 3.3.1. The assignment of second frame's pixels according to Jeffrey's divergence.



Figure 3.3.2. The assignment of second frame's pixels according to L_2 distance.

In those figures each pixel represents that group (background, skin or grey zone) with which the pixel was matched. It is clear that Jeffrey's number provided more satisfying results than L_2 distance (actually it provided the most satisfying results among all the distance measures we tried as we will see on the experimental results chapter). The L_2 distance has completely misclassified the background pixels in the lower part of the figure.



CHAPTER 4

UPDATING THE PARAMETERS

4.1 Introduction

Once we determined a way to find which of the distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$, say $N(\mu_i, \sigma_i^2)$ is closest to the distribution of incoming pixel $N(\mu_p, \sigma_p^2)$ and have ascribed the pixel to this Normal, the pixel will contribute to the process of awareness of the parameters of distribution $N(\mu_i, \sigma_i^2)$. The parameters of distribution $N(\mu_p, \sigma_p^2)$ from which the pixel is supposed to come from will give extra information about the parameters of $N(\mu_i, \sigma_i^2)$. In practice we will approximate $N(\mu_i, \sigma_i^2)$ and $N(\mu_p, \sigma_p^2)$ with a single Normal distribution. The next issue to be clarified is what will be the parameters (weight, mean, and variance) of this new updated distribution.

The problem of estimating those parameters is one of the oldest estimation problems in the statistics literature. It was first considered by Pearson, (1894) but it continues to be of interest as it has been witnessed by many recent papers; it attracts a great deal of attention which probably reflects its difficulty and the lack of completely satisfactory solution.

One general method for estimation is the method of moments, which provide reasonable parameter estimates provided the sample size is large enough. However, the method of moments lacks some optimal properties of other estimation methods, for example, the maximum likelihood estimation.



4.2 The method of moments

Assume we have two Normal distributions, $f_1 \sim N(\mu_1, \sigma_1^2)$ with weight w_1 and $f_2 \sim N(\mu_2, \sigma_2^2)$ with weight w_2 . We merge them, and want to approximate their mixture with a single Normal distribution $f \sim N(\mu, \sigma^2)$ with weight w .

Thus, $Y \sim w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2)$ will be approximated by $Z \sim w N(\mu, \sigma^2)$.

First of all, it is clear that the new density will have weight equal to the sum of weights of the pooled densities. So:

$$w = w_1 + w_2 \quad (4.2.1)$$

By equating the theoretical first moments of Y and Z we have:

$$w_1 \mu_1 + w_2 \mu_2 = w \mu$$

or,

$$\mu = \frac{w_1}{w} \mu_1 + \frac{w_2}{w} \mu_2 \quad (4.2.2)$$

and by setting $\rho = \frac{w_2}{w}$ in (4.2.2) we have:

$$\mu = (1 - \rho) \mu_1 + \rho \mu_2 \quad (4.2.3)$$

Similarly, by equating the theoretical second moments of Y and Z we have:

$$w_1 (\sigma_1^2 + \mu_1^2) + w_2 (\sigma_2^2 + \mu_2^2) = w (\sigma^2 + \mu^2)$$

or, using (4.2.2)

$$\frac{w_1}{w} \sigma_1^2 + \frac{w_1}{w} \mu_1^2 + \frac{w_2}{w} \sigma_2^2 + \frac{w_2}{w} \mu_2^2 = \sigma^2 + \left(\frac{w_1}{w} \mu_1 + \frac{w_2}{w} \mu_2 \right)^2$$

or,

$$\sigma^2 = \frac{w_1}{w} \sigma_1^2 + \frac{w_1}{w} \mu_1^2 + \frac{w_2}{w} \sigma_2^2 + \frac{w_2}{w} \mu_2^2 - \frac{w_1^2}{w^2} \mu_1^2 - \frac{w_2^2}{w^2} \mu_2^2 - 2 \frac{w_1 w_2}{w^2} \mu_1 \mu_2$$

and because of (4.2.1) we have,

$$\sigma^2 = \frac{w_1}{w} \sigma_1^2 + \frac{w_2}{w} \sigma_2^2 + \frac{w_1(w_1 + w_2) - w_1^2}{w^2} \mu_1^2 + \frac{w_2(w_1 + w_2) - w_2^2}{w^2} \mu_2^2 - 2 \frac{w_1}{w} \frac{w_2}{w} \mu_1 \mu_2$$

or,

$$\sigma^2 = \frac{w_1}{w} \sigma_1^2 + \frac{w_2}{w} \sigma_2^2 + \frac{w_1 w_2}{w^2} \mu_1^2 + \frac{w_1 w_2}{w^2} \mu_2^2 - 2 \frac{w_1}{w} \frac{w_2}{w} \mu_1 \mu_2$$

or,

$$\sigma^2 = \frac{w_1}{w} \sigma_1^2 + \frac{w_2}{w} \sigma_2^2 + \frac{w_1}{w} \frac{w_2}{w} (\mu_1^2 + \mu_2^2 - 2\mu_1 \mu_2)$$

Finally, as we have set $\rho = \frac{w_2}{w}$, we get:

$$\sigma^2 = (1-\rho) \sigma_1^2 + \rho \sigma_2^2 + (1-\rho) \rho (\mu_1 - \mu_2)^2 \quad (4.2.4)$$

(4.2.1), (4.2.2) and (4.2.4) give the parameters of the single Normal distribution which approximates the mixture of the two Normal distributions.

4.3 Method of moments in our example

In our example, we first introduce some learning parameter a , which stands for the weight of $N(\mu_p, \sigma_p^2)$; i.e. the distribution of the incoming pixel. Obviously, we need to have $0 < a < 1$ and as a weighs on the weights of the three existing distributions, we subtract $100a\%$ from each one of the three existing weights w_1, w_2, w_3 , $\left(\sum_{i=1}^3 w_i = 1 \right)$ and assign it to the incoming

distribution's weight. The general principal about the parameter α is that it is related to the rate of recording. The faster the rate of recording is, the smaller the value of α we take. In our case, based upon the rate of recording of the camera used this was set $\alpha = 0.05$.

In other words, the three existing distributions have weights $w_i(1-\alpha)$, $i=1,2,3$ and the incoming distribution $N(\mu_p, \sigma_p^2)$ has weight $\alpha = \sum_{i=1}^3 \alpha w_i = \alpha \sum_{i=1}^3 w_i$.

By this definition of α , the sum of all the existing weights is equal to one:

$$\alpha + \sum_{i=1}^3 w_i(1-\alpha) = \alpha + \sum_{i=1}^3 w_i - \sum_{i=1}^3 \alpha w_i = \sum_{i=1}^3 w_i = 1$$

Let us assume now that we have a match between the new distribution $N(\mu_p, \sigma_p^2)$ and one of the existing distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$, say the distribution $N(\mu_j, \sigma_j^2)$ $1 \leq j \leq 3$ (which we will call from now on the winner distribution).

We update the weights of the mixture model as follows:

$$\begin{aligned} w'_i &= (1-\alpha)w_i, & i=1,2,3 \text{ and } i \neq j \\ w'_j &= (1-\alpha)w_j + \alpha \end{aligned}$$

i.e. we assign the weight of the incoming distribution to the winner.

Then update the means and variances μ_i and σ_i^2 $i=1,2,3$. If we call w_1 the $(1-\alpha)w_j$, i.e., w_1 is the weight of the j -th component (which is the winner in the match) before pooling it with the incoming distribution p , and if we call $w_2 = \alpha$, i.e., the weight of the new observation, then we define

$$\rho = \frac{w_2}{w_1 + w_2} = \frac{\alpha}{(1-\alpha)w_j + \alpha}$$

and using the method of moments, i.e., the formulas (4.2.3) and (4.2.4), we get:

$$\mu_j' = (1 - \rho)\mu_j + \rho\mu_p$$

$$\sigma_j'^2 = (1 - \rho)\sigma_j^2 + \rho\sigma_p^2 + (1 - \rho)\rho(\mu_j - \mu_p)^2$$

while the other two (unmatched) distributions keep the same mean and variance that they had before the matching.

The updating of parameters described above is being applied for all pixels and all frames. In what follows, a mathematical algorithm is given.

After Steps 1 and 2 of the algorithm given in the introduction (from which we obtained estimators $\hat{\mu}_i, \hat{\sigma}_i^2, \hat{w}_i, i=1, 2, 3$), set :

$$\hat{\mu}_{i,j,l} := \hat{\mu}_i, \hat{\sigma}_{i,j,l}^2 := \hat{\sigma}_i^2, \hat{w}_{i,j,l} := \hat{w}_i, \text{ for } i=1, 2, 3, \text{ and } j=1, \dots, rc$$

Then perform the following steps for all pixels ($j=1, \dots, rc$) and all frames ($k=1, \dots, N$):

- Let $i_0 := \arg \min_{i=1,2,3} D[N(\hat{\mu}_{i,j,k-1}, \hat{\sigma}_{i,j,k-1}^2), N(x_{jk}, \sigma_p^2)]$, where x_{jk} denotes the value of pixel j in the k -th frame and $D[.,.]$ denotes the chosen discrepancy.
- For $i \neq i_0$, set $\hat{w}_{i,j,k} := (1 - \alpha)\hat{w}_{i,j,k-1}$, $\hat{\mu}_{i,j,k} := \hat{\mu}_{i,j,k-1}$, $\hat{\sigma}_{i,j,k}^2 := \hat{\sigma}_{i,j,k-1}^2$.
- Calculate $\hat{w}_{i_0,j,k}$, $\hat{\mu}_{i_0,j,k}$, $\hat{\sigma}_{i_0,j,k}^2$ from the formulas in section 4.2, setting:

$$w_1 = (1 - \alpha)\hat{w}_{i_0,j,k-1}, \mu_1 = \hat{\mu}_{i_0,j,k-1}, \sigma_1^2 = \hat{\sigma}_{i_0,j,k-1}^2$$

$$w_2 = \alpha, \mu_2 = x_{jk}, \sigma_2^2 = \sigma_p^2$$

4.4 Simulation study.

In this section we present a simulation study in order to try some other formulas for ρ instead of $\rho = \frac{w_2}{w_1 + w_2}$ where we will allow both the means and variances to influence the parameter ρ which is used in:

$$\mu = (1 - \rho)\mu_1 + \rho\mu_2$$
$$\sigma^2 = (1 - \rho)\sigma_1^2 + \rho\sigma_2^2 + (1 - \rho)\rho(\mu_1 - \mu_2)^2$$

In order to check the quality of the approximation for each one of the formulas of ρ we used the qq plots (of the exact versus the approximating distributions).

We first try the following formula for ρ :

$$\rho = \frac{w_2\sigma_2}{w_1\sigma_1 + w_2\sigma_2} \quad (4.4.1)$$

where the variances of the two distributions have been introduced.

In order to check the appropriateness of the above formula we simulated data from mixtures of Normal distributions:

$$w_1N(\mu_1, \sigma_1^2) + w_2N(\mu_2, \sigma_2^2)$$

We define the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, w_1, w_2 = 1 - w_1$ of each one of the mixtures to take values:

$$\mu_1 = 0, 5, 10, 15, 20, 25, 30$$

$$\mu_2 = 0$$

$$\sigma_1 = 1, 5, 9$$

$$\sigma_2 = 1, 5, 9$$

$$w_1 = 0.1, 0.5, 0.9$$

$$w_2 = 1 - w_1 = 0.9, 0.5, 0.1$$

The full factorial of all the possible combinations for the parameter values gave us 189 cases.

We then calculated the parameters μ and σ^2 of the distribution that approximates each mixture based on the formulas:

$$\mu = (1 - \rho)\mu_1 + \rho\mu_2$$

$$\sigma^2 = (1 - \rho)\sigma_1^2 + \rho\sigma_2^2 + (1 - \rho)\rho(\mu_1 - \mu_2)^2$$

where ρ is given in (4.4.1).

To check the quality of the approximation of the distribution $N(\mu, \sigma^2)$ to the mixture of two components, in each of the above described cases, we used the qq (quantile-quantile) plots (figure 4.4.1), where the quantiles of the approximating versus exact distribution are plotted.

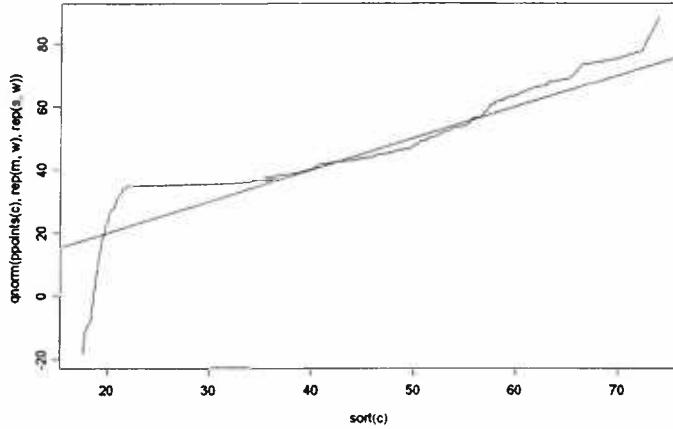


Figure 4.4.1. A qq plot.

We calculated the area between the dichotomous of the axis and the line of the plot and compared it with the corresponding area provided when using the method of moments with $\rho = \frac{w_2}{w_1 + w_2}$. It was found out that in some

mixtures $\rho = \frac{w_2\sigma_2}{w_1\sigma_1 + w_2\sigma_2}$ provided smaller area than $\rho = \frac{w_2}{w_1 + w_2}$, while the

opposite happened in other mixtures. It is clear that in a qq plot, the less area

found between the dichotomous of the axis and the line of the plot, the better the approximation is. We then observed that the approximation when using

$$\rho = \frac{w_2 \sigma_2}{w_1 \sigma_1 + w_2 \sigma_2} \quad \text{was not good enough for mixtures with big deviations}$$

between σ_1 and σ_2 . Thus, we decided to use the following formula for ρ , where σ 's and w 's weigh less.

$$\rho = \frac{\sqrt{w_2 \sigma_2}}{\sqrt{w_1 \sigma_1} + \sqrt{w_2 \sigma_2}} \quad (4.4.2)$$

We checked the performance of distribution $N(\mu, \sigma^2)$ by same way as before and concluded that w_1 and w_2 should weigh more than σ_1 and σ_2 in ρ 's formula. Thus, we tried the above formula:

$$\rho = \frac{w_2 \sqrt{\sigma_2}}{w_1 \sqrt{\sigma_1} + w_2 \sqrt{\sigma_2}} \quad (4.4.3)$$

This formula provided smaller areas between the dichotomous of the axis and the line of qq plots comparing with the areas calculated for formulas (4.4.1) and (4.4.2). However, in those mixtures where the means μ_1 and μ_2 differ a lot, none of the three formulas for ρ seemed to be good enough. We then decided to let the means μ_1 and μ_2 to be used in ρ as well:

$$\rho = \frac{\mu_2 w_2 \sqrt{\sigma_2}}{\mu_1 w_1 \sqrt{\sigma_1} + \mu_2 w_2 \sqrt{\sigma_2}} \quad (4.4.4)$$

This formula operated much better in approximating the mixtures with a single distribution.

Pednekar et al. (2002) proposed Gaussians functions to be used for defining the percentage of participation of each distribution of a mixture in the parameters of a single distribution which approximates that mixture. Thus, we next used the following formula for ρ :

$$\rho = \frac{\exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right)}{\exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2}\right) + \exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right)} \quad (4.4.5)$$

This formula provided even better results than the previous formulas, but we made one more effort to improve ρ by inducing in (4.4.5) the weights w_1 and w_2 :

$$\rho = \frac{\exp\left(-\frac{w_2 (\mu_1 - \mu_2)^2}{\sigma_2^2}\right)}{\exp\left(-\frac{w_1 (\mu_1 - \mu_2)^2}{\sigma_1^2}\right) + \exp\left(-\frac{w_2 (\mu_1 - \mu_2)^2}{\sigma_2^2}\right)} \quad (4.4.6)$$

At the end, we made an overall comparison of the results of this simulation study. We compared the average areas found for each simulated mixture with the use of each one of ρ s given by (4.4.1), (4.4.2), (4.4.3), (4.4.4), (4.4.5), (4.4.6) and of course by $\rho = \frac{w_2}{w_1 + w_2}$. We concluded that formulas (4.4.4), (4.4.5) together with $\rho = \frac{w_2}{w_1 + w_2}$ provided the minimum areas among all the different ρ s.



CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Introduction

In the previous section we described a statistical procedure for pooling together the distribution of an incoming pixel to the ‘closest’ one of three existing Normal distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ and $N(\mu_3, \sigma_3^2)$, which account for human body, background and grey zone pixels. When the procedure is applied to all $r \times c$ pixels of second frame, the three distributions accounting for each pixel are different from the distributions accounting for another pixel.

Then, the same procedure is applied to all following up frames. Every time, the pixels are assigned to that Normal from which they desist less. We used several different measures to define the distance between distributions; Jeffrey’s divergence, Kullback Leibler divergence, Hellinger H_2 , L_2 , Bhattacharyya and Chi-Squared divergence. Some of them appeared to be quite unsuitable, as they did not manage to result in proper ascription of pixels in distributions.

5.2 The results

We will judge the performance of several measures of distance on an inframed video segment. More precisely, we select an area 94×57 pixels corresponding to face during the entire video segment and we count the number of face pixels at the last frame of the video, to judge how well each divergence does.



The area of pixels selected is shown in figures 5.2.1 and 5.2.2, for the first and the last frame correspondingly.

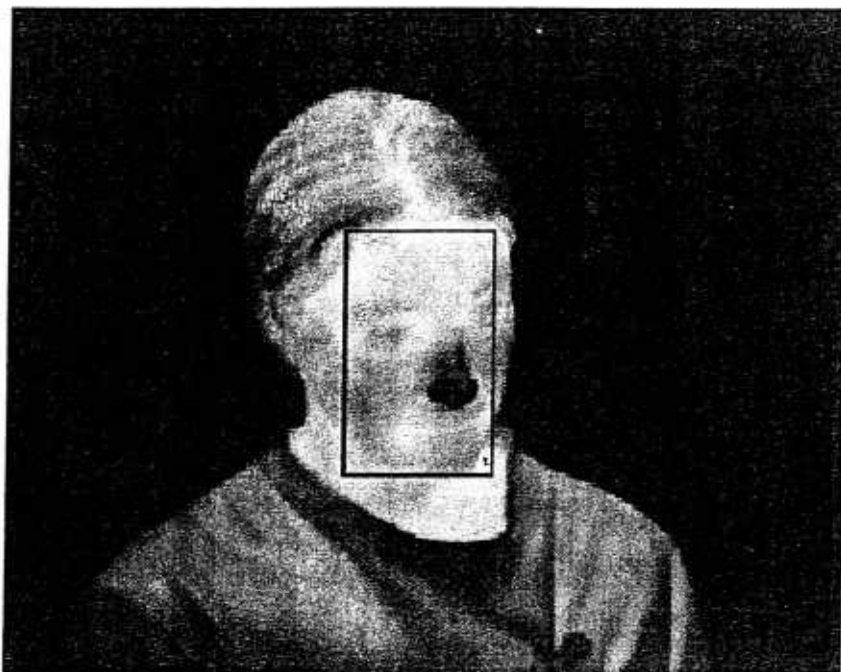


Figure 5.2.1. The area of pixels in the first frame in which our results were checked.

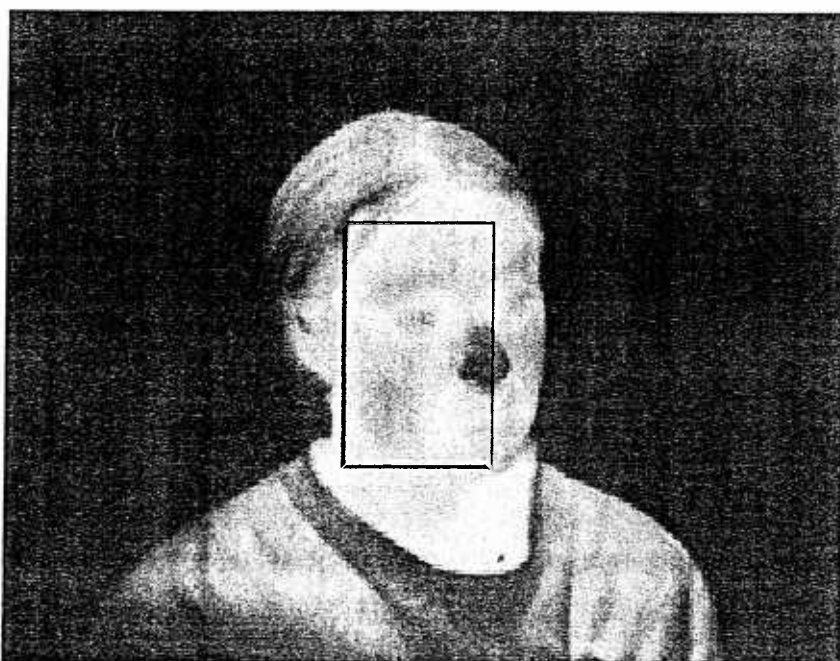


Figure 5.2.2. The area of pixels in the last frame in which our results were checked.

The number of pixels of the last frame in this area which were assigned to each of Normal distributions accounting for human body, background and grey zone were then counted. The results for Jeffrey's number, Kullback-Leibler divergence, Hellinger H_2 , L_2 , Bhattacharyya and Chi-Squared divergence are presented in Table 5.2.1. Inside the parentheses the corresponding percentages are given.

	L2	H2	Chi	BHAT	KULLB	JEFFR
wf	2322 (0.4334)	4307 (0.8038)	4310 (0.8044)	4307 (0.8038)	4562 (0.8514)	4561 (0.8513)
wb	1015 (0.1894)	66 (0.0123)	66 (0.0123)	66 (0.0123)	79 (0.0147)	79 (0.0147)
wgr	2021 (0.3772)	985 (0.1838)	982 (0.1833)	985 (0.1838)	717 (0.1338)	718 (0.1340)
wfalse	3036 (0.5666)	1051 (0.1962)	1048 (0.1956)	1051 (0.1962)	796 (0.1486)	797 (0.1487)
wtrue	2322 (0.4334)	4307 (0.8038)	4310 (0.8044)	4307 (0.8038)	4562 (0.8514)	4561 (0.1487)
n	5358	5358	5358	5358	5358	5358

Table 5.2.1. The number of pixels of the last frame in the area of interest which were assigned to human body, background and grey zone for some divergences, with the corresponding percentages.

where:

wf: the number of pixels into the specific area which were matched to the distribution of face.

wb: the number of pixels into the specific area which were matched to the distribution of background.

wgr: the number of pixels into the specific area which were matched to the distribution of grey zone.

wfalse: the number of pixels into the specific area which were wrongly matched. ($wfalse = wb + wgr$)

wtrue: the number of pixels into the specific area which were correctly matched. ($w_{true} = w_f$)

n: the number of pixels into the area

Next, in figure 5.2.3 the assignment of the last frame is given after applying the L_2 , H_2 , Kullback Leibler divergence, Jeffrey's divergence, Bhattacharrya and Chi-Squared divergence.

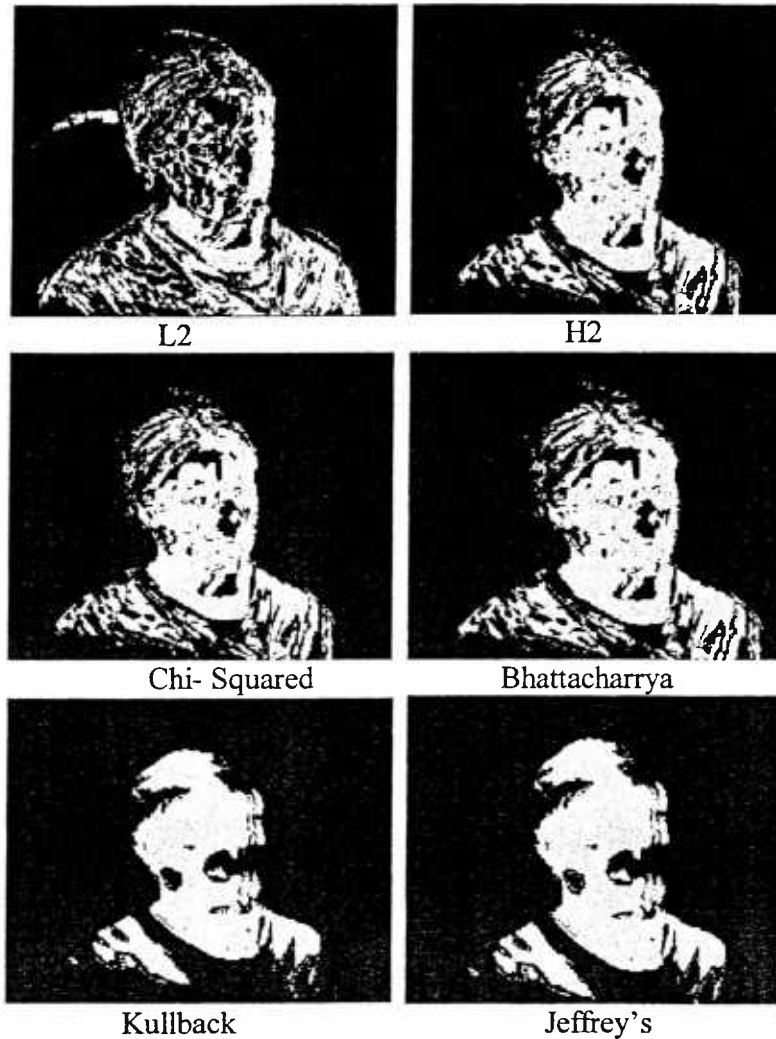


Figure 5.2.3. The classification of last frame's pixels using L_2 , H_2 , Kullback divergence, Jeffrey's divergence, Bhattacharrya and Chi-Squared divergence.

It is obvious from the Table above that L_2 did not manage at all to assign the skin pixels to proper distribution. Kullback and Jeffrey's divergence measures provided similar results, which were the best among all the other measures.

We then checked the results provided by the use of different ρ s in the formulas of method of moments instead of $\rho = \frac{w_2}{w_1 + w_2}$ when using Jeffrey's divergence. If we name

FORMULA 1:
$$\rho = \frac{w_2 \sigma_2}{w_1 \sigma_1 + w_2 \sigma_2},$$

FORMULA 2:
$$\rho = \frac{\sqrt{w_2 \sigma_2}}{\sqrt{w_1 \sigma_1} + \sqrt{w_2 \sigma_2}},$$

FORMULA 3:
$$\rho = \frac{w_2 \sqrt{\sigma_2}}{w_1 \sqrt{\sigma_1} + w_2 \sqrt{\sigma_2}},$$

FORMULA 4:
$$\rho = \frac{\mu_2 w_2 \sqrt{\sigma_2}}{\mu_1 w_1 \sqrt{\sigma_1} + \mu_2 w_2 \sqrt{\sigma_2}},$$

FORMULA 5:
$$\rho = \frac{\exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right)}{\exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2}\right) + \exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right)},$$

FORMULA 6:
$$\rho = \frac{\exp\left(-\frac{w_2 (\mu_1 - \mu_2)^2}{\sigma_2^2}\right)}{\exp\left(-\frac{w_1 (\mu_1 - \mu_2)^2}{\sigma_1^2}\right) + \exp\left(-\frac{w_2 (\mu_1 - \mu_2)^2}{\sigma_2^2}\right)},$$

then the number of pixels of the last frame in the area defined earlier which were assigned to each of Normal distributions accounting for human body, background and grey zone along with the corresponding percentages are given in Table 5.2.2.

	FORM 1	FORM 2	FORM 3	FORM 4	FORM 5	FORM 6
wf	4419 (0.8247)	4598 (0.8582)	4612 (0.8608)	4834 (0.9022)	4908 (0.9160)	4923 (0.9188)
wb	62 (0.0116)	74 (0.0138)	63 (0.0118)	91 (0.0170)	80 (0.0149)	97 (0.0181)
wgr	877 (0.1637)	686 (0.1280)	683 (0.1275)	433 (0.0808)	370 (0.0691)	338 (0.0631)
wfalse	939 (0.1753)	760 (0.1418)	746 (0.1392)	524 (0.0978)	450 (0.0840)	435 (0.0812)
wtrue	4419 (0.8247)	4598 (0.8582)	4612 (0.8608)	4834 (0.9022)	4908 (0.9160)	4923 (0.9188)
n	5358	5358	5358	5358	5358	5358

Table 5.2.2. The number of pixels of the last frame in the area of interest which were assigned to human body, background and grey zone for some formulas for ρ , with the corresponding percentages.

In figure 5.2.4 the assignment of the last frame is given after applying each one of formulas 1, 2, 3, 4, 5, and 6 for ρ (when the Jeffreys divergence measure is used).

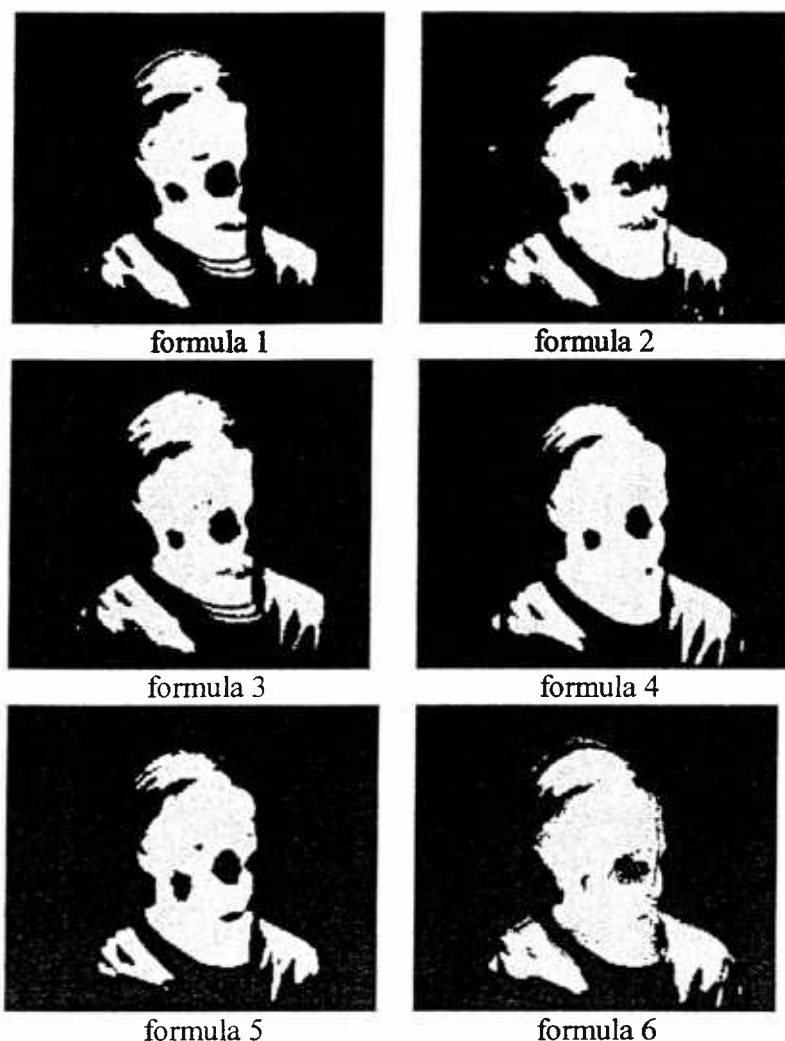


Figure 5.2.4. The classification of last frame's pixels using formulas 1, 2, 3, 4, 5 and 6 for ρ .

Comparing the results obtained by the various formulas of ρ with the corresponding result of Jeffrey's number given in Table 5.2.2 (where

$$\rho = \frac{w_2}{w_1 + w_2} \text{ was used) we can see that all formulas except for formula 1}$$

provided greater percentages of correct ascriptions than $\rho = \frac{w_2}{w_1 + w_2}$.

Obviously, the best results were obtained when formulas 4, 5 and 6 were used.



CHAPTER 6

CONCLUSIONS

6.1 Conclusions

We have presented a general problem of image segmentation when images come from a video sequence. The data values were the temperature of each pixel. First of all we isolated the first frame's pixels that indicated human body from those ones that indicated background and those ones that indicated grey zone. Two different cluster analysis techniques were employed: the k-means and the EM method. After this first phase we have provided statistically valid values for the temperatures of first frame's pixel corresponding to the scene. Then, each pixel x_j of the first frame was considered to be a mixture of the three Normal distributions:

$$x_j \sim w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + w_3 N(\mu_3, \sigma_3^2)$$

Next phase was concerned with the classification of next incoming frames' pixels. We assumed that every new pixel comes from a Normal distribution $N(\mu_p, \sigma_p^2)$ and measured the distance between the distribution of the incoming pixel $N(\mu_p, \sigma_p^2)$ and each one of the three Normal distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$. Several distance measures were presented and applied. We concluded that the best results were provided by Jeffrey's divergence.

After finding which of the distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, $N(\mu_3, \sigma_3^2)$ is closest to the distribution of the incoming pixel $N(\mu_p, \sigma_p^2)$ we had to



approximate those two ‘closest’ densities with one single distribution. So, the next issue was to define the parameters of this new distribution. If $f_1 \sim N(\mu_1, \sigma_1^2)$ with weight w_1 and $f_2 \sim N(\mu_2, \sigma_2^2)$ with weight w_2 are the two distributions merged we need to approximate their mixture with a single Normal distribution $f \sim N(\mu, \sigma^2)$ with weight w . The method of moments provides the weight, mean and variance of the new distribution by:

$$w = w_1 + w_2$$

$$\mu = (1 - \rho)\mu_1 + \rho\mu_2$$

$$\sigma^2 = (1 - \rho)\sigma_1^2 + \rho\sigma_2^2 + (1 - \rho)\rho(\mu_1 - \mu_2)^2$$

where $\rho = \frac{w_2}{w}$. However, we tried some other formulas for ρ , instead of

$\rho = \frac{w_2}{w}$ which provided slightly better results.

6.2 Future research

All of the above open up a wide range of issues that require further research. Most notably the fact that the pixels of each frame considered to be independent from their neighborhood pixels violates the smoothness of the visual world. Use of spatial statistics along with some Bayesian approach would probably be the next step in improving the segmentation while keeping in mind that we are interested in maintaining the results online.

Appendix

Presentation of the Algorithm

The algorithm presented below was written in Matlab. The first part is a function which calculates the weights, mean values and variances of the groups created by k-means method. It accepts as input the values of temperatures of first frame's pixels.

```
function [W0,M0,S0]=EM_initial3(X);

% Calculates the k-means estimates.
% They will be used as initial values in EM.

km=kmeans(X,3);
j=1;
for i=1:80772
    if km(i)==1
        km1(j)=X(i);
        j=j+1;
    end
end

j=1;
for i=1:80772
    if km(i)==2
        km2(j)=X(i);
        j=j+1;
    end
end

j=1;
for i=1:80772
    if km(i)==3
        km3(j)=X(i);
        j=j+1;
    end
end

[N c]=size(X);
[r1 c1]=size(km1);
[r2 c2]=size(km2);
[r3 c3]=size(km3);
c1+c2+c3;

w10=c1/(c1+c2+c3);
w20=c2/(c1+c2+c3);
w30=c3/(c1+c2+c3);
```



```
W0=[w10;w20;w30];
```

```
m10=mean(km1);
m20=mean(km2);
m30=mean(km3);
```

```
M0=[m10;m20;m30];
```

```
s10=sqrt(var(km1));
s20=sqrt(var(km2));
s30=sqrt(var(km3));
```

```
S0=[s10;s20;s30];
```

The returned k-means estimates of the above function are stored into 3 matrices called W0, M0, S0.

The next function gives the image of the woman after the isolation of pixels that indicate human body from those ones that indicate background and those ones that indicate grey zone after the implementation of k-means method.

```
function [km]=km3_image_frame1(X);
```

```
% Gives the image of frame 1 after kmeans method.
```

```
km=kmeans(X,3);
imagesc(reshape(km,254,318));
```

After that, the algorithm continues with a function which calculates the EM estimates of the weights, means and variances. This function accepts as input the values of temperatures of first frame's pixels and the algorithm's terminating condition. The function calls as a subfunction the one that calculates the k-means estimates, i.e. it uses W0, M0, S0. The returned estimates are stored in matrices Wem, Mem, Sem.

```
function [Wem,Mem,Sem] = EM3(X,k);
```

```
% Provides the EM estimates for the weights, means and variances of 3 Normal distributions.
% You must provide the data set and the terminating condition.
```

```
[W0,M0,S0]=EM_initial3(X);
[N c]=size(X);
W0V=[W0(1)*ones(1,N);W0(2)*ones(1,N);W0(3)*ones(1,N)];
M0V=[M0(1)*ones(1,N);M0(2)*ones(1,N);M0(3)*ones(1,N)];
S0V=[S0(1)*ones(1,N);S0(2)*ones(1,N);S0(3)*ones(1,N)];
S01=S0V.^(-1);
S02=(-1/2)*S0V.^(-2);
ZZ=[X';X';X'];
```

```
Y0=W0V.*S01.*exp(S02.*(ZZ-M0V).^2);
```

```

SUMY0=[sum(Y0);sum(Y0);sum(Y0)];
Z0=Y0./SUMY0;
W1=(sum(Z0'))'./N;
M1=(Z0*X)./(N*W1);
M1V=[M1(1,1)*ones(1,N);M1(2,1)*ones(1,N);M1(3,1)*ones(1,N)];
SV=Z0.*(ZZ-M1V).^2;
S1=sqrt((sum(SV'))'./N);
while abs(W1(1,1)-W0(1,1))>k | abs(W1(2,1)-W0(2,1))>k | abs(W1(3,1)-W0(3,1))>k
W0=W1;
M0=M1;
S0=S1;

W0V=[W0(1,1)*ones(1,N);W0(2,1)*ones(1,N);W0(3,1)*ones(1,N)];
M0V=[M0(1,1)*ones(1,N);M0(2,1)*ones(1,N);M0(3,1)*ones(1,N)];
S0V=[S0(1,1)*ones(1,N);S0(2,1)*ones(1,N);S0(3,1)*ones(1,N)];
S01=S0V.^(-1);
S02=(-1/2)*S0V.^(-2);
Y0=W0V.*S01.*exp(S02.*(ZZ-M0V).^2);
SUMY0=[sum(Y0);sum(Y0);sum(Y0)];
Z0=Y0./SUMY0;
W1=(sum(Z0'))'./N;
M1=(Z0*X)./(N*W1);
M1V=[M1(1,1)*ones(1,N);M1(2,1)*ones(1,N);M1(3,1)*ones(1,N)];
SV=Z0.*(ZZ-M1V).^2;
S1=sqrt((sum(SV'))'./N);
end

Wem=W1;
Mem=M1;
Sem=S1;

```

The next two functions give the image of the woman after the isolation of pixels that indicate human body from those ones that indicate background and those

ones that indicate grey zone after the implementation of EM method and the differences between the results of k-means and EM method.

```
function [WemV,MemV,SemV,im]=EM3image_fr1(X,k);

% Gives the image of frame 1 after EM algorithm.
% It is based upon the posterior probabilities.
% The terminating condition must be given.

[Wem,Mem,Sem] = EM3(X(:,1),k);
[N c]=size(X);
WemV=[Wem(1,1)*ones(1,N);Wem(2,1)*ones(1,N);Wem(3,1)*ones(1,N)];

MemV=[Mem(1,1)*ones(1,N);Mem(2,1)*ones(1,N);Mem(3,1)*ones(1,N)];

SemV=[Sem(1,1)*ones(1,N);Sem(2,1)*ones(1,N);Sem(3,1)*ones(1,N)];
Sem1=SemV.^(-1);
Sem2=(-1/2)*SemV.^(-2);

ZZ=[X(:,1)';X(:,1)';X(:,1)'];

Yem=WemV.*Sem1.*exp(Sem2.*(ZZ-MemV).^2);

SUMYem=[sum(Yem);sum(Yem);sum(Yem)];

Z1NEWem=Yem./SUMYem;

[i,j]=max(Z1NEWem);

im=j;
for i=1:N
if j(i)==2
im(i)=3;
else if j(i)==3
im(i)=2;
end
end
end

imagesc(reshape(im,254,318));
```

```
function [km_EM]=dif_km_EM3(X,k);
```

% Gives the differences in frame 1 between the kmeans's and EM' results.

```
[km]=km3_image_frame1(X(:,1));
[WemV,MemV,SemV,im]=EM3image_fr1(X(:,1),k);
km_EM=abs(im-km');
imagesc(reshape(km_EM,254,318));
```

Next, function JEF_3cl is given which calculates the Jeffrey's divergences of the data of a frame from each one of the Normal distributions which have the



parameters obtained after the implementation of EM method. It takes as inputs the two vectors of the data of the first two frames.

```
function [J]=JEF_3cl(X,Y,k);
```

```
% Calculates the Jeffreys divergence of the data of a frame from each one of the distributions  
% of a mixture.
```

```
Z2V=[Y';Y';Y'];  
[N c]=size(X);  
SXV=0.025.*ones(3,N);  
[WemV,MemV,SemV]=EM3image_fr1(X,k);  
  
J=(1/2).*(SXV./SemV-SemV./SXV).^2+((Z2V-MemV).^2./2).*(1./(SXV.^2)+1./(SemV.^2));
```

Similarly to the above function, functions dL2_3cl, dH2_3cl, dK_3cl, BHAT_3cl, CHI_3cl were applied to the data which calculated the L2, H2, Kullback, Bhattacharyya and Chi-squared distances. They differ from JEF_3cl function only to the last row which becomes correspondingly for each function:

```
L2=sqrt((ones(3,N)./sqrt(2*pi*ones(3,N))).*(ones(3,N)./sqrt(2*SXV.^2)+  
ones(3,N)./sqrt(2*SemV.^2)-2*(ones(3,N)./sqrt(SXV.^2+SemV.^2)).*  
exp(-(Z2V-MemV).^2./(2*(SXV.^2+SemV.^2)))));  
  
H2=sqrt(2*ones(3,N)-2*sqrt(2)*sqrt(SXV.*SemV./(SXV.^2+SemV.^2)).*  
exp(-(Z2V-MemV).^2./(4*(SXV.^2+SemV.^2))));  
  
K=log(SXV./SemV)-(1/2).*(ones(3,N)+(SemV.^2)/(2*SXV.^2)+  
((MemV-Z2V).^2)/(2.*(SXV.^2)));  
  
B=(1/2).*log((SXV.^2+SemV.^2)/(2.*sqrt(SXV.^2.*SemV.^2)))+  
(Z2V-MemV).^2/(4.*(SXV.^2+SemV.^2));  
  
C=(SemV.^2./(SXV.*sqrt(2.*SemV.^2-SXV.^2))).*  
exp((Z2V-MemV).^2/(2.*SemV.^2-SXV.^2))-ones(3,N);
```

Next function gives the assignment of second frame's pixels according to Jeffrey's divergences. Similar functions could have been constructed to give the assignment of second frame's pixels according to other divergences as well. The function below accepts as inputs the values of temperatures of first and second frames' pixels, i.e. the two first columns of our matrix of data.

```
function winners_frame2_cl3(X,Y,k);
```

```
% Gives the image of the winners (the distribution of minimum distance)
```

```
[J]=JEF_3cl(X,Y,k);  
[p,r]=min(J);  
imagesc(reshape(r,254,318));
```

In what follows the function which makes the updating of the parameters of the Normal distribution which approximates the mixture of the two 'closest' Normals

according to the method of moments is presented. It uses Jeffrey's divergences. It accepts as inputs the whole matrix of the data of all frames X, the EM algorithm's terminating condition k and the learning parameter which stands for the weight of the distribution of the incoming pixel a.

```
function [WNEW,MNEW,S2NEW]=updatingJ_3cl(X,k,a);

% Gives the parameters of the distributions of each one of the pixels using Jeffrey's
% divergence.

[Wem,Mem,Sem] = EM3(X(:,1),k);
[WemV,MemV,SemV]=EM3image_fr1(X(:,1),k);
[N c]=size(X(:,1));
Z_fr2=[X(:,2)';X(:,2)';X(:,2)'];
SXV=0.025.*ones(3,N);
[J]=JEF_3cl(X(:,1),X(:,2),k);
[p,r]=min(J);
R=full(sparse(r,1:N,ones(1,N)));

WNEW=WemV-a.*WemV+a.*R;
p=((a.*ones(3,N))./WNEW).*R;
MNEW=(ones(3,N)-p).*MemV+p.*Z_fr2;
S2NEW=(ones(3,N)-p).*SemV.^2+p.*SXV.^2+p.*(ones(3,N)-p).*(MemV-Z_fr2).^2;
```

Similar functions were applied to the data which used the L2, H2, Kullback, Bhattacharyya and Chi-squared distances. Moreover, the next function makes the update using Jeffrey's divergence and calculates ρ by formula $\rho = \frac{w_2 \sigma_2}{w_1 \sigma_1 + w_2 \sigma_2}$ instead of the method of moments.

```
function [WNEW,MNEW,S2NEW]=updatingJ_3cl_form1(X,k,a);

% Gives the parameters of the distributions of each one of the pixels.

[Wem,Mem,Sem] = EM3(X(:,1),k);
[WemV,MemV,SemV]=EM3image_fr1(X(:,1),k);
[N c]=size(X(:,1));
Z_fr2=[X(:,2)';X(:,2)';X(:,2)'];
SXV=0.025.*ones(3,N);
[J]=JEF_3cl(X(:,1),X(:,2),k);
[p,r]=min(J);
R=full(sparse(r,1:N,ones(1,N)));

WNEW=WemV.*SemV-(a.*WemV).*SemV+(a.*R).*SXV;
p=((a.*ones(3,N)).*SXV./WNEW).*R;
MNEW=(ones(3,N)-p).*MemV+p.*Z_fr2;
S2NEW=(ones(3,N)-p).*SemV.^2+p.*SXV.^2+p.*(ones(3,N)-p).*(MemV-Z_fr2).^2;
```

For using formula $\rho = \frac{\sqrt{w_2\sigma_2}}{\sqrt{w_1\sigma_1} + \sqrt{w_2\sigma_2}}$ the 11th and 12th line of the above

function become:

$$\begin{aligned} \text{WNEW} &= \text{sqrt}(\text{WemV} * \text{SemV} - (a * \text{WemV}) * \text{SemV}) + \text{sqrt}((a * R) * \text{SXV}); \\ p &= (\text{sqrt}((a * \text{ones}(3, N)) * \text{SXV}) / \text{WNEW}) * R; \end{aligned}$$

For using formula $\rho = \frac{w_2\sqrt{\sigma_2}}{w_1\sqrt{\sigma_1} + w_2\sqrt{\sigma_2}}$ those two lines become:

$$\begin{aligned} \text{WNEW} &= \text{WemV} * \text{sqrt}(\text{SemV}) - (a * \text{WemV}) * \text{sqrt}(\text{SemV}) + (a * R) * \text{sqrt}(\text{SXV}); \\ p &= (((a * \text{ones}(3, N)) * \text{sqrt}(\text{SXV})) / \text{WNEW}) * R; \end{aligned}$$

For using formula $\rho = \frac{\mu_2 w_2 \sqrt{\sigma_2}}{\mu_1 w_1 \sqrt{\sigma_1} + \mu_2 w_2 \sqrt{\sigma_2}}$ those two lines become:

$$\begin{aligned} \text{WNEW} &= \text{MemV} * \text{sqrt}(\text{WemV} * \text{SemV} - (a * \text{WemV}) * \text{SemV}) + Z_fr2 * \text{sqrt}((a * R) * \text{SXV}); \\ p &= (Z_fr2 * \text{sqrt}((a * \text{ones}(3, N)) * \text{SXV}) / \text{WNEW}) * R; \end{aligned}$$

For using formula $\rho = \frac{\exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right)}{\exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2}\right) + \exp\left(-\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right)}$ those two

lines become:

$$\begin{aligned} \text{WNEW} &= \exp(((-1/2) * (\text{MemV} - Z_fr2).^2) / \text{SemV}.^2) + R * \exp(((-1/2) * (\text{MemV} - \\ &\quad Z_fr2).^2) / \text{SXV}.^2); \\ p &= (R * \exp(((-1/2) * (\text{MemV} - Z_fr2).^2) / \text{SXV}.^2)) / \text{WNEW}; \end{aligned}$$

For using formula $\rho = \frac{\exp\left(-\frac{w_2 (\mu_1 - \mu_2)^2}{\sigma_2^2}\right)}{\exp\left(-\frac{w_1 (\mu_1 - \mu_2)^2}{\sigma_1^2}\right) + \exp\left(-\frac{w_2 (\mu_1 - \mu_2)^2}{\sigma_2^2}\right)}$ those

two lines become:

$$\begin{aligned} \text{WNEW} &= \exp((- \text{WemV} * (\text{MemV} - Z_fr2).^2 + (a * \text{WemV}) * (\text{MemV} - \\ &\quad Z_fr2).^2) / \text{SemV}.^2) + R * \exp((- a * (\text{MemV} - Z_fr2).^2) / \text{SXV}.^2); \\ p &= (R * \exp((- a * (\text{MemV} - Z_fr2).^2) / \text{SXV}.^2)) / \text{WNEW}; \end{aligned}$$

Next a function which makes the video of our data is given. It accepts as inputs the matrix of our data, the name we want to give to the video and the duration of it.

```
function makeavi(Z,name,sec)

n=size(Z,2);
fps=min(n/sec,n);
aviobj = avifile(name,'compression','Cinepak','fps',fps,'quality',75)
for i=1:n;
    imagesc(reshape(Z(:,i),254,318));
    frame = getframe(gca);
    aviobj = addframe(aviobj,frame);
end;
aviobj
aviobj=close(aviobj);
close;
```

Using the above function and all the previous functions, a new function is presented which repeats the same procedure for all frames and makes the video where for each frame the pixels are assigned to that normal from which they desist less.

```
function winners_imagesJef_3cl(X,k,a,name,sec);

% Gives the images of the winners through frames:2_98
% (the distribution of minimum Jeffrey's distance)
% set k=1e-10, a=0.05, sec=30
[WNEW,MNEW,S2NEW]=updatingJ_3cl(X,k,a);
[N c]=size(X(:,1));
SXV=0.025.*ones(3,N);

[J1_2]=JEF_3cl(X(:,1),X(:,2),k);
[p,v]=min(J1_2);
Vid(:,1)=v';

for fr=3:98
    Z2V=[X(:,fr)';X(:,fr)';X(:,fr)'];
    J=(1/2).*(SXV./(sqrt(S2NEW))-(sqrt(S2NEW))./SXV).^2+((Z2V-
    MNEW).^2./2).*(1./SXV.^2+1./S2NEW);
    [p,r(fr-2,:)] = min(J);
    Vid(:,fr-1)=r(fr-2,:);
    R=full(sparse(r(fr-2,:),1:N,ones(1,N)));

    WNEW=WNEW-a*WNEW+a*R;
    p=((a*ones(3,N))./WNEW).*R;
    MNEW=(ones(3,N)-p).*MNEW+p.*Z2V;
    S2NEW=(ones(3,N)-p).*S2NEW+p.*SXV.^2+p.*(ones(3,N)-p).*(MNEW-Z2V).^2;

end

makeavi(Vid,name,sec);
```



That function uses Jeffrey's divergence and the method of moments. Similar functions were applied to the data using other measures of divergence and different

$$\rho \text{'s instead of } \rho = \frac{w_2}{w_1 + w_2}.$$

In order to check the result obtained an area was chosen which in all frames actually included skin pixels. Next function counts the number of pixels of the last frame which were assigned to each one of Normal distributions accounting for human body, background and grey zone. It also gives the corresponding percentages.

```
function [ww1,ww2,ww3,pf,pb,pgr]=perc_square3Jef_3cl(X,k,a);
% Gives the images of the winners through frames:2_98
% (the distribution of minimum Jeffreys distance)
% set k=1e-10, a=0.05, sec=30
[WNEW,MNEW,S2NEW]=updatingJ_3cl(X,k,a);
[N c]=size(X(:,1));
SXV=0.025.*ones(3,N);

for fr=3:98
Z2V=[X(:,fr)';X(:,fr)';X(:,fr)'];
J=(1/2).*(SXV./(sqrt(S2NEW))-(sqrt(S2NEW))./SXV).^2+((Z2V-
MNEW).^2./2).*(1./SXV.^2+1./S2NEW);
[p,r(fr-2,:)] = min(J);
R=full(sparse(r(fr-2,:),1:N,ones(1,N)));

WNEW=WNEW-a*WNEW+a*R;
p=((a*ones(3,N))./WNEW).*R;
MNEW=(ones(3,N)-p).*MNEW+p.*Z2V;
S2NEW=(ones(3,N)-p).*S2NEW+p.*SXV.^2+p.*(ones(3,N)-p).*(MNEW-Z2V).^2;

end

gr=r(96,:);
ww1=0;
ww2=0;
ww3=0;
for p=127:183
    for pp=85:178
        if r(96,p*254+pp)==1
            ww1=ww1+1;
        elseif r(96,p*254+pp)==2
            ww2=ww2+1;
        else
            ww3=ww3+1;
        end
    end
end
end

n=(183-127+1)*(178-85+1);
pf=max(max(ww1,ww2),ww3)/n;
pb=min(min(ww1,ww2),ww3)/n;
pgr=(n-max(max(ww1,ww2),ww3)-min(min(ww1,ww2),ww3))/n;
```



The above function uses Jeffrey's divergence and the method of moments. Similar functions were applied to the data using other measures of divergence and different ρ 's instead of $\rho = \frac{w_2}{w_1 + w_2}$.



References

- Adhikari, B. P. and Joshi, D. D. (1956).** Distance, discrimination et resume exhaustif. *Publ. Inst. Statist. Univ. Paris*, 5, 57-74
- Ali, S. M. and Silvey, S. D. (1966).** A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society*, B, 28, 131-142
- Amari, S. (1982).** Differential geometry of curved exponential families- curvatures and information loss. *Annals of Statistics*, 10, 357-385
- Amari, S. (1985).** *Differential geometric methods in statistics*. New York: Springer-Verlag
- Basseville, M. (1988).** Distance measures for signal processing and pattern recognition. *Institute de recherché en informatique et systemes aleatoires*, 349-369
- Bretagnolle, J and Huber, C. (1979)** Estimation des densites. *Risque minimax. Z. Wahrsch. Verw. Gebiete*, 47, 119-137
- Chernoff, H. (1952).** A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23, 493-507
- Csiszar, I. (1967a).** I-Divergence Geometry of Probability Distributions and Minimization Problems. *Annals of Probability*, 3, 146-158
- Csiszar, I. (1967b).** Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungar*, 2, 299-318
- Dempster, A. P., Laird, N. M. and Rubin D. B. (1977)** Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Statist. Soc.*, B, 39, 1-38
- Devroye L. (1987)** A course in density estimation, Boston: Birkhaeuser
- Grimson, W., Stauffer, C., Romano, R. and Lee, L. (1998).** Using adaptive tracking to classify and monitor activities in a site. In: *Proceedings of the 1998 IEEE conference on computer vision and pattern recognition*, Santa Barbara, CA, 23-25 June 1998, pp 22-29



- Hasseblad, V. (1966)**, Estimation of Parameters for a Mixture of Normal Distributions. *Technometrics*, 3, 431-444.
- Hasseblad, V. (1969)**, Estimation of Finite Mixtures from the Exponential Family. *Journal of the American Statistical Association*, 64, 1459-1471.
- Hero, A. O., Ma, B., Michel, O. and Gorman, J. (2001)**. Alpha-Divergence for classification, indexing and retrieval. *Communications and Signal Processing Laboratory Technical Report CSPL-328*
- Jeffreys, H. (1948)** Theory of Probability, Second Edition, University Press, Oxford.
- Karlis, D. (2004)** Notes about the course “Multivariate techniques”, Athens University of Economics and Business, 22.
- Kemperman, J.H.B. (1969)** An optimum rate of transmitting information. lower bound for discrimination in terms of variation. *Annals of Mathematical Statistics*, 40, 2156-2177
- Kolmogorov, A. N. (1963)**. On the approximation of distributions of sums of independent summands by infinitely divisible distributions. *Sankhya*, 25, 159-174
- Kullback, S. and Leibler, R. A. (1951)** Of information and Sufficiency. *Annals of Mathematical Statistics*, 22, 79-86
- Kullback, S. (1959)** *Information theory and statistics*. New York: Dover Publications
- Kullback, S. (1967)** A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory*, 13, 126-127
- Lehmann, E. L. (1959)**. *Testing Statistical Hypotheses*. New York: Wiley
- Lin, J. (1991)** Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37, 145-151
- McLachlan, G. J. and Basford, K. E. (1988)**. *Mixture Models: inference and application to clustering*. New York: Marcel Dekker
- McLachlan, G. J. (1992)**. *Discriminant analysis and statistical pattern recognition*. New York: Wiley
- McLachlan, G. J. and Krishnan, N. (1997)**. *The EM algorithm and its extensions*. New York: Wiley



- Morellas, V., Pavlidis, I. and Tsiamyrtsis, P. (2003).** DETER: Detection of events for threat evaluation and recognition. *Machine Vision and Application*, 15, 29-45
- Onishi, K. and Imai, H. (1997).** Voronoi diagram in statistical parametric space by Kullback-Leibler divergence. *Proceedings of the thirteenth annual symposium on Computational geometry*, 463-465
- Pearson, K. (1894).** Contribution to the mathematical theory of evolution. *JPhil. Trans. A*, 185, 71-110
- Pednekar, A., Kakadiaris, I. and Kurkure, U. (2002).** Adaptive fuzzy connectedness-Based medical image segmentation. *Proceedings of the Indian Conference on Computer Vision, Graphics, and Image Processing*, 457-462
- Rao, C. R. (1952).** *Advanced Statistical Methods in Biometric Research*. New York: Wiley
- Rigau, J., Feixas, M., Sbert M. (2003).** Refinement criteria based on f -divergence. *Proceedings of the 13th Eurographics Symposium on Rendering*, 260-269
- Toussaint G. T. (1975)** Sharper lower bounds for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, IT-21, 99-100
- Vajda, I. (1970)** Note on discrimination information and variation. *IEEE Transactions on Information Theory*, IT-16, 771-773
- Vajda, I. (1972)** On the f -divergence and singularity of probability measures. *Periodica Mathem. Hungarica*, 2, 223-234
- Zhang, J. (2004).** Divergence Function, Duality and Convex Analysis. *Neural Computation*, 16, 159-195



