Statistical Modeling for Simultaneous Data Clustering, Features Selection, and Outliers Rejection

Khaled Almakadmeh

A Thesis

in

The Concordia Institute

for

Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science in Information Systems Security at

Concordia University

Montréal, Québec, Canada

May 2010

Your file Votre référence
ISBN: 978-0-494-71008-1
Our file Notre référence
ISBN: 978-0-494-71008-1

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Khaled Almakadmeh

Entitled: Statistical Modeling for Simultaneous Data Clustering Features Selection and Outliers Rejection

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science In Information Systems Security

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Simon Li ________ Chair

Dr. Jamal Bentahar ________ Examiner

Dr. Abdelwahab Hamou-Lhadj ________ Examiner

Dr. Nizar Bouguila ________ Supervisor

Approved by ________
Chair of Department or Graduate Program Director

____
Dean of Faculty

Date ________

Abstract

Statistical Modeling for Simultaneous Data Clustering, Features Selection, and Outliers Rejection

Khaled Almakadmeh

Model-based approaches and in particular finite mixture models are widely used for data clustering, which is a crucial step in several applications of practical importance. Indeed, many pattern recognition, computer vision, and image processing applications can be approached as feature space clustering problems. However, the use of these approaches for complex high-dimensional data presents several challenges such as the presence of many irrelevant features, which may affect the speed, and compromise the accuracy of the used learning algorithm. Another problem is the presence of outliers which potentially influence the resulting model parameters. Generally, clustering, features selection, and outliers detection problems have been approached separately. In this thesis, we propose a unified statistical framework to address the three problems simultaneously. The proposed statistical model partitions a given data set without a priori information about the number of clusters, the saliency of the features, or the number of outliers. We illustrate the performance of our approach using different applications involving synthetic data, real data, and objects shape clustering.

Acknowledgements

First and foremost, I would like to say that there are no words to express my greatest gratitude to my supervisor Dr. Nizar Bouguila. He has proven to be a very supportive advisor, mentor and motivator. From his valuable tutoring, I have not only gained technical knowledge, but I have also learned to handle real life situations. I am grateful to him for his guidance, and persistent confidence in me.

Furthermore, I wish to extend my sincere appreciation to all the professors in CIISE, especially Prof. Mourad Debbabi and Dr. A. Ben Hamza for their assistance. I was always treated with the type of respect and kindness that I believed could only be offered to one's own son.

Thanks to my colleagues in the lab, for their helpful suggestions during my two years at Concordia University.

Finally, I thank my family, especially my father and mother, for their unconditional support throughout my studies. Your endless love and care continue to encourage me.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Introduction and Related works

A recurring subject in pattern recognition and computer vision applications is the separation of data (images, videos, objects, etc.) into homogeneous clusters [1, 2]. This topic has been extensively studied and different approaches and algorithms have been proposed and applied to several problems. Generally, an important step in these problems is the representation of a given image by a vector of features which is generally high-dimensional [3]. Although different, the majority of approaches agree that a good clustering model should be sensitive to the extracted features but not to the noise (i.e. outliers) which may be present.

A challenging problem in this case is to determine if all the features are necessary, and relevant for the clustering task [4, 5]. Many methods have been developed to estimate the usefulness of features for clustering and prediction. Reducing the number of features not only speeds up the learning and training process, but also prevents over-fitting; allowing the generation of the most optimal model to represent the data and to reflect its regularities [6]. Feature selection is sensitive to the choice of the number of clusters describing the data. Indeed, if the selected number of clusters is very incorrect; the feature selection may be inaccurate as well. Moreover, it is often the case that some of the data is not representative, and may deteriorate clustering performance [7]. Thus, it is crucial to automatically detect these data commonly called "outliers", and which can be described [1] as the observations that do not come from the model [9].

[1] There are many other definitions that have been discussed, for instance, in [8].

To achieve maximum utility and flexibility, finite mixture models are widely used and are well-known for their efficiency in clustering heterogenous data sets. In mixture models, data is supposed to be described by a number of distributions mixed in varying proportions. A fundamental element when using finite mixture models is then the choice of the components densities, which has to take into account the characteristics of the data. Works dealing with finite mixtures vary in the assumptions that they make about the mixture distributions. Generally, data is assumed to be normally distributed [10]. However, the normal distribution isn't always appropriate in pattern recognition, signal and image processing applications [2]. One of the drawbacks is the rigidity of its shape; which prevents it from yielding a good modeling and adequate representation of actual data [12]. Thus, it is important to take into account the underlying structural characteristics of the data domain, and the nature of the patterns that we would like to discover.

This thesis addresses unsupervised learning of data defined over the positive interval $[0, \infty)$, which naturally appear in several image processing, pattern recognition and computer vision applications, in which the Gamma distribution is known to be a good flexible choice, and an accurate alternative to the Gaussian distribution [13–19]. Hence, we propose a novel statistical mixture model based on the Gamma distribution [3], which simultaneously select features by explicitly introducing a common background distribution to explain non-salient features, and detect outliers, by explicitly introducing a new class to model outliers. The background distribution, the outlier component, and their respective weights are repeatedly adjusted to maximize the total integrated likelihood of the model.

Determining the relevant features is one of the central problems in machine learning and pattern recognition, and several approaches have been proposed in the past. However, most of the approaches have focused on the supervised case, which is justified by the difficulty to assess feature relevancy without resorting class labels, and more challenge is added when the number of clusters is unknown. The literature about supervised feature selection is considerable (see, for instance, [21]).

[2] For instance, previous studies of natural images have revealed that their statistics are not Gaussian. Such studies suggest the use of more flexible models (see, for example, [11]).

[3] This distribution has a long history. For instance, in 1895, Karl Pearson directed attention to the Gamma as a model for skewed data [20].

Recently there has been some renewed interest in the problem of selecting features when data are unlabeled, and arise from sampling a mixture of distributions. For instance, the authors in [22] proposed a simultaneous feature selection and clustering algorithm using finite Gaussian mixture model that assume independence of features, and minimum message length (MML) criterion. The MML criterion was used also in [23] where the authors proposed an unsupervised approach for feature selection and extraction to model non-Gaussian data. The author in [24] proposed a probabilistic approach that assigns relevancy weights to discrete features that are considered as random variables modeled by finite discrete mixture models learned using stochastic complexity. In [25] an algorithm for subspace clustering based on expectation-maximization and using a minimum description length (MDL) criterion was proposed to select the relevant features subset and the number of clusters which best describe a continuous Gaussian or discrete data set. All the aforementioned approaches are based on global selection of features (i.e. produce a common feature subset for all the mixture components). Recently, a localized variational feature selection approach has been proposed in [26] for finite Gaussian mixtures, where different feature subsets are associated to the different mixture components, and has been shown to generally outperform global feature selection approach. The effectiveness of these feature selection techniques can be compromised by the presence of outliers. Removing outliers often enhances the generalization ability of the learning algorithms, since it is well-known that even a few outliers can have a dramatic negative effect on estimation [27]. Detection of outliers and the development of approaches insensitive to their presence is an old problem [4], and also has been the goal of much research [32–34]. In particular, in recent years, there has been considerable interest in the effect of outlying observations in finite mixture models. In [35], for instance, a novel statistical mixture model has been proposed to reject outliers in the case of mixed labeled/unlabeled samples. In [36] a robust method that reduces the effect of statistical outliers for parameters estimation under the Gaussian mixture model has been proposed and implemented to classify multi-spectral data. A sequential algorithm for fitting mixture models using an outlier component has been proposed in [37].

It is clear that both problems (feature selection and outlier rejection) have a long history, but have been addressed separately in the past. To the best of our knowledge no models, based on mixture of distributions,

[4] For instance, criteria for the rejection of outliers have been proposed when dealing with univariate Gaussian observations in [28–30]. Also, an approach in the case of univariate Gamma samples has been proposed in [31].

have been proposed to tackle at the same time the problems of feature selection and outlier rejection, which are closely and intimately inter-related in unsupervised settings [5].

1.2 Contributions

The contributions of this thesis are as follows:

☞ **A Novel approach for robust high-dimensional data clustering:** Our approach is simultaneously capable of clustering high-dimensional data sets, by considering the relevancy of features, and the presence of outlier data. Our approach is applied mainly using localized feature selection, and then using globalized feature selection for the purpose of comparison. The integrated likelihood criterion is used to estimate the number of clusters.

☞ **Shape modeling and representation:** We propose a new approach for accurate shape modeling and representation, which is considered as an important step in several applications such as content-based image retrieval, indexing, and segmentation using Zernike Moments Magnitudes ($ZMMs$). Indeed, it allows invariant (i.e. regardless of the position, size, and orientation) recognition of objects.

☞ **Comparison of our approach performance with finite Gaussian mixture model:** We compare the performance of our approach with a finite Gaussian mixture model in terms of clustering, selection of relevant features, and detection of outliers, through applying localized and globalized feature selection methods.

1.3 Thesis Overview

The organization of this thesis is as follows:

❑ Chapter 1 introduces the dilemma of clustering high-dimensional data sets, methods for selecting relevant features, and the presence/effect of outlying data in the model selection process.

[5] It is noteworthy that some techniques have been proposed in the case of subspace clustering and feature extraction (see, for instance, [38, 39])

❑ Chapter 2 proposes a new finite multivariate Gamma mixture model-based approach, which is simultaneously capable of clustering high-dimensional data sets, select relevant features, and remove outlying data.

❑ Chapter 3 presents the experimental results using synthetic and real data sets. We investigate also the problem of 2D shape modeling and clustering.

❑ Chapter 4 summarizes our contributions.

❑ Chapter 5 (Appendices): presents the derivation equations of the proposed model using multivariate mixtures of Gamma and Gaussian distributions. Further, the derivation equations for the integrated likelihood criterion are developed.

CHAPTER 2

Finite Mixture Model for Simultaneous Data Clustering, Features Selection, and Outliers Rejection

2.1 Introduction

In the previous chapter, we presented the dilemma of high-dimensional data clustering, and the difficulties related to this issue, such as determining the relevant features, and the effect of outliers presence on the model selection process. In this chapter, we propose a novel approach for simultaneous high-dimensional data clustering, localized feature selection, and outlier rejection using finite multivariate Gamma mixture model. We present the model learning, maximum likelihood estimation, and model selection based on the integrated likelihood criterion. Finally, we present the expectation-maximization (EM) algorithm, and define the outlier distribution that will detect the outlying data.

2.2 The Model

Let us consider a data set of N vectors $\mathcal{X} = \{\vec{X}_1, \dots, \vec{X}_N\}$, where each $\vec{X}_i = (X_{i1}, \dots, X_{iD})$ is a D-dimensional vector of features representing a given image. Set of vectors generally contains examples that belong to many clusters and can be modeled by a finite mixture of distributions:

$$p(\vec{X}_i|\Theta_M) = \sum_{j=1}^{M} p_j p(\vec{X}_i|\theta_j) \tag{1}$$

where $p_j > 0$, are the mixing proportions, M is the number of mixture components, and $\Theta_M = \{\vec{P} = (p_1, \ldots, p_M), \vec{\theta} = (\theta_1, \ldots, \theta_M)\}$ is the set of parameters in the mixture model. Eq. 1 constitutes actually a family of models which can be viewed as additive models in statistics [40]. A critical problem in this case is the choice of the probability density function to represent each component. In several applications, the vectors$'$ elements X_{id}, $d = 1, \ldots, D$ are positive. In this case one of the most useful probability density functions is the multivariate Gamma. A simple multivariate widely used form assuming independence of variables is given as the following [41, 42] [1]:

$$p(\vec{X}_i|\theta_j) = \prod_{d=1}^{D} p(X_{id}|\theta_{jd}) = \prod_{d=1}^{D} \frac{X_{id}^{\alpha_{jd}-1} \exp(-\frac{X_{id}}{\beta_{jd}})}{\beta_{jd}^{\alpha_{jd}} \Gamma(\alpha_{jd})} \tag{2}$$

where $\theta_j = (\theta_{j1}, \ldots, \theta_{jD}), p(X_{id}|\theta_{jd})$ is the univariate Gamma distribution, and $\theta_{jd} = (\alpha_{jd}, \beta_{jd})$, $\alpha_{jd} > 0$, $\beta_{jd} > 0$, $d = 1, \ldots, D$ represent shape and scale parameters, respectively. Also $\Gamma(.)$ denotes the Gamma function.

As we have mentioned in chapter 1, an important step in pattern recognition applications in general is clustering, and in particular is feature selection. The main objective is to choose only those features which are better suited and relevant for the problem under study (i.e. features should be selected according to their discrimination power). As a classic problem, feature selection has been defined in different ways in the literature (see, for instance, [44, 45]). A widely used approach to perform feature selection, in order to take into account the fact that different features may have different weights, in the case of finite mixture models is to approximate Eq. 1 as following: [22, 23, 25, 46]

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^{M} p_j \prod_{d=1}^{D} \left(\rho_d p(X_{id}|\theta_{jd}) + (1-\rho_d) p(X_{id}|\lambda_d)\right) \tag{3}$$

where $\Theta = \{\Theta_M, \{\rho_d\}, \{\lambda_d\}\}$ is the set of all the model parameters, ρ_d represents the probability that the d^{th} feature is relevant for clustering, and $p(X_{id}|\lambda_d)$ is an univariate Gamma with parameters $\lambda_d =$

[1] Many parameterizations do exist for the Gamma distribution [43]. In this thesis, we consider the parametrization given in Eq. 2.

$(\alpha_{\lambda|d}, \beta_{\lambda|d})$ and can be viewed as a common background distribution to explain nonsalient features. Notice that if $\rho_d = 0, d = 1, \ldots, D$, the model in Eq. 3 is reduced to the one in Eq. 1. Notice also that feature saliency is defined globally (i.e. a given feature is relevant or not to all the mixture components), which can be an invalid assumption in practical clustering problems as shown in [26]. A better approach to take into account the local intrinsic property of the data, which plays an important role [47], by assuming that the relevancy of features is different for different classes which can be modeled as following [26]:

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^{M} p_j \prod_{d=1}^{D} \left(\rho_{jd} p(X_{id}|\theta_{jd}) + (1 - \rho_{jd}) p(X_{id}|\lambda_{jd})\right) \tag{4}$$

where $\Theta = \{\Theta_M, \{\rho_{jd}\}, \{\lambda_{jd}\}\}$, and ρ_{jd} denotes the weight of the d^{th} feature on cluster j and $p(X_{id}|\lambda_{jd})$, $\lambda_{jd} = (\alpha_{\lambda|jd}, \beta_{\lambda|jd})$, is the Gamma distribution from which the feature is drawn if it is irrelevant. Notice that the previous model is reduced to the one in Eq. 3 when $\rho_{jd} = \rho_d$ and $\lambda_{jd} = \lambda_d$, $j = 1, \ldots, M$, $d = 1, \ldots, D$.

Generally no knowledge is available as to which vector $\vec{X}_i$ is not representative and then is not really generated from our assumed statistical model. Outliers do not only make the model learning more complex, but also corrupt the parameters estimates, hence, compromise the performance of the final model. Classic outliers identification methods have been generally based on the sample mean and covariance matrix which are themselves compromised by the outliers as shown in [48], especially in the case of high-dimensional non-Gaussian data. A better technique is to approach the problem by incorporating an auxiliary outlier component to which we associate a uniform density [2] [37, 38, 50, 51] into the model:

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^{M} p_j \prod_{d=1}^{D} \left(\rho_{jd} p(X_{id}|\theta_{jd}) + (1 - \rho_{jd}) p(X_{id}|\lambda_{jd})\right) + p_{M+1} U(\vec{X}_i) \tag{5}$$

where $p_{M+1} = 1 - \sum_{j=1}^{M} p_j$ is the probability that $\vec{X}_i$ was not generated by the central mixture model and $U(\vec{X}_i)$ is a uniform distribution common for all data to model isolated vectors, which are not in any of the M clusters, and which show significantly less differentiation among clusters. The previous model can be viewed as a way to robustify unsupervised feature selection to learn the right meaning from the right observations (i.e. inliers). Notice that when $p_{M+1} = 0$ the outlier component is removed and the previous equation is reduced to Eq. 4.

[2] Some researchers have proposed nonuniform outlier distributions, also (see, for instance, [49]).

2.3 Model Learning

2.3.1 Maximum Likelihood Estimation

A well-known approach for unknown parameters estimation is the technique of maximum likelihood (ML), which properties have been extensively examined in the past (see, for instance, [52]). Using ML the parameters are estimated by maximizing the log-likelihood function as following:

$$\hat{\Theta} = \arg\max_{\Theta} \left\{ \log p(\mathcal{X}|\Theta) = \sum_{i=1}^{N} \log \left[\sum_{j=1}^{M} \left(p_j \prod_{d=1}^{D} \left(\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd}) \right) \right) + p_{M+1} U(\vec{X}_i) \right] \right\} \tag{6}$$

which gives us (See Appendices 1, 2, 3):

$$p_j = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)}{N} \qquad j = 1, \dots, M+1 \tag{7}$$

$$\rho_{jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd})}{\sum_{i=1}^{N} p(j|\vec{X}_i)} \qquad j = 1, \dots, M \quad d = 1, \dots, D \tag{8}$$

$$\beta_{jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd}) X_{id}}{\alpha_{jd} \sum_{i=1}^{N} p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd})} \qquad j = 1, \dots, M \quad d = 1, \dots, D \tag{9}$$

$$\alpha_{jd} = \Psi^{-1} \left(\frac{\sum_{i=1}^{N} p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd}) (\log X_{id} - \log \beta_{jd})}{\sum_{i=1}^{N} p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd})} \right) \qquad j = 1, \dots, M \quad d = 1, \dots, D \tag{10}$$

$$\beta_{\lambda|jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i) f(1 - \rho_{jd}, \theta_{jd}, \lambda_{jd}) X_{id}}{\alpha_{\lambda|jd} \sum_{i=1}^{N} p(j|\vec{X}_i) f(1 - \rho_{jd}, \theta_{jd}, \lambda_{jd})} \qquad j = 1, \dots, M \quad d = 1, \dots, D \tag{11}$$

$$\alpha_{\lambda|jd} = \Psi^{-1} \left(\frac{\sum_{i=1}^{N} p(j|\vec{X}_i) f(1 - \rho_{jd}, \theta_{jd}, \lambda_{jd}) (\log X_{id} - \log \beta_{\lambda|jd})}{\sum_{i=1}^{N} p(j|\vec{X}_i) f(1 - \rho_{jd}, \theta_{jd}, \lambda_{jd})} \right) \qquad j = 1, \dots, M \quad d = 1, \dots, D \tag{12}$$

where $\Psi^{-1}()$ is the inverse digamma function.

$$p(j|\vec{X}_i) = \begin{cases} \dfrac{p_j \prod_{d=1}^{D} \left(\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd}) \right)}{\sum_{j=1}^{M} \left(p_j \prod_{d=1}^{D} \left(\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd}) \right) \right) + p_{M+1} U(\vec{X}_i)} & \text{if } j = 1, \dots, M \\ \dfrac{p_{M+1} U(\vec{X}_i)}{\sum_{j=1}^{M} \left(p_j \prod_{d=1}^{D} \left(\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd}) \right) \right) + p_{M+1} U(\vec{X}_i)} & \text{if } j = M+1 \end{cases} \tag{13}$$

is the posterior probability that a vector $\vec{X}_i$ will be considered as an inlier and then assigned to a cluster $j, j = 1, \ldots, M$, or as an outlier and then affected to cluster $M+1$ which allows to safeguard against erroneous feature selection, and:

$$f(\rho_{jd}, \theta_{jd}, \lambda_{jd}) = \frac{\rho_{jd} p(X_{id}|\theta_{jd})}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})} \tag{14}$$

$$f(1-\rho_{jd}, \theta_{jd}, \lambda_{jd}) = \frac{(1-\rho_{jd}) p(X_{id}|\lambda_{jd})}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})} \tag{15}$$

It is noteworthy that $p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd})$ $\big(p(j|\vec{X}_i) f(1-\rho_{jd}, \theta_{jd}, \lambda_{jd})\big)$ can be viewed as the posterior probability that a given feature d is relevant (irrelevant) for a cluster j.

2.3.2 Model Selection Based on the Integrated Likelihood Criterion

It is desirable to determine the simplest model that can explain the data accurately [3]. It is noteworthy that the simplicity is measured in our case by the number of mixture components, and the number of relevant features. The simplest model can be viewed as the one that maximizes the integrated (or marginal) likelihood, to reach an acceptable balance between model complexity and goodness of fit, given as following when approximated using Laplace's method [53]:

$$\log p(\mathcal{X}) \simeq \log p(\mathcal{X}|\hat{\Theta}) + \log p(\hat{\Theta}) - \frac{1}{2}\log(|H(\hat{\Theta})|) + \frac{N_p}{2}\log(2\pi) \tag{16}$$

where $\hat{\Theta}$ is the posterior mode, $p(\Theta)$ is the prior density for Θ, $H(\hat{\Theta})$ is the negative Hessian matrix evaluated at $\hat{\Theta}$ and $|H(\hat{\Theta})|$ is its determinant, and N_p is the number of free parameters to be estimated which is equal to $M(5D+1)$ in our case. More details and discussions about the integrated likelihood can be found in [53–55] and references therein. In the following, we develop the required terms to determine the integrated likelihood of our model; namely the prior density and the determinant of the Hessian matrix.

Concerning the prior density $p(\Theta)$, in the absence of knowledge about the parameters (or complete ignorance), a widely applied reasonable assumption is to consider that the mixture parameters are independent:

$$p(\Theta) = p(p_1, \ldots, p_{M+1}) \prod_{j=1}^{M} \prod_{d=1}^{D} p(\rho_{jd}) p(\beta_{jd}) p(\alpha_{jd}) p(\beta_{\lambda|jd}) p(\alpha_{\lambda|jd}) \tag{17}$$

[3] Generally it is easier to understand simple models which are often the best solutions in accordance with Occam's Razor philosophy.

It is common to assume a symmetric Dirichlet distribution with parameters η as a prior for the mixing parameters, since they are defined on the simplex $(p_1, \dots, p_M : \sum_{j=1}^{M} p_j < 1)$. Generally η is set to 1 which gives the following prior $p(p_1, \dots, p_{M+1}) = M!$. We know also that each ρ_{jd} is defined in the compact support $[0, 1]$; thus a common widely used prior is the Beta distribution. Taking a symmetric Beta with parameters set to 1 gives us a uniform prior $p(\rho_{jd}) = \mathcal{U}_{[0,1]}$. For the scale parameters $p(\beta_{jd})$ and $\beta_{\lambda|jd}$, we consider $p(\beta_{jd}) = \frac{1}{\beta_{jd}}$ and $p(\beta_{\lambda|jd}) = \frac{1}{\beta_{\lambda|jd}}$ as priors, respectively. Moreover, we consider exponential priors with parameters set to 10^{-2}. For the shape parameters: $p(\alpha_{jd}) = 10^{-2}\exp(-10^{-2}\alpha_{jd})$ and $p(\alpha_{\lambda|jd}) = 10^{-2}\exp(-10^{-2}\alpha_{\lambda|jd})$. Notice that these specific choices of priors express actually our uncertainty about the model's parameters and were found convenient according to our experiments. By substituting all these priors into Eq. 17 we obtain:

$$p(\Theta) = M! \prod_{j=1}^{M} \prod_{d=1}^{D} \frac{10^{-4}\exp\big(- 10^{-2}(\alpha_{jd} + \alpha_{\lambda|jd})\big)}{\beta_{jd}\beta_{\lambda|jd}} \tag{18}$$

Concerning the determinant of the Hessian matrix it is common to assume in the case of finite mixture models that the components decouple [56], thus we may write $|H(\Theta)|$ as:

$$|H(\Theta)| = |H(p_1, \dots, p_{M+1})| \prod_{j=1}^{M} \left[|H(\theta_j)||H(\lambda_j)| \prod_{d=1}^{M} |H(\rho_{jd})| \right] \tag{19}$$

where $|H(p_1, \dots, p_{M+1})|$, $|H(\theta_j)|$, $|H(\lambda_j)|$ and $|H(\rho_{jd})|$ are the determinants of the Hessians with respect to the mixing parameters, θ_j, λ_j and (ρ_{jd}), respectively, and are given by (See Appendix 4):

$$|H(p_1, \dots, p_{M+1})| = \prod_{j=1}^{M} \sum_{i=1}^{N} \left(\frac{p(M+1|\vec{X}_i)}{p_{M+1}} - \frac{p(j|\vec{X}_i)}{p_j} \right)^2 \tag{20}$$

$$|H(\rho_{jd})| = \sum_{i=1}^{N} p(j|\vec{X}_i)^2 \left(\frac{f(1-\rho_{jd}, \theta_{jd}, \lambda_{jd})}{1-\rho_{jd}} - \frac{f(\rho_{jd}, \theta_{jd}, \lambda_{jd})}{\rho_{jd}} \right)^2 \tag{21}$$

$$\begin{aligned} |H(\theta_j)| \;&=\; \prod_{d=1}^{D} \Bigg[\Big(\sum_{i=1}^{N} p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd}) \big(\frac{\alpha_{jd}}{\beta_{jd}^2} - \frac{X_{id}}{\beta_{jd}^3}\big) \Big) \Psi'(\alpha_{jd}) \Big(\sum_{i=1}^{N} p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd}) \Big) \\ &- \; \Big(\frac{1}{\beta_{jd}} \sum_{i=1}^{N} p(j|\vec{X}_i) f(\rho_{jd}, \theta_{jd}, \lambda_{jd}) \Big)^2 \Bigg] \end{aligned} \tag{22}$$

$$
\begin{aligned}
|H(\lambda_j)| &= \prod_{d=1}^{D}\Bigg[\Bigg(\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})(\frac{\alpha_{\lambda|jd}}{\beta^2_{\lambda|jd}} - \frac{X_{id}}{\beta^3_{\lambda|jd}})\Bigg)\Psi'(\alpha_{\lambda|jd})\Bigg(\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})\Bigg) \\
&- \Bigg(\frac{1}{\beta_{\lambda|jd}}\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})\Bigg)^2\Bigg] \qquad (23)
\end{aligned}
$$

2.3.3 The Expectation Maximization (EM) Algorithm

Having all the estimation equations and the integrated likelihood expression in hand, our model learning will be performed under the standard two-phase paradigm employed by the expectation maximization (EM) framework as follows:

For each candidate value of M:

1. Set $\rho_{jd} \leftarrow 0.5, d = 1, \ldots, D, j = 1, \ldots, M$ and initialization of the rest of parameters [4].
2. Iterate the two following steps until convergence:
   (a) E-Step: Update $p(j|\vec{X}_n)$ using Eq. 13 and $f(\rho_{jd}, \theta_{jd}, \lambda_{jd})$ using Eq. 14
   (b) M-Step: Update the p_j, ρ_{jd}, β_{jd}, α_{jd}, $\beta_{\lambda|jd}$ and $\alpha_{\lambda|jd}$ using Eqs. 7, 8, 9, 10, 11 and 12, respectively.
3. Calculate the associated integrated likelihood using Eq. 16.
4. Select the optimal model that yields the largest integrated likelihood.

The previous algorithm is based on the EM approach and both E- and M-steps have a complexity of $O(NMD)$. According to our operational definition of outliers, they should have a uniform distribution, since they do not follow the pattern of the majority of the data. A common approach, to define this uniform distribution, is to suppose that the data follow a single component model averaged over all the observation [37]. Thus, in our case, we choose the following [5] $U(\vec{X}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{d=1}^{D}\big(\hat{\rho}_d p(X_{id}|\hat{\theta}_d) + (1 -$

[4] The initialization is based on the K-Means algorithm and the method of moments by considering that $M+1$ clusters are present in the data.

[5] Of course other choices are possible, but in our case this specific choice was found appropriate according to our experimental results.

$\hat{\rho}_d)p(X_{id}|\hat{\lambda}_d))$, where the parameters $\hat{\rho}$, $\hat{\theta}_d$ and $\hat{\lambda}_d$ are estimated using ML technique, which takes into account the fact that outliers should be sparsely distributed. It is noteworthy that the previous algorithm allows to first detect outlying data. Then, the remaining ones (i.e. inliers) are used to identify the optimal clustering structure in terms of number of clusters, relevant features and optimal parameters.

CHAPTER 3

Experimental Results

3.1 Introduction

In this chapter, experiments are carried out to evaluate the usefulness of our model. The experiments are performed on both synthetic and real data. Also, we investigate our model on a challenging problem namely; objects shape clustering. For the purpose of comparison, we have implemented our model with Gaussian distributions (see appendix 5 for the derivation equations of the maximum likelihood estimation), and using localized and globalized feature selection methods, as well. In all our experiments, we investigate the advantages of performing simultaneous feature selection and outliers detection. Moreover, all results are averaged over 10 runs of the algorithm. The experimental framework is defined by the following:

- ❑ First, we compare the performance of these finite mixtures without taking into consideration the relevancy of features nor the presence of outlier data.
- ❑ Then, we consider the relevancy of features for both mixtures. Hence, we got better partitioning and model selection for the tested data sets.
- ❑ Finally, we consider the relevancy of features, and the presence of outlier data.

3.2 Synthetic Data

The first application evaluates the performance of the proposed model using three 2D synthetic data sets generated from 2- 3- and 4-components bivariate Gamma mixture models. The parameters used to generate these data sets are given in table 3.1.

Table 3.1: Parameters used to generate the synthetic data sets (n_j represents the number of elements in cluster *j*).

Data set	*j*	α_{j1}	β_{j1}	α_{j2}	β_{j2}	n_j
Data set 1	1	18	4	13	6	100
	2	6	3	5	4	100
Data set 2	1	18	4	13	6	100
	2	6	3	5	4	100
	3	38	7	39	6	100
Data set 3	1	18	4	13	6	100
	2	6	3	5	4	100
	3	38	7	39	6	100
	4	49	9	46	8	100

Experiment 1

The first experiment is conducted by appending eight “noisy” features to the generated data sets which increases the dimensionality of the data to 10. The goal of this experiment is to evaluate the ability of the algorithm in selecting features when no outliers are present. Table 3.2 contains the classification results for the three synthetic data sets using the localized and globalized feature selection methods. The results in this table show that the localized feature selection method improves significantly the classification accuracy for the three generated data sets compared to the globalized one. Figures 3.1 and 3.2 show the localized and the globalized saliency of features using our approach for the three generated data sets. According to these figures, we can see that the localized feature selection method gives higher weights to the first two features

Table 3.2: Classification results for the three synthetic data sets using localized and globalized feature selection methods.

Data Set	j	Localized feature selection: n_j	Accuracy	Globalized feature selection: n_j	Accuracy
Data set 1	1	86	93.00%	79	89.50%
	2	114		121	
Data set 2	1	111	96.30%	100	92.00%
	2	89		124	
	3	100		76	
Data set 3	1	99	95.25%	84	92.50%
	2	112		100	
	3	82		86	
	4	107		130	

(i.e. the relevant features). Figure 3.3 shows the number of clusters selected for each data set using the integrated likelihood criterion using localized and globalized feature selection methods. According to these figures, it is clear that our algorithm is able to select the correct number of clusters in both cases.

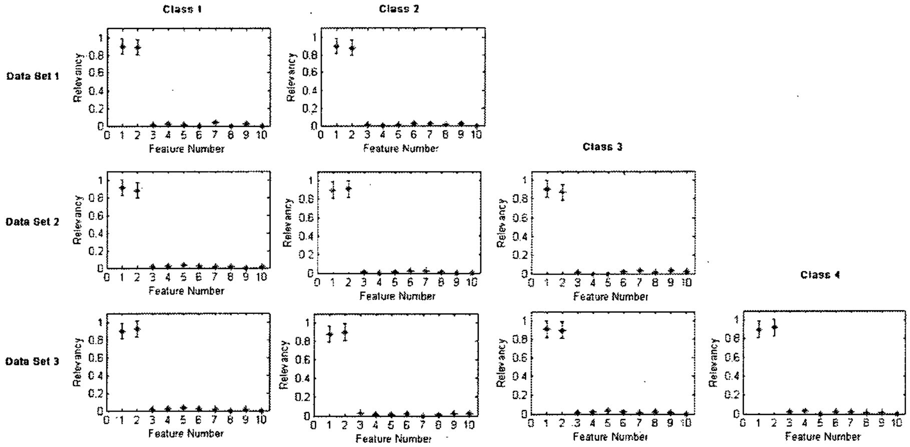

Figure 3.1: Localized features saliency for the synthetic data sets.

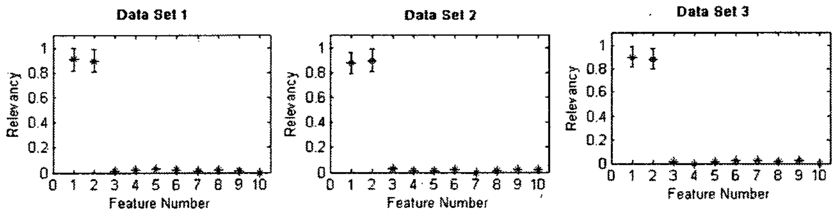

Figure 3.2: Globalized features saliency for the synthetic data sets.

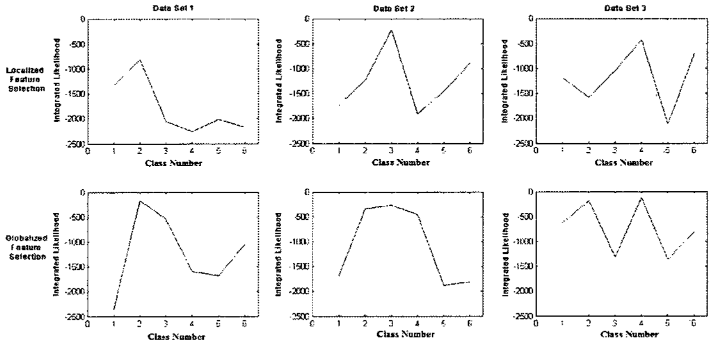

Figure 3.3: Integrated likelihood as a function of the number of clusters for the three synthetic data sets when relevancy of features is considered. Row 1: using localized feature selection, Row 2: using globalized feature selection.

Experiment 2

The second experiment is conducted to evaluate the presence of outlying data in the model selection process; through the introduction of 5, 10 and 15 outlying 10-dimensional vectors into the first, second, and third generated data sets, respectively. Table 3.3 contains the classification results for the three synthetic data sets using the localized and globalized feature selection methods. It is clear that the presence of outliers compromises the feature selection process by affecting high weights to some irrelevant features and by decreasing the relevancy weights of the two first features as shown in figures 3.4 and 3.5. Figure 3.6 shows the number of clusters found using the integrated likelihood criterion after performing feature selection by considering that all the vectors are actually inliers (i.e. without performing outliers detection). According to this figure, the algorithm is unable to determine the exact number of clusters only for the third data set when performing globalized feature selection.

Table 3.3: Classification results for the three synthetic data sets using localized and globalized feature selection methods, without outliers detection.

Data Set	j	Localized feature selection: n_j	Accuracy	Globalized feature selection: n_j	Accuracy
Data set 1	1	129	85.85%	66	80.97%
	2	76		139	
Data set 2	1	107	92.90%	62	84.51%
	2	88		116	
	3	115		132	
Data set 3	1	123	93.97%	121	91.32%
	2	100		80	
	3	108		99	
	4	82		115	

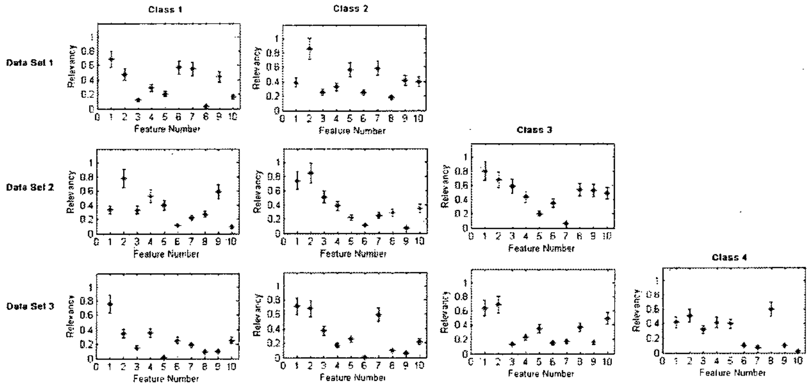

Figure 3.4: Localized saliency of features for the synthetic data sets without outliers detection.

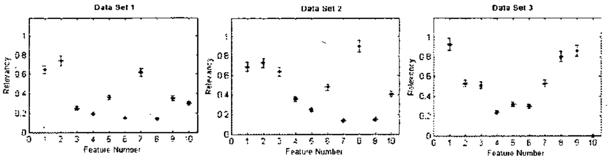

Figure 3.5: Globalized saliency of features for the synthetic data sets without outliers detection.

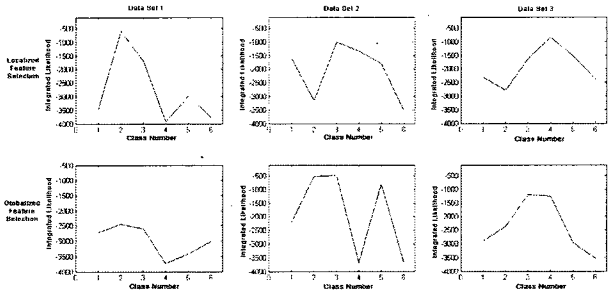

Figure 3.6: Integrated likelihood as a function of the number of clusters for the three synthetic data sets using feature selection and without outliers detection. Row 1: using localized feature selection, Row 2: using globalized feature selection.

Experiment 3

When performing outliers detection, our algorithm is able to detect all the outlying data for the three generated data sets and to select the relevant features necessary for model selection process as shown in table 3.4, and in figures 3.7 and 3.8. Also, figure 3.9 shows that our algorithm is able to detect the exact number of clusters in all cases using the integrated likelihood criterion.

Table 3.4: Classification results of the three synthetic data sets using features selection and outliers detection.

Data Set	j	Localized feature selection: n_j	Accuracy	Globalized feature selection: n_j	Accuracy
Data set 1	1	105	97.56%	80	90.24%
	2	95		120	
	Outlier	5		5	
Data set 2	1	100	97.41%	87	92.90%
	2	92		122	
	3	108		91	
	Outlier	10		10	
Data set 3	1	94	95.90%	82	92.04%
	2	117		96	
	3	91		133	
	4	100		89	
	Outlier	13		15	

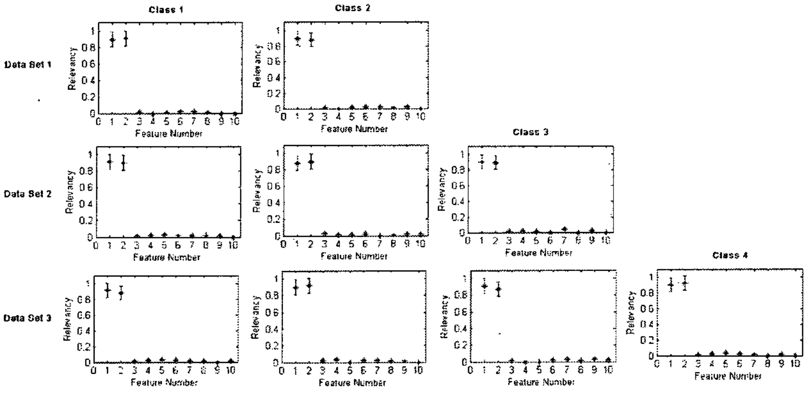

Figure 3.7: Localized saliency of features for the synthetic data sets with outliers detection.

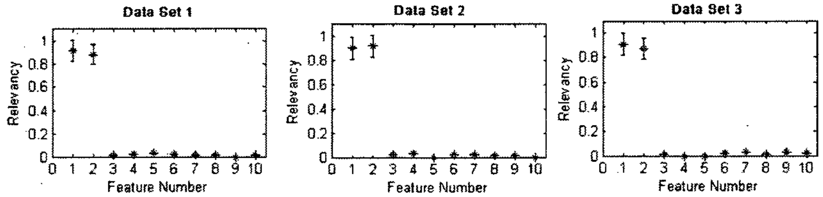

Figure 3.8: Globalized saliency of features for the synthetic data sets with outliers detection.

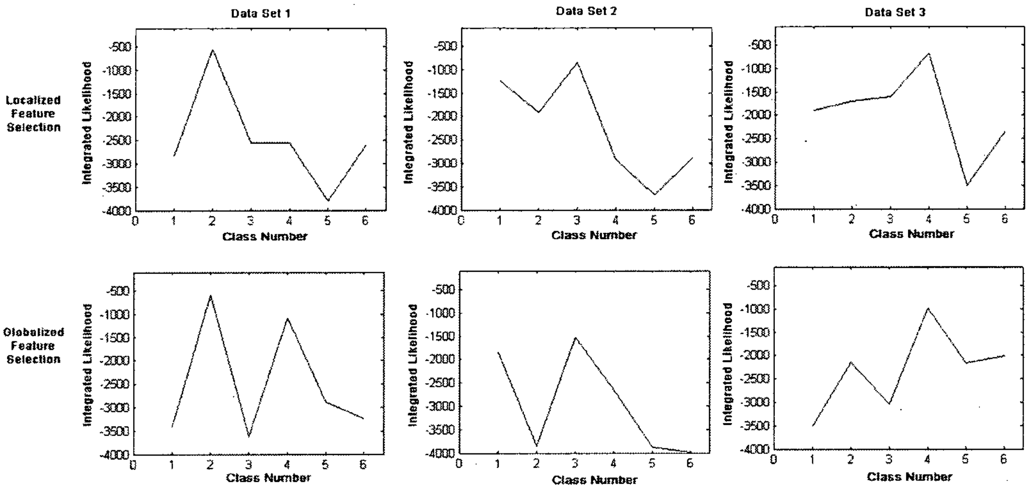

Figure 3.9: Integrated likelihood as a function of the number of clusters for the three synthetic data sets using features selection and outliers detection. Row 1: using localized feature selection, Row 2: using globalized feature selection.

3.3 Real Data

The second application concerns the classification of handwritten numerals using a data set composed of 10 classes ("0"-"9") obtained from [57]. Each class contains 200 patterns where each pattern is represented by a 47-dimensional positive vector describing extracted Zernike moments magnitudes features. We add to this data set 20 47-dimensional vectors, representing Zernike moments magnitudes features, extracted from 20 textural images from the MIT Vistex gray level texture database [1] (see figure 3.10). These 20 added vectors are considered as our outlier data. The main goal of this application is to compare the performance of Gamma and Gaussian mixture models when dealing with features defined over the positive reals.

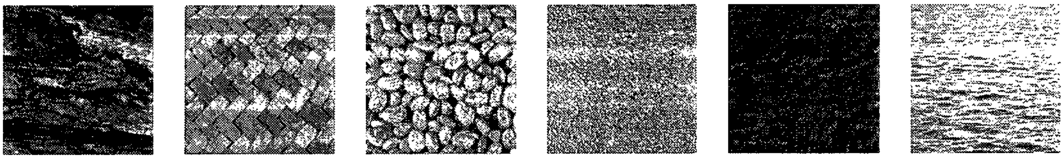

Figure 3.10: Sample images from the Vistex data set.

Experiment 1

The first experiment is conducted to compare mixtures of Gamma and Gaussian distributions without taking into consideration the relevancy of features nor the presence of outlier data. Table 3.5 shows the confusion matrix when applying mixture of Gamma distributions; the number of miss-classified shapes is 281 which represents an accuracy of 86.08%. On the other hand, table 3.6 shows the confusion matrix using mixture of Gaussian distributions; the number of miss-classified shapes is 389 which represents an accuracy of 80.74%. In both cases, we are unable to detect the outlier data as a different class. Indeed, all outliers are affected to the 10 selected classes. Figure 3.11 shows the number of clusters found using the integrated likelihood criterion.

Experiment 2

The second experiment is conducted to compare mixtures of Gamma and Gaussian distributions through considering the relevancy of features using localized feature selection method. Table 3.7 shows the confusion matrix when applying mixture of Gamma distributions; the number of miss-classified shapes is 195 which represents an accuracy of 90.3%. On the other hand, table 3.8 shows the confusion matrix when applying mixture of Gaussian distributions; the number of miss-classified shapes is 311 which represents an

[1] http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html

Table 3.5: Confusion matrix for the handwritten numerals data set using mixture of Gamma distributions, without performing features selection nor outliers detection.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9
0	183	0	0	0	0	0	13	0	0	4
1	0	198	0	0	0	0	0	2	0	0
2	0	0	143	8	0	49	0	0	0	0
3	0	0	0	175	0	0	0	0	23	2
4	0	9	0	0	132	59	0	0	0	0
5	0	0	12	0	0	186	2	0	0	0
6	0	0	0	0	0	0	191	0	0	9
7	0	3	0	0	0	0	0	158	0	39
8	0	0	0	21	0	0	0	0	179	0
9	6	0	0	0	0	0	0	0	0	194
Outlier Class	1	14	0	0	3	0	0	2	0	0

Table 3.6: Confusion matrix for the handwritten numerals data set using mixture of Gaussian distributions, without performing features selection nor outliers detection.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9
0	152	0	0	5	0	0	9	0	34	0
1	0	187	0	0	8	0	0	5	0	0
2	0	42	147	0	0	0	0	0	11	0
3	0	0	0	163	0	0	0	0	31	6
4	3	0	0	0	197	0	0	0	0	0
5	0	0	5	1	6	125	63	0	0	0
6	0	0	0	0	3	46	151	0	0	0
7	8	18	0	0	0	0	0	174	0	0
8	6	0	0	12	0	0	0	0	182	0
9	0	0	0	39	0	0	0	0	8	153
Outlier Class	0	0	0	8	0	0	12	0	0	0

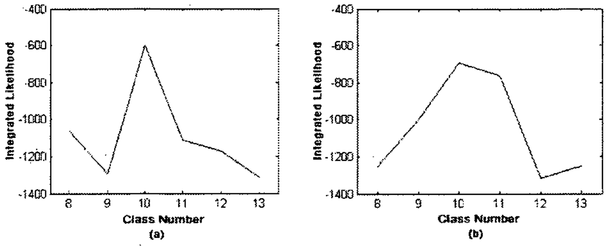

Figure 3.11: Integrated likelihood as a function of the number of clusters for the handwritten numerals data set, without performing features selection nor outliers detection. (a) Using Mixture of Gamma Distributions. (b) Using Mixture of Gaussian Distributions.

accuracy of 84.6%, which show that the classification accuracy has increased using both mixtures with the effect of outliers presence. Figure 3.12 and 3.13 shows the localized saliency of features for the classified shapes for both mixtures. Figure 3.14 shows the number of clusters found using the integrated likelihood criterion.

Table 3.7: Confusion matrix for the handwritten numerals data set using mixture of Gamma distributions, and localized feature selection method.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9
0	193	0	0	0	0	0	0	0	7	0
1	0	181	0	0	9	0	0	10	0	0
2	0	0	166	26	0	8	0	0	0	0
3	0	0	0	184	0	0	0	0	16	0
4	0	0	6	6	188	0	0	0	0	0
5	0	0	0	0	0	181	15	0	4	0
6	0	0	0	0	0	11	189	0	0	0
7	0	12	0	0	0	0	0	184	0	4
8	29	0	0	0	0	8	0	0	163	0
9	0	0	0	4	0	0	0	0	0	196
Outlier Class	0	0	4	0	3	0	9	0	4	0

Table 3.8: Confusion matrix for the handwritten numerals data set using mixture of Gaussian distributions, and localized feature selection method.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9
0	199	0	0	0	0	0	0	1	0	0
1	0	122	0	0	0	0	24	0	38	16
2	0	6	194	0	0	0	0	0	0	0
3	0	0	0	119	0	0	0	0	55	26
4	0	11	0	0	184	0	0	5	0	0
5	0	0	2	0	0	167	28	0	3	0
6	0	0	0	0	0	19	177	0	4	0
7	0	6	0	0	0	0	0	181	0	13
8	0	0	13	0	0	0	0	0	187	0
9	0	0	13	0	0	0	0	0	8	179
Outlier Class	15	0	2	0	0	3	0	0	0	0

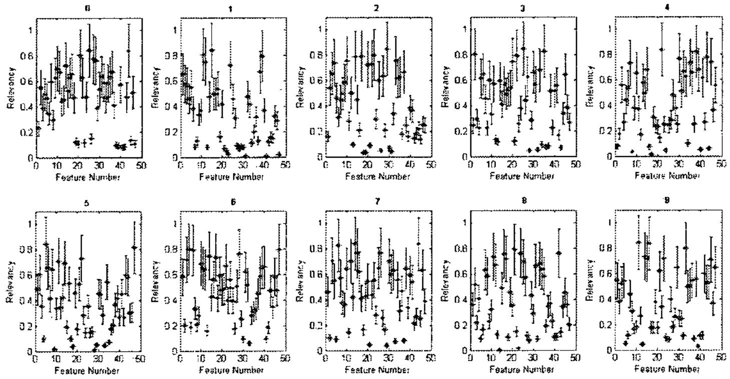

Figure 3.12: Localized saliency of features for the handwritten data set using mixture of Gamma distributions.

Experiment 3

The third experiment is conducted to compare mixtures of Gamma and Gaussian distributions using globalized feature selection method. Table 3.9 shows the confusion matrix when applying mixture of Gamma

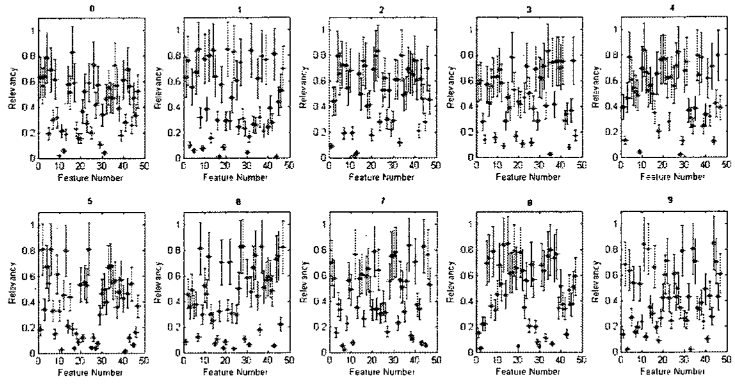

Figure 3.13: Localized saliency of features for the handwritten data set using mixture of Gaussian distributions.

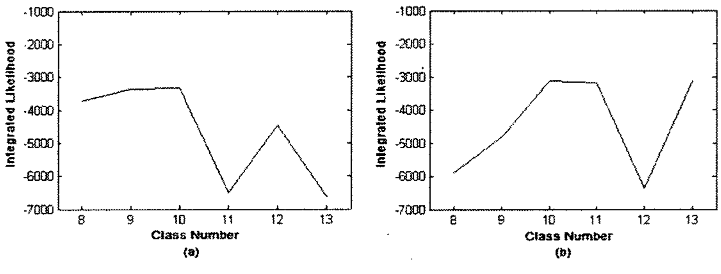

Figure 3.14: Integrated likelihood as a function of the number of clusters for the handwritten numerals data set using localized feature selection method. (a) Using Mixture of Gamma Distributions. (b) Using Mixture of Gaussian Distributions.

distributions; the number of miss-classified shapes is 213 which represents an accuracy of 89.4%. On the other hand, table 3.10 shows the confusion matrix when applying mixture of Gaussian distributions; the number of miss-classified shapes is 337 which represents an accuracy of 83.3%, which show that the classification accuracy is lower using the globalized feature selection method comparably to localized feature selection in the previous experiment. Figure 3.15 shows the globalized saliency of features for the classified shapes for both mixtures, and figure 3.16 shows the number of clusters found using the integrated likelihood criterion.

Table 3.9: Confusion matrix for the handwritten numerals data set using mixture of Gamma distributions, and globalized feature selection method.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9
0	178	0	0	0	0	0	0	0	22	0
1	0	192	8	0	0	0	0	0	0	0
2	0	5	158	37	0	0	0	0	0	0
3	0	0	9	184	0	0	0	0	7	0
4	0	26	0	0	166	0	8	0	0	0
5	0	0	0	0	0	184	16	0	0	0
6	0	0	0	0	0	11	189	0	0	0
7	0	6	0	0	0	0	0	175	0	19
8	0	0	7	0	0	0	0	0	193	0
9	0	0	12	0	0	0	0	0	0	188
Outlier Class	4	0	7	0	6	1	2	0	0	0

Table 3.10: Confusion matrix for the handwritten numerals data set using mixture of Gaussian distributions, and globalized feature selection method.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9
0	118	0	0	0	0	0	0	0	82	0
1	0	167	33	0	0	0	0	0	0	0
2	0	9	191	0	0	0	0	0	0	0
3	0	0	0	176	0	0	0	0	24	0
4	0	19	0	0	181	0	0	0	0	0
5	0	0	0	0	0	119	81	0	0	0
6	0	0	0	0	0	13	187	0	0	0
7	0	0	0	0	0	0	0	171	0	29
8	6	0	0	0	0	0	0	0	194	0
9	0	0	0	21	0	0	0	0	0	179
Outlier Class	2	8	0	0	0	3	0	1	0	6

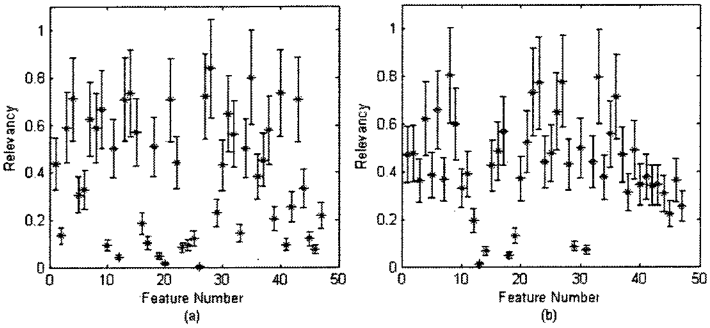

Figure 3.15: Globalized saliency of features for the handwritten numerals data set. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

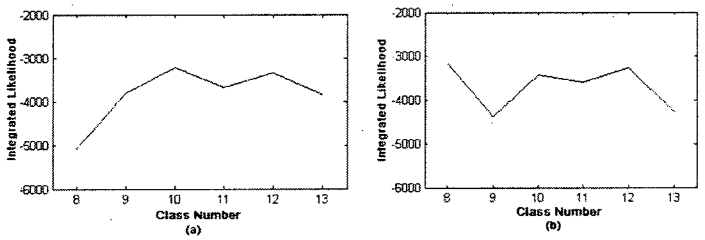

Figure 3.16: Integrated likelihood as a function of the number of clusters for the handwritten numerals data set using globalized feature selection method. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

Experiment 4

The fourth experiment is conducted using mixtures of Gamma and Gaussian distributions using localized features selection method, and with outliers detection. Table 3.11 shows the confusion matrix for the mixture of Gamma distributions; the number of miss-classified shapes is 104 which represents an accuracy of 94.85%. On the other hand, table 3.12 shows the confusion matrix for mixture of Gaussian distributions; the number of miss-classified shapes is 241 which represents an accuracy of 88.06%. It is noteworthy that using our approach applied with the Gamma mixture we are able to detect all the outliers which is not the case with the Gaussian mixture (2 outliers are assigned to class 9). Figures 3.17 and 3.18 show the localized saliency of features for both mixtures, and figure 3.19 shows the number of clusters found using the integrated likelihood criterion.

Table 3.11: Confusion matrix for the handwritten numerals data set using mixture of Gamma distributions, localized features selection method, and with outliers detection.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9	Outlier Class
0	186	0	0	0	0	0	0	0	0	14	0
1	0	192	8	0	0	0	0	0	0	0	0
2	0	0	179	0	3	18	0	0	0	0	0
3	0	0	0	200	0	0	0	0	0	0	0
4	0	2	0	0	198	0	0	0	0	0	0
5	0	0	0	0	0	200	0	0	0	0	0
6	0	0	0	0	0	0	182	0	9	9	0
7	0	5	11	0	0	0	0	184	0	0	0
8	0	0	0	6	0	0	0	0	194	0	0
9	0	0	0	0	0	0	0	0	19	181	0
Outlier Class	0	0	0	0	0	0	0	0	0	0	20

Experiment 5

The fifth experiment is conducted using mixtures of Gamma and Gaussian distributions using globalized features selection method, and with outliers detection. Table 3.13 shows the confusion matrix for the mixture of Gamma distributions; the number of miss-classified shapes is 124 which represents an accuracy of 93.8%. On the other hand, table 3.14 shows the confusion matrix for mixture of Gaussian distributions; the number of miss-classified shapes is 257 which represents an accuracy of 87.2%. It is noteworthy that using our approach applied with the Gamma mixture we are able to detect all the outliers which is not the case with the Gaussian mixture (1 outlier is assigned to class 7). Figures 3.20 and 3.21 show the globalized saliency of features for both mixtures, and the number of clusters found using the integrated likelihood criterion.

Table 3.12: Confusion matrix for the handwritten numerals data set using mixture of Gaussian distributions, localized features selection method, and with outliers detection.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9	Outlier Class
0	161	0	0	0	0	0	0	0	0	39	0
1	0	189	11	0	0	0	0	0	0	0	0
2	0	4	129	0	0	67	0	0	0	0	0
3	0	0	0	172	0	0	0	0	0	28	0
4	0	15	0	0	185	0	0	0	0	0	0
5	0	0	0	0	0	162	19	0	0	19	0
6	0	0	0	0	0	6	194	0	0	0	0
7	0	2	16	0	0	0	0	182	0	0	0
8	0	0	0	13	0	0	0	0	187	0	0
9	0	0	0	0	0	0	0	0	0	200	0
Outlier Class	0	0	0	0	0	0	0	0	0	2	18

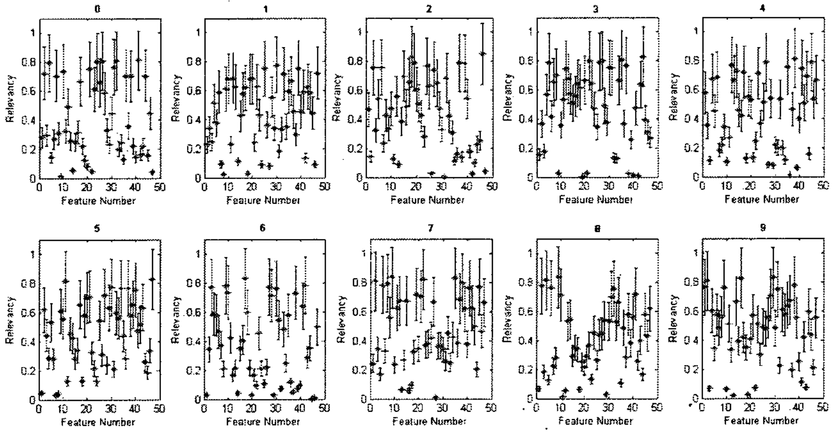

Figure 3.17: Localized saliency of features for the handwritten numerals data set using mixture of Gamma distributions, and with outliers detection.

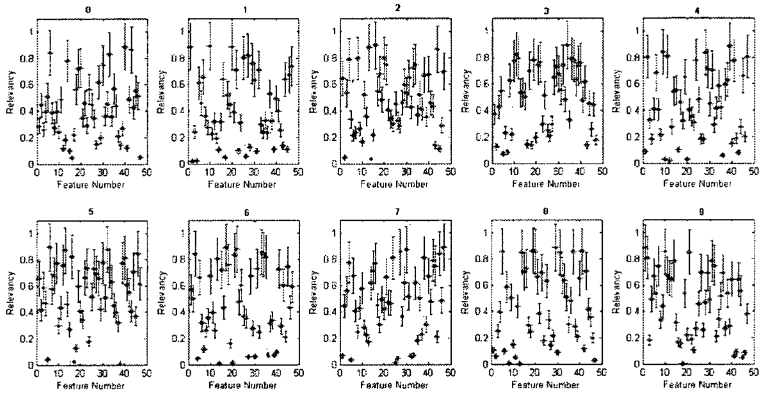

Figure 3.18: Localized saliency of features for the handwritten numerals data set using mixture of Gaussian distributions, and with outliers detection.

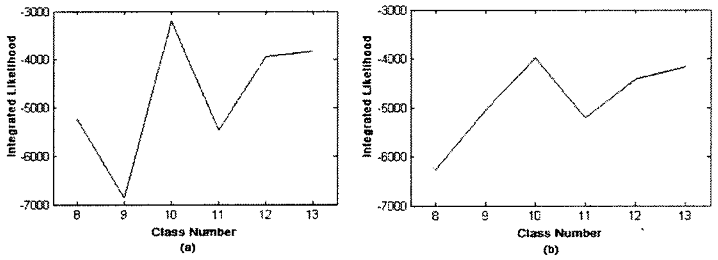

Figure 3.19: Integrated likelihood as a function of the number of clusters for the handwritten numerals data using localized feature selection method, and outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

Table 3.13: Confusion matrix for the handwritten numerals data set using mixture of Gamma distributions, globalized features selection method, and outliers detection.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9	Outlier Class
0	200	0	0	0	0	0	0	0	0	0	0
1	0	192	8	0	0	0	0	0	0	0	0
2	0	0	174	26	0	0	0	0	0	0	0
3	0	0	0	192	0	0	0	0	8	0	0
4	0	17	0	0	183	0	0	0	0	0	0
5	0	0	0	0	0	182	18	0	0	0	0
6	0	0	0	0	0	10	190	0	0	0	0
7	0	0	0	0	0	0	0	179	0	21	0
8	0	0	0	14	0	0	0	0	186	0	0
9	0	0	0	2	0	0	0	0	0	198	0
Outlier Class	0	0	0	0	0	0	0	0	0	0	20

Table 3.14: Confusion matrix for the handwritten numerals data set using mixture of Gaussian distributions, globalized features selection method, and outliers detection.

Classified as ⇒	0	1	2	3	4	5	6	7	8	9	Outlier Class
0	134	0	0	0	0	0	0	0	66	0	0
1	0	189	11	0	0	0	0	0	0	0	0
2	0	0	200	0	0	0	0	0	0	0	0
3	0	0	0	157	0	0	0	0	43	0	0
4	0	0	0	0	185	0	0	6	0	9	0
5	0	0	0	0	0	152	48	0	0	0	0
6	0	0	0	0	0	6	194	0	0	0	0
7	0	0	0	0	0	0	0	182	0	18	0
8	13	0	0	0	0	0	0	0	187	0	0
9	0	0	0	36	0	0	0	0	0	164	0
Outlier Class	0	0	0	0	0	0	0	1	0	0	19

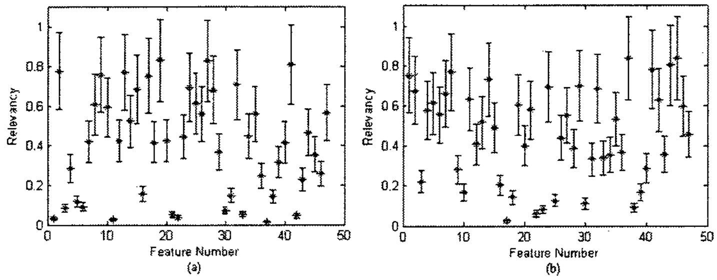

Figure 3.20: Globalized saliency of features for the handwritten numerals data set using mixtures of Gamma and Gaussian distributions with outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

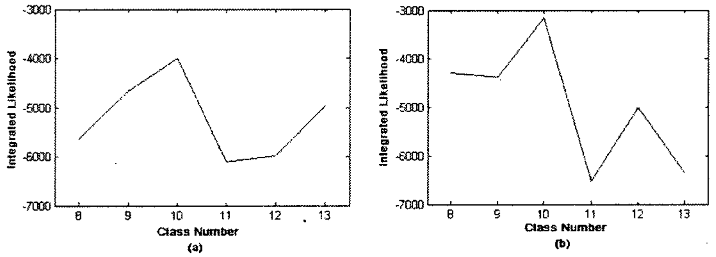

Figure 3.21: Integrated likelihood as a function of the number of clusters for the handwritten numerals using globalized feature selection method, and outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

3.4 2D Objects Shape Clustering

Shape modeling and representation is an important step in several applications such as content-based image retrieval, indexing, and image segmentation [58]. Zernike Moments Magnitudes (ZMMs), which have been the subject of extensive theoretical and experimental research in the past [59–61], are known to be very effective to model objects shapes. Indeed, they allow invariant (i.e. regardless the position, size and orientation) recognition of objects [62]. Moreover, the Gamma distribution was found suitable to model ZMMs [63]. Thus, we present in the following results of applying our model on the clustering of shape images represented by ZMMs. The data set used is a subset of the MPEG-7 CE Shape-1 Part-B data set that consists of seven classes, where each class includes 20 shape samples [2]. Figure 3.22 shows examples of images from this data set. We also add to this data set 5 textural images from the MIT Vistex gray level database as outliers. After normalizing all the images [64, 65], the vector of characteristics (ZMMs) is computed for each image (shape) using the method proposed in [63]. Thus, each image is represented by a 36-dimensional vector.

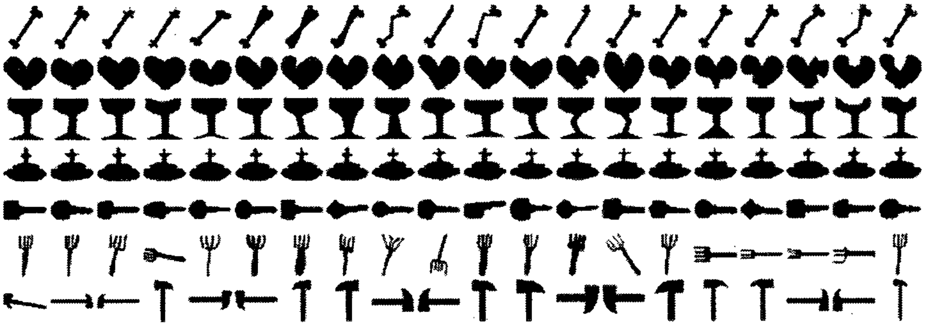

Figure 3.22: Samples of the MPEG-7 CE Shape-1 Part-B data set.

Experiment 1

The first experiment is conducted using mixtures of Gamma and Gaussian distributions without performing nor feature selection neither outliers detection. Table 3.15 shows the confusion matrix for mixture of Gamma distributions; the number of miss-classified shapes is 29 which represents an accuracy of 80.0%. On the other hand, table 3.16 shows the confusion matrix for mixture of Gaussian distributions; the number of miss-classified shapes is 50 which represents an accuracy of 65.5%. For both mixtures, the number of clusters selected by the integrated likelihood criterion is 7 as shown in figure 3.23.

[2] http://www.cis.temple.edu/ latecki/TestData/mpeg7shapeB.tar.gz

Table 3.15: Confusion matrix for MPEG-7 data set using mixture of Gamma distributions without features selection nor outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer
Bone	17	0	3	0	0	0	0
Heart	0	15	0	0	0	1	4
Glass	0	0	19	1	0	0	0
Fountain	0	0	9	11	0	0	0
Key	0	0	0	0	18	0	2
Fork	0	0	0	0	0	20	0
Hummer	4	0	0	0	0	0	16
Outlier Class	0	0	5	0	0	0	0

Table 3.16: Confusion matrix for MPEG-7 data set using mixture of Gaussian distributions without features selection nor outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer
Bone	19	1	0	0	0	0	0
Heart	4	16	0	0	0	0	0
Glass	0	6	10	0	0	2	2
Fountain	0	0	0	14	0	3	3
Key	5	0	0	0	15	0	0
Fork	1	6	0	0	0	13	0
Hummer	12	0	0	0	0	0	8
Outlier Class	3	0	0	1	1	0	0

Experiment 2

The second experiment is conducted using mixtures of Gamma and Gaussian distributions by performing localized feature selection, and without outliers detection. Tables 3.17 and 3.18 show the confusion matrices for both mixtures in this case. The number of miss-classified shapes using mixture of Gamma is 22 which represents an accuracy of 84.82%. On the other hand, the number of miss-classified shapes using mixture of

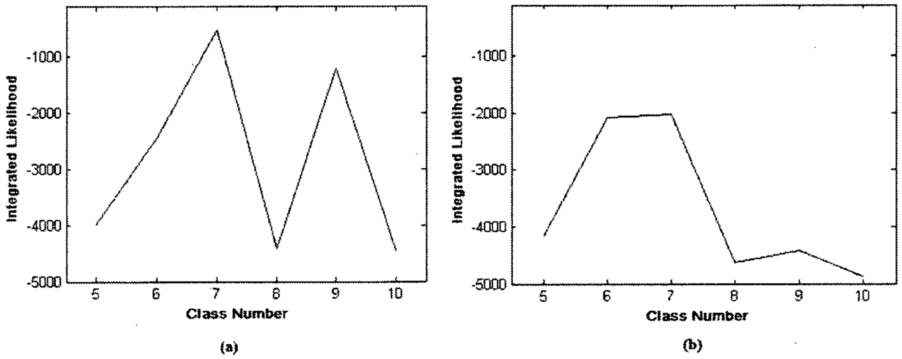

Figure 3.23: Integrated likelihood as a function of the number of clusters for the MPEG-7 data set without features selection nor outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

Gaussian is 40 which represents an accuracy of 72.41%. Figures 3.24 and 3.25 show the localized saliency of features of the classified shapes using both mixtures. Comparing the first and second experiments, we can deduce that the feature selection process improves significantly the clustering of shapes accuracy. However, as 7 is considered as the optimal number of clusters for both mixtures as shown in figure 3.26. The outliers are mixed with the inliers and affected to these clusters.

Experiment 3

The third experiment is conducted using mixtures of Gamma and Gaussian distributions by performing globalized feature selection, and without outliers detection. Tables 3.19 and 3.20 show the confusion matrices for both mixtures in this case. The number of miss-classified shapes using mixture of Gamma is 33 which represents an accuracy of 77.2%. On the other hand, the number of miss-classified shapes using mixture of Gaussian is 46 which represents an accuracy of 68.2%. Figures 3.27 and 3.28 show the globalized saliency of features of the classified shapes using both mixtures, and number of clusters found using the integrated likelihood criterion. Comparing the second and the third experiments; we can deduce that the localized features selection process improves the clustering of shapes accuracy compared to the globalized one.

Table 3.17: Confusion matrix for MPEG-7 data set using mixture of Gamma distributions, localized feature selection method, and without outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer
Bone	18	2	0	0	0	0	0
Heart	0	17	0	0	0	0	3
Glass	0	0	19	1	0	0	0
Fountain	0	2	0	18	0	0	0
Key	0	0	2	0	16	0	2
Fork	0	0	0	0	5	15	0
Hummer	0	0	0	0	0	0	20
Outlier Class	0	0	0	0	3	0	2

Table 3.18: Confusion matrix for MPEG-7 data set using mixture of Gaussian distributions, localized feature selection method, and without outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer
Bone	14	0	0	1	0	0	5
Heart	0	19	0	0	1	0	0
Glass	0	0	16	0	0	0	4
Fountain	0	8	0	12	0	0	0
Key	0	0	0	3	17	0	0
Fork	0	0	0	0	2	18	0
Hummer	2	0	0	0	5	4	9
Outlier Class	0	0	0	0	4	0	1

Experiment 4

The fourth experiment is conducted by taking into consideration both the relevancy of features and the presence of outlier data using localized feature selection method. Both mixtures are able to detect the outlier data (in the case of the Gaussian, however, one outlier was considered as an inlier); hence, the classification accuracy has increased. Tables 3.21 and 3.22 show the confusion matrices for both mixtures. The number

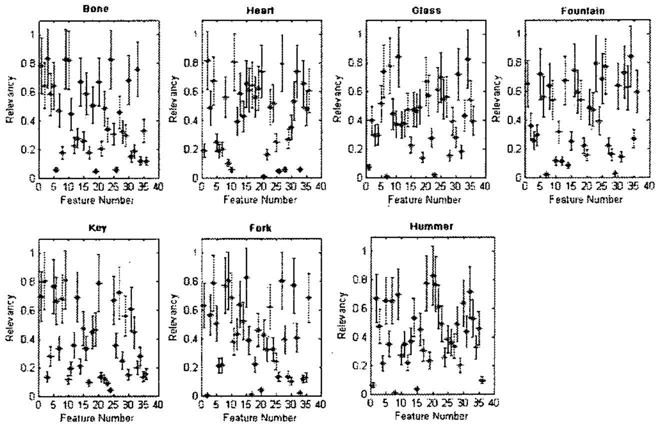

Figure 3.24: Localized saliency of features using mixture of Gamma distributions, and without outliers detection.

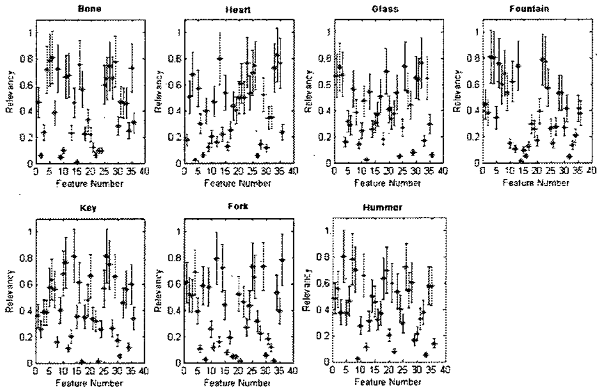

Figure 3.25: Localized saliency of features using mixture of Gaussian distributions, and without outliers detection.

of miss-classified shapes are 9 and 21 which represent accuracies of 93.7% and 85.5% when using Gamma and Gaussian mixtures, respectively. Figures 3.29 and 3.30 show the localized saliency of features for both mixtures. Figure 3.31 shows the number of clusters found using the integrated likelihood criterion.

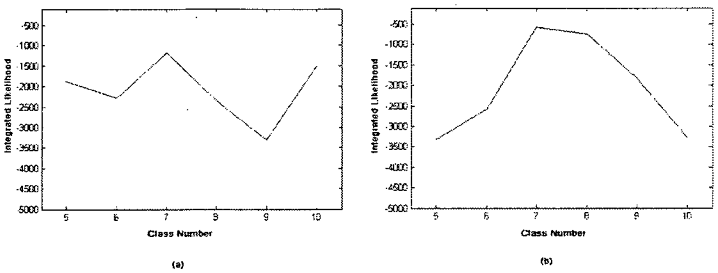

Figure 3.26: Integrated likelihood as a function of the number of clusters for the MPEG-7 data set using localized feature selection method, and without outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

Table 3.19: Confusion matrix for MPEG-7 data set using mixture of Gamma distributions, globalized features selection method, and without outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer
Bone	19	1	0	0	0	0	0
Heart	0	16	0	0	1	3	0
Glass	3	3	14	0	0	0	0
Fountain	0	2	0	18	0	0	0
Key	0	0	0	0	15	0	5
Fork	0	0	0	0	3	17	0
Hummer	0	0	0	0	0	7	13
Outlier Class	0	2	2	0	1	0	0

Experiment 5

The fifth experiment is conducted by taking into consideration both the relevancy of features and the presence of outlier data using globalized feature selection method. Both mixtures are able to detect the outlier data (in the case of the Gaussian, however, two outliers were considered as an inliers); hence, the classification accuracy has increased. Tables 3.23 and 3.24 show the confusion matrices for both mixtures. The numbers of miss-classified shapes is 9 and 23 which represent accuracies of 93.7% and 83.4% percent when

Table 3.20: Confusion matrix for MPEG-7 data set using mixture of Gaussian distributions, globalized features selection method, and without outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer
Bone	5	0	0	0	0	12	3
Heart	0	18	2	0	0	0	0
Glass	0	0	18	2	0	0	0
Fountain	0	0	0	9	0	0	11
Key	0	0	0	1	19	0	0
Fork	0	0	0	0	0	14	6
Hummer	0	1	0	0	3	0	16
Outlier Class	0	0	0	0	4	0	1

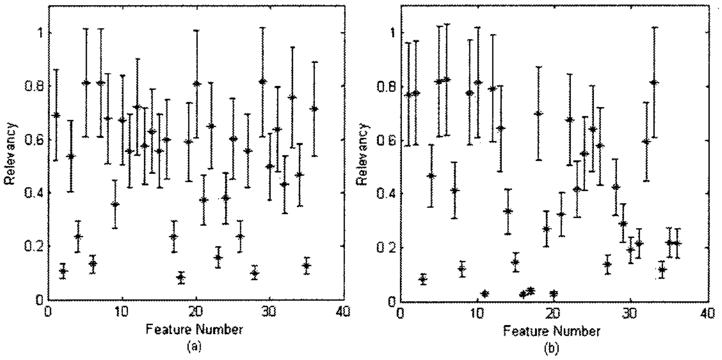

Figure 3.27: Globalized saliency of features for the MPEG-7 data set using globalized features selection method, and without outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

using Gamma and Gaussian mixtures, respectively. Figures 3.32 and 3.33 show the globalized saliency of features for both mixtures, and the number of clusters found using the integrated likelihood criterion.

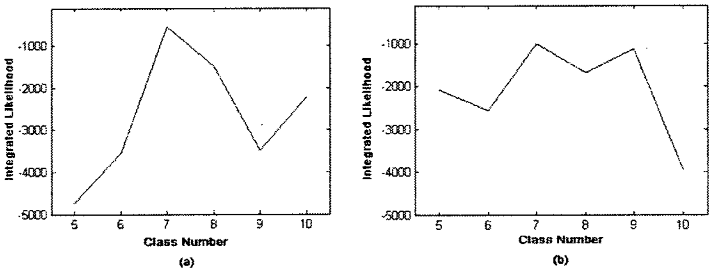

Figure 3.28: Integrated likelihood as a function of the number of clusters for the MPEG-7 data set using globalized features selection method, and without outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

Table 3.21: Confusion matrix for MPEG-7 data set using mixture of Gamma distributions, localized features selection method, and outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer	Outlier Class
Bone	19	0	0	0	0	0	1	0
Heart	0	20	0	0	0	0	0	0
Glass	0	0	18	0	2	0	0	0
Fountain	0	1	0	19	0	0	0	0
Key	0	0	0	0	19	0	1	0
Fork	0	0	0	0	0	18	2	0
Hummer	0	0	0	0	0	2	18	0
Outlier Class	0	0	0	0	0	0	0	5

Table 3.22: Confusion matrix for MPEG-7 data set using mixture of Gaussian distributions, localized features selection method, and outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer	Outlier Class
Bone	13	0	0	0	0	0	7	0
Heart	0	18	2	0	0	0	0	0
Glass	0	0	19	0	1	0	0	0
Fountain	0	0	3	17	0	0	0	0
Key	0	0	0	0	18	0	2	0
Fork	0	4	0	0	0	16	0	0
Hummer	0	0	0	0	0	1	19	0
Outlier Class	0	0	0	0	0	1	0	4

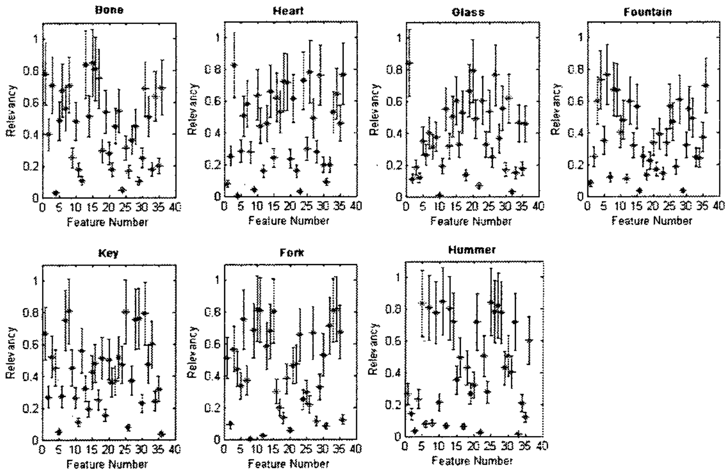

Figure 3.29: Localized saliency of features using mixture of Gamma distributions, and outliers detection.

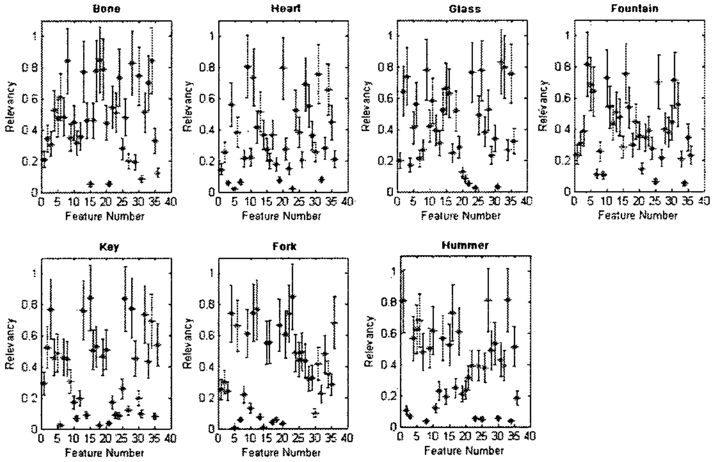

Figure 3.30: Localized saliency of features using mixture of Gaussian distributions, and outliers detection .

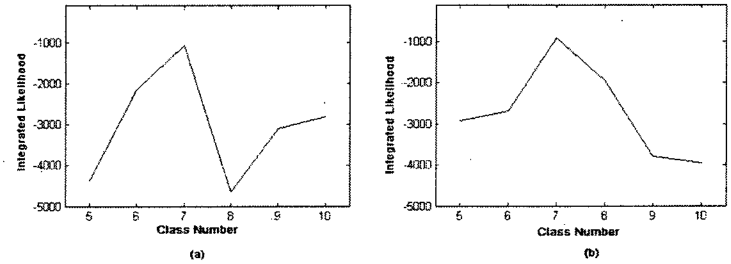

Figure 3.31: Integrated likelihood as a function of the number of clusters for the MPEG-7 data set using localized feature selection method, and outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

Table 3.23: Confusion matrix for MPEG-7 data set using mixture of Gamma distributions, globalized features selection, and outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer	Outlier Class
Bone	20	0	0	0	0	0	0	0
Heart	0	18	2	0	0	0	0	0
Glass	0	0	18	0	2	0	0	0
Fountain	0	1	0	19	0	0	0	0
Key	0	0	0	0	18	0	2	0
Fork	0	0	0	0	0	19	1	0
Hummer	0	0	0	0	0	1	19	0
Outlier Class	0	0	0	0	0	0	0	5

Table 3.24: Confusion matrix for MPEG-7 data set using mixture of Gaussian distributions, globalized features selection method, and outliers detection.

Classified as ⇒	Bone	Heart	Glass	Fountain	Key	Fork	Hummer	Outlier Class
Bone	19	0	0	0	0	0	1	0
Heart	3	16	0	1	0	0	0	0
Glass	0	0	18	2	0	0	0	0
Fountain	0	0	3	13	0	0	4	0
Key	0	0	0	0	19	1	0	0
Fork	2	0	0	0	0	18	0	0
Hummer	0	0	0	3	1	1	15	0
Outlier Class	0	0	0	0	0	2	0	3

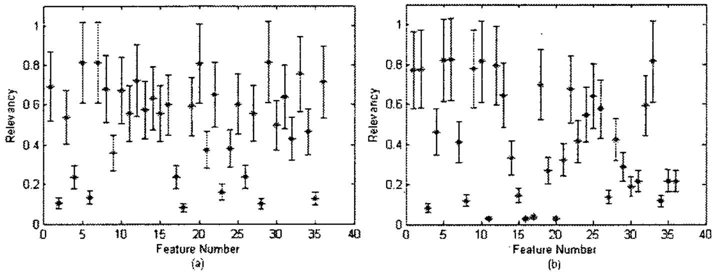

Figure 3.32: Globalized saliency of features using globalized features selection method, and outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

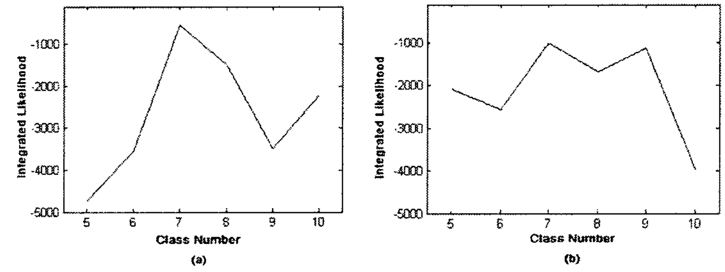

Figure 3.33: Integrated likelihood as a function of the number of clusters for the MPEG-7 data set using globalized feature selection method, and outliers detection. (a) Mixture of Gamma Distributions. (b) Mixture of Gaussian Distributions.

CHAPTER 4

Conclusions

Recently, there has been increased number of unlabeled databases, this is due to the large amount of data generated by human activities and scientific disciplines, which make a challenge for humans to assign (i.e. manually label) each instance into its category. Such a process is a subjective and expensive one. More challenge is added when each element of these databases is high-dimensional (text, images, etc). Typically, some features used to describe each element are not relevant for classification, or are "noisy" features. Hence, the process of selecting the most appropriate features to represent such an element becomes a necessary task, in order to remove noisy features and keep the most discriminating ones. Another more challenging issue is the presence of outlier data; which makes its difficult even for the most efficient learning algorithms to accurately identify the natural groupings of these databases.

In this thesis, we have presented a principled unsupervised generative model-based approach, for simultaneous clustering feature selection and outlier detection, to achieve robust data modeling using finite multivariate Gamma mixture model. Our proposal was to use a statistical model that makes explicit what data or features have to be ignored and what information has to be retained.

Our work is mainly driven by the increased collection of high-dimensional non-Gaussian data in various domains, and by the complexity of both feature selection and outlier detection problems in such domains. It has been shown through extensive experiments involving synthetic and real data that the proposed approach has excellent modeling capabilities and that feature selection mixed with outliers detection influences significantly the clustering performance.

Future works can be devoted to extend the proposed model to online settings using a variational approach, for instance, since we generally deal with dynamically changing environments; the proposed model can be extended to the semi-supervised case, such a model will use both labeled and unlabeled data for training. The main motivation toward moving to semi-supervised learning model is the difficulty of acquiring labeled data for learning. Such a process, is hard, expensive and subjective for a human being to assign each element of the data sets into its category. On the other hand, acquiring unlabeled data is an inexpensive and more feasible process. Hence, semi-supervised approaches use a limited amount of labeled data and a large amount of unlabeled data, and seem to provide more support in such a demanding environment.

CHAPTER 5

Appendices

5.1 Appendix 1: Proof of Equation 7

Note that we have to introduce a lagrange multiplier Λ to incorporate the constraint $\sum_{j=1}^{M+1} p_j = 1$. Computing the derivative of $\log p(\mathcal{X}|\Theta)$ w.r.t $p_j, j = 1, \dots, M+1$, we obtain

$$\frac{\partial\Big[\sum_{i=1}^{N} \log\Big(\sum_{j=1}^{M}\Big(p_j \prod_{d=1}^{D}\big(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\big)\Big) + p_{M+1}U(\vec{X}_i)\Big) + \Lambda(1 - \sum_{j=1}^{M+1} p_j)\Big]}{\partial p_j}$$

$$\begin{aligned}
&= \left[\sum_{i=1}^{N} \frac{\prod_{d=1}^{D}\big(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\big)}{\sum_{j=1}^{M}\Big(p_j \prod_{d=1}^{D}\big(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\big)\Big) + p_{M+1}U(\vec{X}_i)} - \Lambda\right] \\
&= \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)}{p_j} - \Lambda = 0
\end{aligned}$$

where

$$p(j|\vec{X}_i) = \begin{cases} \dfrac{p_j \prod_{d=1}^{D}\big(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\big)}{\sum_{j=1}^{M}\Big(p_j \prod_{d=1}^{D}\big(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\big)\Big) + p_{M+1}U(\vec{X}_i)} & \text{if } j = 1, \dots, M \\ \dfrac{p_{M+1}U(\vec{X}_i)}{\sum_{j=1}^{M}\Big(p_j \prod_{d=1}^{D}\big(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\big)\Big) + p_{M+1}U(\vec{X}_i)} & \text{if } j = M+1 \end{cases}$$

$p_j = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)}{\Lambda}$. Taking the derivative w.r.t Λ, we obtain $1 - \sum_{j=1}^{M+1} p_j = 0$ which gives us $\sum_{j=1}^{M+1} p_j = 1$. Thus,

$$\sum_{j=1}^{M+1} \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)}{\Lambda} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M+1} p(j|\vec{X}_i)}{\Lambda} = 1$$

since $\sum_{j=1}^{M+1} p(j|\vec{X}_i) = 1$, we obtain $L = N$, then $p_j = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)}{N}$

5.2 Appendix 2: Proof of Equation 8

Let $\rho_{jd} = \rho_{jd,1}$ and $1 - \rho_{jd} = \rho_{jd,2}$. By computing the derivative w.r.t $\rho_{jd,1}, j = 1, \ldots, M$, $d = 1, \ldots, D$ and introducing the lagrange multiplier Λ to take into account the fact that $\rho_{jd,1} + \rho_{jd,2} = 1$, we obtain

$$\frac{\partial \Big[\sum_{i=1}^{N} \log \Big(\sum_{j=1}^{M} \Big(p_j \prod_{d=1}^{D} \big(\rho_{jd,1} p(X_{id}|\theta_{jd}) + \rho_{jd,2} p(X_{id}|\lambda_{jd}) \big) \Big) + p_{M+1} U(\vec{X}_i) \Big) + \Lambda(1 - \rho_{jd,1} - \rho_{jd,2}) \Big]}{\partial \rho_{jd,1}}$$

$$= \sum_{i=1}^{N} p(j|\vec{X}_i) \left(\frac{p(X_{id}|\theta_{jd})}{\rho_{jd,1} p(X_{id}|\theta_{jd}) + \rho_{jd,2} p(X_{id}|\lambda_{jd})} \right) - \Lambda = 0$$

Multiplying by $\rho_{jd,1}$, we obtain

$$\sum_{i=1}^{N} p(j|\vec{X}_i) \left(\frac{\rho_{jd,1} p(X_{id}|\theta_{jd})}{\rho_{jd,1} p(X_{id}|\theta_{jd}) + \rho_{jd,2} p(X_{id}|\lambda_{jd})} \right) - \Lambda \rho_{jd,1} \Big] \quad = \quad 0 \tag{1}$$

By computing the derivative w.r.t $\rho_{jd,2}$, we obtain $\sum_{i=1}^{N} p(j|\vec{X}_i) \left(\frac{p(X_{id}|\lambda_{jd})}{\rho_{jd,1} p(X_{id}|\theta_{jd}) + \rho_{jd,2} p(X_{id}|\lambda_{jd})} \right) - \Lambda = 0$. Multiplying by $\rho_{jd,2}$, we obtain

$$\sum_{i=1}^{N} p(j|\vec{X}_i) \left(\frac{\rho_{jd,2} p(X_{id}|\lambda_{jd})}{\rho_{jd,1} p(X_{id}|\theta_{jd}) + \rho_{jd,2} p(X_{id}|\lambda_{jd})} \right) - \Lambda \rho_{jd,2} = 0 \tag{2}$$

Summing equations 1 and 2, we obtain $\sum_{i=1}^{N} p(j|\vec{X}_i) = \Lambda$, then according to equation 1, we have

$$\rho_{jd,1} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i) \left(\frac{\rho_{jd,1} p(X_{id}|\theta_{jd})}{\rho_{jd,1} p(X_{id}|\theta_{jd}) + \rho_{jd,2} p(X_{id}|\lambda_{jd})} \right)}{\sum_{i=1}^{N} p(j|\vec{X}_i)} \tag{3}$$

and

$$\rho_{jd,2} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i) \left(\frac{\rho_{jd,2} p(X_{id}|\lambda_{jd})}{\rho_{jd,1} p(X_{id}|\theta_{jd}) + \rho_{jd,2} p(X_{id}|\lambda_{jd})} \right)}{\sum_{i=1}^{N} p(j|\vec{X}_i)} \tag{4}$$

5.3 Appendix 3: Proof of Equations 9 and 10

Computing the derivative of $\log p(\mathcal{X}|\Theta)$ w.r.t β_{jd}, we obtain

$$\frac{\partial \log p(\mathcal{X}|\Theta)}{\partial \beta_{jd}} = \sum_{i=1}^{N} p(j|\vec{X}_i) \frac{\partial}{\partial \beta_{jd}} \log \Big[\big(\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd}) \big) \Big]$$

$$= \sum_{i=1}^{N} p(j|\vec{X}_i) \frac{\frac{\partial\left(\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})\right)}{\partial \beta_{jd}}}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})} = \sum_{i=1}^{N} p(j|\vec{X}_i) \frac{\rho_{jd} p(X_{id}|\theta_{jd}) \frac{\partial}{\partial \beta_{jd}} \log p(X_{id}|\theta_{jd})}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})}$$

we have $\frac{\partial}{\partial \beta_{jd}} \log p(X_{id}|\theta_{jd}) = \frac{\partial}{\partial \beta_{jd}} \left[(\alpha_{jd}-1) \log X_{id} - \frac{X_{id}}{\beta_{jd}} - \alpha_{jd} \log \beta_{jd} - \log \Gamma(\alpha_{jd}) \right] = \frac{X_{id}}{\beta_{jd}^2} - \frac{\alpha_{jd}}{\beta_{jd}}$.
Then, $\frac{\partial \log p(\mathcal{X}|\Theta)}{\partial \beta_{jd}} = 0$ gives us $\beta_{jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i) \frac{\rho_{jd} p(X_{id}|\theta_{jd}) X_{id}}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})}}{\alpha_{jd} \sum_{i=1}^{N} p(j|\vec{X}_i) \frac{\rho_{jd} p(X_{id}|\theta_{jd})}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})}}$.
Computing the derivative of $\log p(\mathcal{X}|\Theta)$ w.r.t α_{jd}, we obtain

$$\frac{\partial \log p(\mathcal{X}|\Theta)}{\partial \alpha_{jd}} = \sum_{i=1}^{N} p(j|\vec{X}_i) \frac{\rho_{jd} p(X_{id}|\theta_{jd}) \frac{\partial}{\partial \alpha_{jd}} \log p(X_{id}|\theta_{jd})}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})}$$

we have $\frac{\partial}{\partial \alpha_{jd}} \log p(X_{id}|\theta_{jd}) = \frac{\partial}{\partial \alpha_{jd}} \left[(\alpha_{jd}-1) \log X_{id} - \frac{X_{id}}{\beta_{jd}} - \alpha_{jd} \log \beta_{jd} - \log \Gamma(\alpha_{jd}) \right] = \log X_{id} - \log \beta_{jd} - \Psi(\alpha_{jd})$, where $\Psi()$ is the digamma function. Thus
$\alpha_{jd} = \Psi^{-1} \left(\frac{\sum_{i=1}^{N} p(j|\vec{X}_i) \frac{\rho_{jd} p(X_{id}|\theta_{jd})}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})} (\log X_{id} - \log \beta_{jd})}{\sum_{i=1}^{N} p(j|\vec{X}_i) \frac{\rho_{jd} p(X_{id}|\theta_{jd})}{\rho_{jd} p(X_{id}|\theta_{jd}) + (1-\rho_{jd}) p(X_{id}|\lambda_{jd})}} \right)$, where $\Psi^{-1}()$ is the inverse digamma function. Notice that it is possible also to use a Newton-Raphson method to estimate α_{jd}:

$$\alpha_{jd}^{new} = \alpha_{jd}^{old} - \frac{\partial \log p(\mathcal{X}|\Theta)}{\partial \alpha_{jd}} \left(\frac{\partial^2 \log p(\mathcal{X}|\Theta)}{\partial^2 \alpha_{jd}} \right)^{-1} \qquad (5)$$

It is straightforward to determine $\alpha_{\lambda|jd}$ and $\beta_{\lambda|jd}$ following the same development as above.

5.4 Appendix 4: Proof of Equations 20, 21, 22 and 23

The Hessian matrix is defined as, $H(\Theta) = \frac{\partial^2 - \log p(\mathcal{X}|\Theta)}{\partial^2 \Theta}$. Its determinant can be defined as follows: $|H(\Theta)| = |H(p_1, \ldots, p_{M+1})| \prod_{j=1}^{M} \left[|H(\theta_j)||H(\lambda_j)| \prod_{d=1}^{M} |H(\rho_{jd})| \right]$. Let's start by the Hessian matrix with respect to the mixing parameters $|H(p_1, \ldots, p_{M+1})|$ which should take into account the fact that $p_{M+1} = 1 - \sum_{j=1}^{M} p_j$. We have

$$\frac{\partial - \log p(\mathcal{X}|\Theta)}{\partial p_j} =$$

$$\left[-\sum_{i=1}^{N} \frac{\prod_{d=1}^{D}\left(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\right) - U(\vec{X}_i)}{\sum_{j=1}^{M}\left(p_j \prod_{d=1}^{D}\left(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\right)\right) + (1-\sum_{j=1}^{M}p_j)U(\vec{X}_i)}\right]$$

$$= \frac{\sum_{i=1}^{N} p(M+1|\vec{X}_i)}{p_{M+1}} - \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)}{p_j}, j = 1, \dots, M$$

By approximating the Hessian, with respect to the mixing parameters, from its diagonal components only, it is possible to show that [56]

$$|H(p_1, \dots, p_{M+1})| = \prod_{j=1}^{M}\sum_{i=1}^{N}\left(\frac{p(M+1|\vec{X}_i)}{p_{M+1}} - \frac{p(j|\vec{X}_i)}{p_j}\right)^2 \tag{6}$$

As for the Hessian, with respect to the ρ_{jd} parameters, we have

$$\frac{\partial\left[-\sum_{i=1}^{N}\log\left(\sum_{j=1}^{M}\left(p_j\prod_{d=1}^{D}\left(\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})\right)\right) + p_{M+1}U(\vec{X}_i)\right)\right]}{\partial\rho_{jd}}$$

$$= \sum_{i=1}^{N}p(j|\vec{X}_i)\left(\frac{p(X_{id}|\lambda_{jd}) - p(X_{id}|\theta_{jd})}{\rho_{jd}p(X_{id}|\theta_{jd}) + (1-\rho_{jd})p(X_{id}|\lambda_{jd})}\right) = \sum_{i=1}^{N}p(j|\vec{X}_i)\left(\frac{f(1-\rho_{jd}, \theta_{jd}, \lambda_{jd})}{1-\rho_{jd}} - \frac{f(\rho_{jd}, \theta_{jd}, \lambda_{jd})}{\rho_{jd}}\right)$$

An the determinant of the Hessian can be approximated as

$$|H(\rho_{jd})| = \sum_{i=1}^{N}p(j|\vec{X}_i)^2\left(\frac{f(1-\rho_{jd}, \theta_{jd}, \lambda_{jd})}{1-\rho_{jd}} - \frac{f(\rho_{jd}, \theta_{jd}, \lambda_{jd})}{\rho_{jd}}\right)^2 \tag{7}$$

According to the previous appendix, it is easy to show that:

$$\frac{\partial^2 - \log p(\mathcal{X}|\Theta)}{\partial^2\beta_{jd}} = -\sum_{i=1}^{N}p(j|\vec{X}_i)f(\rho_{jd}, \theta_{jd}, \lambda_{jd})\left(\frac{X_{id}}{\beta_{jd}^3} - \frac{\alpha_{jd}}{\beta_{jd}^2}\right) \tag{8}$$

$$\frac{\partial^2 - \log p(\mathcal{X}|\Theta)}{\partial\beta_{jd}\partial\alpha_{jd}} = \frac{\partial^2 - \log p(\mathcal{X}|\Theta)}{\partial\alpha_{jd}\partial\beta_{jd}} = \frac{1}{\beta_{jd}}\sum_{i=1}^{N}p(j|\vec{X}_i)f(\rho_{jd}, \theta_{jd}, \lambda_{jd}) \tag{9}$$

$$\frac{\partial^2 - \log p(\mathcal{X}|\Theta)}{\partial^2\alpha_{jd}} = \Psi'(\alpha_{jd})\sum_{i=1}^{N}p(j|\vec{X}_i)f(\rho_{jd}, \theta_{jd}, \lambda_{jd}) \tag{10}$$

where $\Psi'()$ is the trigamma function. Then,

$$\begin{aligned}|H(\theta_j)| &= \prod_{d=1}^{D}\Bigg[\left(\sum_{i=1}^{N}p(j|\vec{X}_i)f(\rho_{jd}, \theta_{jd}, \lambda_{jd})\left(\frac{\alpha_{jd}}{\beta_{jd}^2} - \frac{X_{id}}{\beta_{jd}^3}\right)\right)\Psi'(\alpha_{jd})\left(\sum_{i=1}^{N}p(j|\vec{X}_i)f(\rho_{jd}, \theta_{jd}, \lambda_{jd})\right) \\ &- \left(\frac{1}{\beta_{jd}}\sum_{i=1}^{N}p(j|\vec{X}_i)f(\rho_{jd}, \theta_{jd}, \lambda_{jd})\right)^2\Bigg]\end{aligned}$$

Using the same approach, we can show also that

$$
\begin{aligned}
|H(\lambda_j)| &= \prod_{d=1}^{D}\Bigg[\Bigg(\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})(\frac{\alpha_{\lambda|jd}}{\beta^2_{\lambda|jd}}-\frac{X_{id}}{\beta^3_{\lambda|jd}})\Bigg)\Psi'(\alpha_{\lambda|jd})\Bigg(\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})\Bigg) \\
&- \Bigg(\frac{1}{\beta_{\lambda|jd}}\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})\Bigg)^2\Bigg]
\end{aligned}
\tag{11}
$$

5.5 Appendix 5: Maximum Likelihood Estimation For Finite Gaussian Mixture Model

The parameters are estimated by maximizing the log-likelihood function using the maximum likelihood approach as following:

$$
\hat{\Theta} = \arg\max_{\Theta}\Bigg\{\log p(\mathcal{X}|\Theta) = \sum_{i=1}^{N}\log\Bigg[\sum_{j=1}^{M}\Bigg(p_j\prod_{d=1}^{D}\big(\rho_{jd}p(X_{id}|\theta_{jd})+(1-\rho_{jd})p(X_{id}|\lambda_{jd})\big)\Bigg)+p_{M+1}U(\vec{X}_i)\Bigg]\Bigg\}
\tag{12}
$$

where $\theta_{jd} = \{\mu_{jd}, \sigma^2_{jd}\}$, $\lambda_{jd} = \{\mu_{\lambda|jd}, \sigma^2_{\lambda|jd}\}$, and

$$
p(X_{id}|\mu_{jd},\sigma^2_{jd}) = \frac{1}{\sqrt{2\pi\sigma^2_{jd}}}e^{\frac{-(X_{id}-\mu_{jd})^2}{2\sigma^2_{jd}}}
\tag{13}
$$

is the Gaussian probability density function. μ_{jd} and σ^2_{jd} are mean and variance, which gives us:

$$
p_j = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)}{N} \qquad j = 1,\dots,M+1
\tag{14}
$$

$$
\rho_{jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)f(\rho_{jd},\theta_{jd},\lambda_{jd})}{\sum_{i=1}^{N} p(j|\vec{X}_i)} \qquad j = 1,\dots,M \quad d = 1,\dots,D
\tag{15}
$$

$$
\mu_{jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)f(\rho_{jd},\theta_{jd},\lambda_{jd})X_{id}}{\sum_{i=1}^{N} p(j|\vec{X}_i)f(\rho_{jd},\theta_{jd},\lambda_{jd})} \qquad j = 1,\dots,M \quad d = 1,\dots,D
\tag{16}
$$

$$
\sigma^2_{jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)f(\rho_{jd},\theta_{jd},\lambda_{jd})(X_{id}-\mu_{jd})^2}{\sum_{i=1}^{N} p(j|\vec{X}_i)f(\rho_{jd},\theta_{jd},\lambda_{jd})} \qquad j = 1,\dots,M \quad d = 1,\dots,D
\tag{17}
$$

$$
\mu_{\lambda|jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})X_{id}}{\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})} \qquad j = 1,\dots,M \quad d = 1,\dots,D
\tag{18}
$$

$$
\sigma^2_{\lambda|jd} = \frac{\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})(X_{id}-\mu_{\lambda|jd})^2}{\sum_{i=1}^{N} p(j|\vec{X}_i)f(1-\rho_{jd},\theta_{jd},\lambda_{jd})} \qquad j = 1,\dots,M \quad d = 1,\dots,D
\tag{19}
$$

List of References

[1] A. K. Jain, M. Murty and P. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[2] H. Xiong, J. Wu and J. Chen. K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(2):318–331, 2009.

[3] G. V. Trunk. A Problem of Dimensionality: A Simple Example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(3):306–307, 1979.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Morgan Kaufmann, second edition, 1990.

[5] Q. Hu, W. Pedrycz, D. Yu and J. Lang. Selecting Discrete and Continuous Features Based on Neighborhood Decision Error Minimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(1):137–150, 2010.

[6] Y. Pang, Y. Yuan and X. Li. Effective Feature Extraction in High-Dimensional Space. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 38(6):1652–1656, 2008.

[7] J-S. Wang and J-C. Chiang. A Cluster Validity Measure with Outlier Detection for Support Vector Clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 38(1):78–89, 2008.

[8] J. Muñoz-Garcia, J. L. Moreno-Rebollo and A. Pascual-Acosta. Outliers: A Formal Approach. *International Statistical Review*, 58(3):215–226, 1990.

[9] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley and Sons Ltd, 1994.

[10] Z. Lu and H. H. S. Ip. Generalized Competetive Learning of Gaussian Mixture Models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(4):901–909, 2009.

[11] A. Srivastava, A. Lee, E. Simoncelli and S. Zhu. On Advances in Statistical Modeling of Natural Images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003.

[12] H. Oja. On Location, Scale, Skewness and Kurtosis of Univariate Distributions. *Scandinavian Journal of Statistics*, 8(3):154–168, 1981.

[13] R. Samadani. A Finite Mixture Algorithm for Finding Proportions in SAR Images. *IEEE Transactions on Image Processing*, 4(8):1182–1186, 1995.

[14] A. El Zaart and D. Ziou and S. Wang and Q. Jiang. Segmentation of SAR Images. *Pattern Recognition*, 35(3):713–724, 2002.

[15] M. Petrou, F. Giorgini, and P. Smits. Modelling the Histograms of Various Classes in SAR Images. *Pattern Recognition Letters*, 23(9):1103–1107, 2002.

[16] I-T. Hsiao, A. Rangarajan and G. Gindi. Joint-MAP Bayesian Tomographic Reconstruction with a Gamma-Mixture Prior. *IEEE Transactions on Image Processing*, 11(12):1466–1477, 2002.

[17] D. Ziou and N. Bouguila. Unsupervised Learning of a Finite Gamma Mixture Using MML: Application to SAR Image Analysis. In *Proc. of 17th International Conference on Pattern Recognition (ICPR)*, pages 68–71, 2004.

[18] M. Petrou and A. Matrucceli. On the Stability of Thresholding SAR Images. *Pattern Recognition*, 31(11):1791–1796, 1998.

[19] D. Ziou, N. Bouguila, M. S. Allili and A. El Zaart. Finite Gamma Mixture Modeling Using Minimum Message Length Inference: Application to SAR Image Analysis. *International Journal of Remote Sensing*, 30(3):771–792, 2009.

[20] K. Pearson. Skew Variation in Homogeneous Material. *Philosophical Transactions, A*, 186:343–414, 1895.

[21] D. Wettschereck, D. W. Aha and T. Mohri. A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. *Artificial Intelligence Review*, 11(1-5):273–314, 1997.

[22] M. H. C. Law, M. A. T. Figueiredo and A. K. Jain. Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004.

[23] S. Boutemedjet, N. Bouguila and D. Ziou. A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1429–1443, 2009.

[24] N. Bouguila. A Model-Based Approach for Discrete Data Clustering and Feature Weighting Using MAP and Stochastic Complexity. *IEEE Transactions on Knowledge and Data Engineering*, 21(12):1649–1664, 2009.

[25] M. W. Graham and D. J. Miller. Unsupervised Learning of Parsimonious Mixtures on Large Spaces with Integrated Feature and Component Selection. *IEEE Transactions on Signal Processing*, 54(4):1289–1303, 2006.

[26] Y. Li, M. Dong and J. Hua. Simultaneous Localized Feature Selection and Model Detection for Gaussian Mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):953–960, 2009.

[27] P. J. Huber. *Robust Statistics*. New York: Wiley, 1981.

[28] B. Peirce. Criterion for the Rejection of Doubtful Observations. *The Astronomical Journal*, 45(21):161–163, 1852.

[29] J. O. Irwin. On a Criterion for the Rejection of Outlying Observations. *Biometrika*, 17(3/4):238–250, 1925.

[30] E. S. Pearson and C. C. Sekar. The Efficiency of Statistical Tools and A Criterion for the Rejection of Outlying. *Biometrika*, 28(3/4):308–320, 1936.

[31] T. Lewis and N. R. J. Fieller. A Recursive Algorithm for Null Distributions for Outliers: I. Gamma Samples. *Technometrics*, 21(3):371–376, 1979.

[32] G. E. P. Box and G. C. Tiao. A Bayesian Approach to Some Outlier Rejection. *Biometrika*, 55(1):119–129, 1968.

[33] D. F. Andrews and D. Pregibon. Finding the Outliers that Matter. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):85–93, 1978.

[34] M. Aitkin and G. T. Wilson. Mixture Models, Outliers, and the EM Algorithm. *Technometrics*, 22(3):325–331, 1980.

[35] D. J. Miller and J. Browning. A Mixture Model and EM-Based Algorithm for Class Discovery, Robust Classification and Outlier Rejection in Mixed Labeled/Unlabeled Data Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1468–1483, 2003.

[36] S. Tadjudin and D. A. Landgrebe. Robust Parameter Estimation for Mixture Model. *IEEE Transactions on Geoscience and Remote Sensing*, 38(1):439–445, 2000.

[37] M. K. Titsias and C. K. I. Williams. Sequentially Fitting Mixture Models using an Outlier Component. In *Proc. of the 6th International Workshop on Advances in Scattering and Biomedical Engineering*, pages 386–393, 2003.

[38] Q. Ke and T. Kanade. Robust Subspace Clustering by Combined Use of kNDD and SVD Algorithm. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 592–599, 2004.

[39] M. Hubert and S. Engelen. Robust PCA and Classification in Biosciences. *Bioinformatics*, 20(11):1728–1736, 2004.

[40] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall Ltd, London, 1990.

[41] A. R. Webb. Gamma Mixture Models for Target Recognition. *Pattern Recognition*, 33(12):2045–2054, 2000.

[42] K. Copsey and A. R. Webb. Bayesian Gamma Mixture Model Approach to Radar Target Recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1201–1217, 2003.

[43] A. R. Webb. *Statistical Pattern Recognition*. John Wiley and Sons Ltd, second edition, 2002.

[44] S. K. Das. Feature Selection with a Linear Dependence Measure. *IEEE Transactions on Computers*, 20(9):1106–1109, 1971.

[45] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. Chapman & Hall/CRC, 2008.

[46] J. Novovičová, P. Pudil and J. Kittler. Divergence Based Feature Selection for Multimodal Class Densities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(2):218–223, 1996.

[47] M. Dong Y. Li and J. Hua. Localized Feature Selection for Clustering. *Pattern Recognition Letters*, 29(1):10–18, 2008.

[48] P. J. Rousseeuw and B. C. Van Zomeren. Unmasking Multivariate Outliers and Leverage Points (with discussion). *Journal of the American Statistical Association*, 85(411):633–651, 1990.

[49] C. V. Stewart. MINPRAN: A New Robust Estimator for Computer Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):925–938, 1995.

[50] M. J. Black and A. D. Jepson. EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 329–342, 1996.

[51] C. K. I. Williams and M. K. Titsias. Greedy Learning of Multiple Objects in Images using Robust Statistics and Factorial Learning. *Neural Computation*, 16(5):1039–1062, 2003.

[52] H. O. Hartley. Maximum Likelihood Estimation from Incomplete Data. *Biometrics*, 14(2):174–194, 1958.

[53] G.J. McLachlan and D. Peel. *Finite Mixture Models*. New York: Wiley, 2000.

[54] D. M. Chickering and D. Heckerman. Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. *Machine Learning*, 29:181–212, 1997.

[55] N. Bouguila and D. Ziou. High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1716–1731, 2007.

[56] S. J. Roberts and L. Rezek. Bayesian Approach to Gaussian Mixture Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

[57] C.L. Blake and C.J. Merz. UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/mlearn/MLRepository.html. 1998.

[58] J. Sklansky. Image Segmentation and Feature Extraction. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(4):237–247, 1978.

[59] E. C. Kintner. On the Mathematical Properties of the Zernike Polynomials. *Journal of Modern Optics*, 23(8):679–680, 1976.

[60] C-H. Teh and R. T. Chin. On Image Analysis by the Method of Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496–513, 1988.

[61] R. Mukundan and K. R. Ramakrishnan. Fast Computation of Legendre and Zernike Moments. *Pattern Recognition*, 28(9):1433–1442, 1995.

[62] A. Khotanzad and Y. H. Hong. Invariant Image Recognition by Zernike Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990.

[63] Y. S. Kim and W. Y. Kim. Content-Based Trademark Retrieval System using a Visually Salient Feature. *Image and Vision Computing*, 16(12-13):931–939, 1998.

[64] R. Mukundan and K. R. Ramakrishnan. *Moment Functions in Image Analysis - Theory and Applications*. World Scientific, 1998.

[65] Y. S. Abu-Mostafa and D. Psaltis. Image Normalization by Complex Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(1):46–55, 1985.