Robust Model-based Learning via Spatial-EM Algorithm

Kai Yu, Xin Dang, Henry Bart, Jr. and Yixin Chen, *Member, IEEE*

Abstract—This paper presents a new robust EM algorithm for the finite mixture learning procedures. The proposed Spatial-EM algorithm utilizes median-based location and rank-based scatter estimators to replace sample mean and sample covariance matrix in each M step, hence enhancing stability and robustness of the algorithm. It is robust to outliers and initial values. Compared with many robust mixture learning methods, the Spatial-EM has the advantages of simplicity in implementation and statistical efficiency. We apply Spatial-EM to supervised and unsupervised learning scenarios. More specifically, robust clustering and outlier detection methods based on Spatial-EM have been proposed. We apply the outlier detection to taxonomic research on fish species novelty discovery. Two real datasets are used for clustering analysis. Compared with the regular EM and many other existing methods such as K-median, X-EM and SVM, our method demonstrates superior performance and high robustness.

Index Terms—Clustering, EM algorithm, finite mixture, spatial rank, outlier detection, robustness

1 Introduction

Finite mixture models are powerful and flexible to represent arbitrarily complex probabilistic distribution of data. Mixture model-based approaches have been increasingly popular. Applications in a wide range of fields have emerged in the past decades. They are used for density estimation in unsupervised clustering [28], [15], [13], [48], [9], [25], for estimating class-conditional densities in supervised learning settings [1], [15], [36], and for outlier detection purposes [38], [30], [52]. Comprehensive surveys on mixture models and their applications can be found in the monographs by Titterington *et al.* [44] and McLachlan and Peel [29].

Usually parameters of a mixture model are estimated by the maximum likelihood estimate (MLE) via the expectation maximization (EM) algorithm [11], [27]. It is well known that the MLE can be very sensitive to outliers. To overcome this limitation, various robust alternatives have been developed. Rather than maximizing the likelihood function of Gaussian mixtures, Markatou [23] used a weighted likelihood with down-weights on outliers. Neykov *et al.* [26] proposed a weighted trimmed likelihood framework to accommodate many interesting cases including the weighted likelihood method. Fujisawa and Eguchi [16] utilized a so-called β-likelihood to overcome the unboundedness of the likelihood function and sensitivity of the maximum likelihood estimator to outliers. Qin and Priebe [35] introduced a maximum L_q-likelihood estimation of mixture models and studied its robustness property. Peel and McLachlan [33], [28] considered modelling a mixture of t distributions to reduce the effects of outliers. Shoham [41] also used t mixtures to handle outliers and agglomerated an annealing approach to deal with the sensitivity with respect to initial values.

Another common technique for robust fitting of mixtures is to update the component estimates on the M-step of the EM algorithm by some robust location and scatter estimates. M-estimator has been considered by Campbell [5], Tadjudin and Landgrebe [43]. Hardin and Rocke [17] used Minimum Covariance Determinant (MCD) estimator for cluster analysis. Bashir and Carter [1] recommended the use of S estimator. In this paper, we propose to apply spatial rank based location and scatter estimators. They are highly robust and are computationally and statistically more efficient than the above robust estimators [54]. We develop a Spatial-EM algorithm for robust finite mixture learning. Based on the Spatial-EM, supervised outlier detection and unsupervised clustering methods are illustrated and compared with other existing techniques.

The remainder of the paper is organized as follows. Section 2 reviews mixture elliptical models and the EM algorithm. Section 3 introduces spatial rank related statistics. Section 4 presents the Spatial-EM algorithm for mixture elliptical models. Section 5

- *K. Yu is with Amazon Web Service,1918 8th Ave, Seattle, WA 98101 E-mail: yukai@amazon.com.*
- *X. Dang is with Department of Mathematics, University of Mississippi, 315 Hume Hall, University, MS 38677, USA. Telephone: (662)915-7409. Fax: (662)915-2361. E-mail: xdang@olemiss.edu. Webpage: http://olemiss.edu/∼xdang*
- *H. Bart, Jr. is with Department of Ecology and Evolutionary Biology, Tulane University, New Orleans, LA 70118, USA. E-mail: hbartjr@tulane.edu.*
- *Y. Chen is with Department of Computer and Information Science, University of Mississippi, 201 Weir Hall, University, MS 38677, USA. Telephone: (662)915-7438. Fax: (662)915-5623. E-mail: ychen@cs.olemiss.edu. Webpage: http://cs.olemiss.edu/∼ychen*

formulates mixture model based novelty detection. We apply the Spatial-EM based outlier detection to new species discovery in taxonomy research. In Section 6, the clustering method based on robust mixture learning is illustrated using two data sets from UCI machine learning repository. We end the paper in Section 7 with some concluding remarks and a discussion of possible future work.

2 Review of EM Algorithm

2.1 Finite Mixture Models

A d-variate random vector $\boldsymbol{X}$ is said to follow a K-component mixture distribution if its density function has the form of

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{K} \tau_j f_j(\boldsymbol{x}|\boldsymbol{\theta}_j),$$

where $f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)$ denotes the conditional probability density function of $\boldsymbol{x}$ belonging to the j^{th} component parametrized by $\boldsymbol{\theta}_j$, $\tau_1, ..., \tau_K$ are the mixing proportions with all $\tau_j > 0$ and $\sum_{j=1}^{K} \tau_j = 1$, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K, \tau_1, ..., \tau_K\}$ is the set of parameters.

For the mixture elliptical distributions, $f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)$ can be written as

$$f_j(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j) = |\Sigma_j|^{-1/2} h_j\{(\boldsymbol{x} - \boldsymbol{\mu_j})^T \Sigma_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu_j})\}, \tag{2.1}$$

for some $\boldsymbol{\mu}_j \in \mathbb{R}^d$, a positive definite symmetric $d \times d$ matrix Σ_j. The family of mixture elliptical distributions contains a quite rich collection of models. The most widely used one is the mixture of Gaussian distributions, in which

$$h_j(t) = (2\pi)^{-d/2} e^{-t/2}. \tag{2.2}$$

The mixture of t distributions and Laplace distributions are commonly used in modeling data with heavy tails. For the mixture t distributions,

$$h_j(t) = c(\nu_j, d)(1 + t/\nu_j)^{-(d+\nu_j)/2},$$

where ν_j is the degree freedom and $c(\nu_j, d)$ is the normalization constant. As a generalization of multivariate mixture Laplace distribution, the mixture of Kotz type distribution [34] has the density

$$h_j(t) = \frac{\Gamma(d/2)}{(2\pi)^{d/2}\Gamma(d)} e^{-\sqrt{t}}. \tag{2.3}$$

For detailed and comprehensive accounts on mixture models, see McLachlan and Peel [29].

2.2 EM algorithm

In the EM framework for finite mixture models, the observed sample $\mathcal{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ are viewed as incomplete. The complete data shall be $\mathcal{Z} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$, where $\boldsymbol{y}_i = (y_{1i}, ..., y_{Ki})^T$ is an "unobserved " indicator vector with $y_{ji} = 1$ if $\boldsymbol{x}_i$ is from component j, zero otherwise. The log-likelihood of $\mathcal{Z}$ is then defined by

$$L_c(\boldsymbol{\theta}|\mathcal{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{K} y_{ji} \log[\tau_j f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)]. \tag{2.4}$$

The EM algorithm obtains a sequence of estimates $\{\boldsymbol{\theta}^{(t)}, t = 0, 1, ...\}$ by alternating two steps until some convergence criterion is met.

E-Step: Calculate Q function, the conditional expectation of the complete log-likelihood, given $\mathcal{X}$ and the current estimate $\boldsymbol{\theta}^{(t)}$. Since Y_{ji} is either 1 or 0, $E(Y_{ji}|\boldsymbol{\theta}^{(t)}, \boldsymbol{x}_i) = \Pr(Y_{ji} = 1|\boldsymbol{\theta}^{(t)}, \boldsymbol{x}_i)$, which is denoted as $T_{ji}^{(t)}$. By the Bayes rule, we have

$$T_{ji}^{(t)} = \frac{\tau_j^{(t)} f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_i^{(t)})}{\sum_{j=1}^{K} \tau_j^{(t)} f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j^{(t)})}. \tag{2.5}$$

$T_{ji}^{(t)}$'s can be interpreted as soft labels at the t^{th} iteration. Replacing y_{ji} with T_{ji} in (2.4), we have $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

M-step: Update the estimate of the parameters by maximizing the Q function

$$\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{2.6}$$

For convenience, we define

$$w_{ji}^{(t)} = \frac{T_{ji}^{(t)}}{\sum_{i=1}^{n} T_{ji}^{(t)}}. \tag{2.7}$$

$w_{ji}^{(t)}$ can be viewed as the current weight of $\boldsymbol{x}_i$ contributing to component j. In the case of Gaussian mixture, maximizing Q with respective to $\{\boldsymbol{\mu}_j, \Sigma_j, \tau_j\}_{j=1}^K$ provides an explicit close-form solution :

$$\tau_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{ji}^{(t)}}{\sum_{j=1}^{K}\sum_{i=1}^{n} T_{ji}^{(t)}} = \frac{1}{n}\sum_{i=1}^{n} T_{ji}^{(t)}, \tag{2.8}$$

$$\boldsymbol{\mu}_j^{(t+1)} = \sum_{i=1}^{n} w_{ji}^{(t)} \boldsymbol{x}_i, \tag{2.9}$$

$$\Sigma_j^{(t+1)} = \sum_{i=1}^{n} w_{ji}^{(t)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})^T. \tag{2.10}$$

EM estimation has been proved to converge to maximum likelihood estimation (MLE) of the mixture parameters under mild conditions [11], [51], [27]. The above simple implementation makes Gaussian mixture models popular. However, a major limitation of Gaussian mixture models is their lack of robustness to outliers. This is easily understood because maximization of likelihood function under an assumed Gaussian distribution is equivalent to finding the least-squares solution, whose lack of robustness is well known. Moreover, from the perspective of robust statistics, using sample mean (2.9) and sample covariance (2.10) of each component in the M-step causes the

sensitivity problem because they have the lowest possible breakdown point. Here the breakdown point is a prevailing quantitative robustness measure proposed by Donoho and Huber [12]. Roughly speaking, the breakdown point is the minimum fraction of "bad" data points that can render the estimator beyond any boundary. It is clear to see that one point $\|\boldsymbol{x}\| \to \infty$ is enough to ruin the sample mean and sample covariance matrix. Thus, their breakdown point is $1/n$.

As a robust alternative, mixtures of t-distributions have been used for modeling data that have wider tails than Gaussian's observations [33], [41]. The EM implementation treats each t-distributed component as a weighted average Gaussian distribution with weight being a gamma distribution parameterized by the degree freedom ν_j. There are two issues in this approach. One is that there is no closed-form expression for $\nu_j^{(t+1)}$ in the M-step. Solving a nonlinear equation for ν_j through a greedy search is time-consuming. The other issue is a non-vanishing effect of an outlier on estimating $\Sigma_j^{(t+1)}$. Although some modifications [20], [21] have been proposed to address these issues for a single t-distribution and applied to the mixtures, those estimators, including M-estimators, are not strictly robust in the sense of the breakdown point, especially in high dimensions. The phenomena of low breakdown point of MLE of t-mixture had been observed by Tadjudin [43] and Shoham [41]. Huber [19] found that the breakdown point of scatter M-estimator in d dimension is less than $1/(d+1)$, which is disappointingly low.

We propose a new robust EM algorithm for mixtures of elliptical distributions, utilizing robust location and scatter estimators in each M-step. The estimators are based on multivariate spatial rank statistics, achieving the high possible breakdown point, which is asymptotically $1/2$. As shown later, our method can be viewed as a least L_1 approach in contrast to a least squared (L_2) approach in the regular EM.

3 Spatial Rank Related Statistics

3.1 Spatial Rank, Depth, and Median

We start the discussion on spatial rank in one dimension. We shall clarify that the term "spatial" refers to data space, not usual geographic space. Given a sample $\mathcal{X} = \{x_1, ..., x_n\}$ from a distribution F, it is well known that the sample mean minimizes the (average) squared distance to the sample, while the sample median minimizes the (average) absolute distance. That is, the sample median is the solution of

$$R(x, \mathcal{X}) = \nabla_x(\frac{1}{n}\sum_{i=1}^{n}|x - x_i|) = \frac{1}{n}\sum_{i=1}^{n} s(x - x_i) := 0, \tag{3.1}$$

where $s(\cdot)$ is the sign function defined as $s(x) = x/|x| = \pm 1$ when $x \neq 0$, $s(0) = 0$. $R(x, \mathcal{X})$ is called the *centered* rank function. The sample median has a centered rank of 0. For an order statistics without a tie $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, their centered ranks are $-1 + 1/n, -1 + 3/n, \cdots, 1 - 3/n, 1 - 1/n$, which are linear transformations from their naturally-ordered ranks $1, ..., n$. Such a center-oriented transformation is of vital importance for a rank concept in high dimensions where the natural ordering in 1D no longer exists.

Replacing $|\cdot|$ in (3.1) by Euclidean norm $\|\cdot\|$, we obtain a multivariate median and rank function.

$$\boldsymbol{R}(\boldsymbol{x}, \mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{x} - \boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{x} - \boldsymbol{x}_i}{||\boldsymbol{x} - \boldsymbol{x}_i||}, \tag{3.2}$$

where $\boldsymbol{s}(\cdot)$ is the *spatial sign* function such that $\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$, ($\boldsymbol{s}(\boldsymbol{0}) = \boldsymbol{0}$). $\boldsymbol{R}(\boldsymbol{x}, \mathcal{X})$ is called the *spatial rank* of $\boldsymbol{x}$ with respective to $\mathcal{X}$, and the solution of $\boldsymbol{R}(\boldsymbol{x}, \mathcal{X}) = \boldsymbol{0}$ is called the *spatial median*. The spatial median, also termed as geometric median, L_1 median, has a century-long history dating back to Weber [50]. Brown [3] has developed many properties of the spatial median. Similar to the univariate median, the spatial median is extremely robust with a breakdown point of $1/2$.

If we replace $|\cdot|$ in (3.1) by the L_1 norm, that is $\|\boldsymbol{x}\|_{L1} = |x_1| + ... + |x_d|$, we obtain the component wise rank and the corresponding median. The component-wise median has been widely used because of its conceptual simplicity and computational ease. But the component-wise median may be a very poor center representative of data, because it disregards the interdependence information among the variables and is calculated separately on each dimension. Like its univariate counterpart, the component-wise median may not unique, is not affine equivariant and not even orthogonal equivariant. For those reasons, the spatial median and spatial rank are more appealing.

The spatial rank provides a relative position of $\boldsymbol{x}$ with respect to $\mathcal{X}$. Its magnitude yields a measure of outlyingness of $\boldsymbol{x}$. It is easy to prove that $\|\boldsymbol{R}(\boldsymbol{x}, \mathcal{X})\| \leq 1$ by simply applying Jensen's inequality. Hence equivalently, we can define the *spatial depth* function as $1 - \|\boldsymbol{R}(\boldsymbol{x}, \mathcal{X})\|$. The spatial median is the deepest point with the maximum spatial depth value of 1. The spatial depth produces, from the "deepest" point (the spatial median), a "center-outward ordering" of multidimensional data [**?**]. It is natural to conduct outlier detection in such a way that an observation with a depth value less than a threshold is declared as an outlier. Dang and Serfing [10] studied properties of depth-based outlier identifiers. Chen *et al.* [8] proposed the kernelized spatial depth (KSD) by generalizing the spatial depth via positive definite kernels and applied the KSD-based outlier detection to taxonomic study. We will compare their results on an experiment of fish species novelty discovery in Section 5.4.

The spatial rank $\boldsymbol{R}(\boldsymbol{x}, \mathcal{X})$ is the average unit directions to $\boldsymbol{x}$ from sample points of $\mathcal{X}$. Unlike its univari-

ate counterpart, the spatial rank is not distribution-free; it characterizes the distribution of $\mathcal{X}$, especially directional information of the distribution. For a better understanding, we also consider the population version

$$\boldsymbol{R}(\boldsymbol{x}, F) = \mathbb{E}\boldsymbol{s}(\boldsymbol{x} - \boldsymbol{X}),$$

where $\boldsymbol{X}$ is a random vector from the distribution F.

3.2 RCM and MRCM

Based on spatial ranks, the rank covariance matrix (RCM) of $\mathcal{X}$, denoted by $\Sigma_R(\mathcal{X})$, is

$$\Sigma_R(\mathcal{X}) = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{R}(\boldsymbol{x}_j, \mathcal{X})\boldsymbol{R}^T(\boldsymbol{x}_j, \mathcal{X}). \tag{3.3}$$

Notice that the spatial ranks of a sample are centered, i.e., $\frac{1}{n}\sum_j \boldsymbol{R}(\boldsymbol{x}_j, \mathcal{X}) = \boldsymbol{0}$. The RCM is nothing but the covariance matrix of the ranks. The corresponding population version is

$$\Sigma_R(F) = \mathbb{E}\boldsymbol{R}(\boldsymbol{X}, F)\boldsymbol{R}^T(\boldsymbol{X}, F) = \mathrm{cov}(\boldsymbol{R}(\boldsymbol{X}, F)).$$

For an elliptical distribution F with a scatter matrix Σ, the rank covariance matrix preserves the orientation information of F. Marden [22] has proved that the eigenvectors of Σ_R are the same as that of Σ. But their eigenvalues are different. Those results are easily understood by features of the spatial rank. Each observation contributes a unit directional vector to the spatial rank. It gains resistance to extreme observations, but in the meantime it trades off some variability measurement. Visuri *et al.* [47] proposed to re-estimate dispersion of the projected data on eigenvectors. The modified spatial rank covariance matrix (MRCM), $\tilde{\Sigma}(\mathcal{X})$, is constructed as follows.

1. Compute the sample RCM, $\Sigma_R(\mathcal{X})$, using (3.2) and (3.3).
2. Find the eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_d$ of $\Sigma_R(\mathcal{X})$ and denote the matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_d]$.
3. Find scale estimates (eigenvalues, principal values) of $\mathcal{X}$ on $\boldsymbol{u}_i$'s directions using an univariate robust scale estimate σ. Let $\hat{\lambda}_i = \sigma(\boldsymbol{u}_i^T\boldsymbol{x}_1, ..., \boldsymbol{u}_i^T\boldsymbol{x}_n)$ and denote $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1^2, ..., \hat{\lambda}_d^2)$.
4. The scatter estimate is $\tilde{\Sigma}(\mathcal{X}) = \boldsymbol{U}\hat{\Lambda}\boldsymbol{U}^T$.

Different choices of robust σ can be used. Here we use median absolute deviation (MAD), a well-known robust dispersion estimator defined as

$$1.486 \times \mathrm{med}_i|x_i - \mathrm{med}(x_1, ..., x_n)|.$$

The scaling factor 1.486 is the reciprocal of the 3^{rd} quartile of the Gaussian distribution. This particular choice of scaling factor makes MAD a consistent estimator of the standard deviation when data are from a Gaussian distribution.

Yu *et al.* [54] developed many nice properties of MRCM. It is affine equivariant under elliptical distributions, i.e.,

$$\tilde{\Sigma}(F_{AX+b}) = A\tilde{\Sigma}(F_X)A^T,$$

which is an important feature for a covariance matrix. $\tilde{\Sigma}(\mathcal{X})$ is statistically and computationally more efficient than other popular robust covariance estimators such as M-estimator, MCD, and S-estimator. It is highly robust with the highest possible breakdown point, i.e., asymptotically $1/2$.

So far, all the merits of spatial median and modified rank covariance matrix we discussed above are limited to one single elliptical distribution. For a mixture elliptical model, we next demonstrate a novel approach that integrate the EM algorithm with spatial rank methods.

4 SPATIAL-EM

4.1 Algorithm

The motivation on strengthening the robustness of regular EM algorithm on a mixture of Gaussian model comes from the closed forms of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ in the M-step. The idea of *Spatial-EM* is to replace sample mean and sample covariance matrix on M-step with the spatial median and MRCM.

Algorithm 1: Spatial-EM Algorithm

{Initialization} $t = 0$, $\boldsymbol{\mu}_j^{(0)}$, $\Sigma_j^{(0)} = \mathbf{I}$, $\tau_j^{(0)} = 1/K$ for $\forall j$
Do Until $\tau_j^{(t)}$'s converge for all j
  For $j = 1$ To K
E-Step:
    Calculate $T_{ji}^{(t)}$ by Equations (2.5), (2.1), (2.2)
M-Step:
    Update $\tau_j^{(t+1)}$ by Equation (2.8)
    Define $w_{ji}^{(t)}$ as Equation (2.7)
    Find $\boldsymbol{\mu}_j^{(t+1)}$ by Algorithm 2
    Find $(\tilde{\Sigma}_j^{(t+1)})^{-1}$ and $|\tilde{\Sigma}_j^{(t+1)}|^{-1/2}$ by Alg. 3
  End
  $t = t + 1$
End

Obviously, we need the following two algorithms for the spatial median and MRCM of j^{th} component.

Algorithm 2: Compute the weighted spatial median $\boldsymbol{\mu}_j^{(t+1)}$

Input $\{\boldsymbol{x}_i\}_{i=1}^n$, $\{w_{ji}^{(t)}\}_{i=1}^n$
For $\ell = 1$ To n
  $\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_\ell) = \sum_{i=1}^n w_{ji}^{(t)}\boldsymbol{s}(\boldsymbol{x}_\ell - \boldsymbol{x}_i)$
End
$\boldsymbol{\mu}_j^{(t+1)} = \arg\min_{\boldsymbol{x}_\ell} \|\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_\ell)\|$
Output $\{\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_\ell)\}_{\ell=1}^n$, $\boldsymbol{\mu}_j^{(t+1)}$

Algorithm 3: Compute the inverse of weighted MRCM $\tilde{\boldsymbol{\Sigma}}_j^{(t+1)}$

1 Input $\{\boldsymbol{x}_i, \boldsymbol{R}_j^{(t)}(\boldsymbol{x}_i), T_{ji}^{(t)}, w_{ji}^{(t)}\}_{i=1}^n$, $\boldsymbol{\mu}_j^{(t+1)}$, $\tau_j^{(t+1)}$

2 $\Sigma_{R,j}^{(t+1)} = \sum_{i=1}^n w_{ji}^{(t)} \big(\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_i)\big)\big(\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_i)\big)^T$

3 Find eigenvectors $\boldsymbol{U}_j = [\boldsymbol{u}_{j,1}, ..., \boldsymbol{u}_{j,d}]$ of $\Sigma_{R,j}^{(t+1)}$

4 For $m = 1$ To d

5 $\quad \boldsymbol{a}_m = \{T_{ji}^{(t)} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})\}_{i=1}^n$

6 $\quad$ Delete the $\lceil n(1-\tau_j^{(t+1)})\rceil$ smallest values of $\boldsymbol{a}_m$, denoted as $\{T_{ji_k}^{(t)} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_{i_k} - \boldsymbol{\mu}_j^{(t+1)})\}_{i_k}$

7 $\quad \hat{\lambda}_{jm} = \text{MAD}\ (\{T_{ji_k}^{(t)} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_{i_k} - \boldsymbol{\mu}_j^{(t+1)})\}_{i_k})$

8 End

9 $\hat{\Lambda}_j = \text{diag}(\hat{\lambda}_{j1}^2, ..., \hat{\lambda}_{jd}^2)$

10 Inverse MRCM $(\tilde{\Sigma}_j^{(t+1)})^{-1} = \boldsymbol{U}_j \hat{\Lambda}_j^{-1} \boldsymbol{U}_j^T$

11 Output $(\tilde{\Sigma}_j^{(t+1)})^{-1}$, $\prod_{m=1}^d \hat{\lambda}_{jm}^{-1}$

The Spatial-EM terminates when $\tau_j^{(t)}$ gets converged for all j or the number of iterations exceeds the pre-specified parameter maxiter. We set maxiter to be 100. K-means or other clustering methods can be used to assign initial values to $\boldsymbol{\mu}_j^{(0)}$.

4.2 On M-step

There are several places worth noting on M-step. The first one is the way to update $\boldsymbol{\mu}_j^{(t+1)}$. Rather than using a modified Weiszfeld algorithm [46] to coordinate component weights w_{ji}, we confine our search of the solution in the pool of sample points. i.e.,

$$\boldsymbol{\mu}_j^{(t+1)} \quad = \quad \arg\min_{\boldsymbol{x}_k} \left\| \sum_{i=1}^n w_{ji}^{(t)} \boldsymbol{s}(\boldsymbol{x}_k - \boldsymbol{x}_i) \right\|. \quad (4.1)$$

This would save a great amount of computational time and works fine when the sample size is relatively large.

Secondly, in defining MRCM for a certain component at the t^{th} iteration, we need to calculate a weighted RCM on Step 2 in Algorithm 3. It is not difficult to see for the points that can be well clustered into different components, $T_{ji}^{(t)}$ would be either close to 1 or 0. It is similar to a binary classification on whether a point belongs to j^{th} component or not. So the factor $w_{ji}^{(t)} = T_{ji}^{(t)} / \sum_{i=1}^n T_{ji}^{(t)}$, can provide a proper weight to average the elements that belongs to the j^{th} component. As the iteration goes on, the j^{th} component RCM would finally stand out by "picking" the correct ranks using $w_{ji}^{(t)}$.

Thirdly, the construction of MRCM becomes tricky because when we compute MAD, we have to consider soft membership T_{ji}. As shown on Step 5 in Algorithm 3, we project the centered data onto each eigen-direction, then multiply the factor $T_{ji}^{(t)}$ to generate the whole sequence of $\{T_{ji} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})\}_{i=1,..,n}$. Because each $T_{ji}^{(t)}$ plays as a classifier and degenerates to 0 if $\boldsymbol{x}_i$ does not belong to the j^{th} component, the above sequence contains many small values (probably sufficiently close to 0). This suggests that the corresponding data points may not belong to component j. Therefore we shall omit the smallest $\lceil n(1-\tau_j^{(t+1)})\rceil$ number of values, and apply MAD on the rest of projected data. Various experiments have shown that this approach performs very well.

Fourthly, there is a close relationship between K-median and our Spatial-EM algorithm. K-median treats the covariance matrix in each component all the same as the identity matrix, while our Spatial-EM estimates the covariance matrix of each component robustly at every iteration. Hence essentially K-median assumes independence among all variables and all variables having the same scale, which are very restrictive. K-median uses the component-wise median for the center of each component, while ours uses the spatial median. K-medoid is closely related to K-median method with the center of each cluster being a sample point. Of course, we have an option to use the spatial median in the K-median algorithm but major correlation information among variables has been lost in the covariance matrix and utilizing the spatial median seems not help much. On the other hand, we want to use the spatial median in our algorithm with the cost of a little extra computation time since we need to compute spatial ranks anyway.

4.3 More on M-Step

It is interesting to find that the proposed estimator is closely related to the maximum likelihood estimator (MLE) of a Kotz-type mixture model. As introduced in Section 2, a Kotz-type distribution belongs to the family of elliptical distributions. It has heavier tail regions than those covered by Gaussion distributions. Hence it is expected that a Kotz-type mixture model is more robust to outliers than Gaussian mixture models.

For a mixture of Kotz-type distribution, one can obtain the MLE by EM algorithm. In each M-step for component j, maximizing the Q function in (2.6) w.r.t $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ would be equivalent to minimizing the objective function $\rho(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, which is

$$\sum_{i=1}^n T_{ji} \big(\frac{1}{2} \ln |\boldsymbol{\Sigma}_j| + \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)}\ \big).$$

Setting the first derivatives of ρ w.r.t $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j^{-1}$ to be a zero vector and a zero matrix respectively, one can derive:

$$\sum_{i=1}^n w_{ji} \boldsymbol{s}(\boldsymbol{\Sigma}_j^{-1/2} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)) = \boldsymbol{0}, \quad (4.2)$$

$$\boldsymbol{\Sigma}_j = \sum_{i=1}^n w_{ji} \frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)}}. \quad (4.3)$$

The above equations look similar to the formula of the spatial median and spatial rank covariance matrix,

	0%		10%		20%	
	# Iters	Time (sec)	# Iters	Time (sec)	# Iters	Time (sec)
Spatial-EM	5.60 (0.68)	2.65 (0.40)	7.20 (1.20)	3.01 (0.56)	12.3 (5.25)	9.07 (4.17)
Reg-EM	6.80 (0.70)	0.04 (0.01)	20.4 (2.70)	0.14 (0.02)	25.0 (8.15)	0.24 (0.08)
Kotz-EM	5.15 (0.37)	2.57 (0.24)	92.3 (26.1)	53.2 (15.3)	96.7 (19.2)	67.8 (13.2)

TABLE 1
Convergence speed and computation time for each method. Standard deviations are included in parentheses.

but their computation is more demanding. Rao [37] proposed an iterative method to solve the above equations. First initialize $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}$, then find the solution of $\hat{\boldsymbol{\mu}}_j$ in (4.2) as $\hat{\boldsymbol{\mu}}_j = \hat{\boldsymbol{\Sigma}}_j^{1/2}\hat{\boldsymbol{\nu}}_j$, where $\hat{\boldsymbol{\nu}}_{\boldsymbol{j}}$ is the spatial median of the transformed data $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}^{-1/2}\boldsymbol{x}_i$'s. Plug $\hat{\boldsymbol{\mu}}_j$ into (4.3) to update $\hat{\boldsymbol{\Sigma}}_j$. The iteration stops until convergence. The MLE of $\boldsymbol{\mu}_j$ is called the generalized spatial median [37]. It minimizes the (average) Mahalanobis distances (associated with a covariance matrix) to sample points. It uses transformation-retransformation technique [40] to gain affine equivariance property. The two versions of spatial medians are not same unless $\boldsymbol{\Sigma}_j = c\boldsymbol{I}$, where $c > 0$ and the $\boldsymbol{I}$ is the identity matrix.

Although the EM algorithm for a mixture of Kotz type distribution is mathematically tractable, it suffers the same problems as that of mixture t-distributions. First of all, it is computationally expensive to compute. In order to solve for $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}$, an inner iteration has to be done in each M-step. It would significantly increase the computation complexity. Secondly, it is not strictly robust in the sense of the breakdown point. Comparing with Equation (2.10), each $\boldsymbol{x}_i$ in (4.3) is weighted by the reciprocal of its Mahalanobis distance, hence an outlier has less effect in the MLE of Kotz mixture than that of Gaussian mixture. However, the effect of an outlier is non-vanishing. For instance, for an extreme outlier $\boldsymbol{x}_i = c\boldsymbol{\mu}_j$ with $c \to \infty$, since the estimator of location μ_j in (4.2) is a median type estimator. It is not effected by a single outlier. Hence $\boldsymbol{x}_i$ is dominated in in (4.3) and the weight (the reciprocal of the Mahalanobis distance) is decreasing linearly to zero, while the numerator increases quadratically to infinite. To get the solution of (4.3), Σ_j should be in the form of $c^2\Sigma$. In this case, one extreme outlier breaks down the estimator. Indeed, if we modify (4.3) as

$$\boldsymbol{\Sigma}_j = \sum_{i=1}^{n} w_{ji} \frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T}{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)}$$

to match up the rate of weights, we obtain Tyler-M estimators [45]. The breakdown point increases to $1/d$, which is still low in high dimensions. The breakdown point for M estimators is not intuitively obvious. For a simplified proof, refer to Maronna *et al.* [24].

4.4 Convergence

Spatial-EM modifies the component estimates on each M-step by spatial median and rank covariance matrix to gain robustness at the cost of increasing computational burden and losing theoretical tractability. For single component elliptically symmetric models, consistency and efficiency of the rank covariance have been established in [47] and [54] with an amount of effort. The extension to a mixture model loses mathematical tractability due to deletion of a portion of smallest values of the projected data in the update of covariance matrix. The whole procedure hybridizes soft and hard labels at each iteration, which makes the connection to maximum likelihood approximation extremely difficult to verify theoretically. In such a desperate situation, demonstrating empirical evidence seems to be the only thing we can do.

We consider a sample consisting initially of 200 simulated points from a 3-component bivariate Gaussian mixture model, to which contamination points with increasing proportions 0%, 10%, 20% are added from a uniform distribution over the range -30 to 30 on each variate. The parameters of the mixture are,

$$\begin{aligned}
\boldsymbol{\mu}_1 &= (\ -6 \quad 6\ )^T & \Sigma_1 &= \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix} \\
\boldsymbol{\mu}_2 &= (\ 6 \quad -6\ )^T & \Sigma_2 &= \begin{pmatrix} 3 & -.5 \\ -.5 & 1 \end{pmatrix} \\
\boldsymbol{\mu}_3 &= (\ 6 \quad 6\ )^T & \Sigma_3 &= \begin{pmatrix} 4 & -.3 \\ -.3 & 1 \end{pmatrix}
\end{aligned}$$

with mixture proportions $\alpha_1 = 0.2$, $\alpha_2 = 0.2$ and $\alpha_3 = 0.6$. The procedure is repeated 20 times. The average and standard deviation of the number of iterations and computation time are reported in Table 1.

Without outliers, three algorithms have a similar convergence speed with average 6 iterations. The computation time of the regular EM is much faster than that of the other two. With 10% contamination, the spatial-EM takes 7 iterations to converge, while the regular EM triples that number and Kotz-EM is even worse. Out of 20 experiments, Kotz-EM only converged twice before 100 iterations. With 30% contamination, the regular EM takes twice of iterations as that of our method. Kotz-EM gets converged only 1 time. Kotz-EM is theoretically sound. The problem is the practical implementation. With inside and outside loops, the speed of convergence will slow down. On

the contrary, our method is not well-principled, but its convergence behavior seems satisfactory practically.

The regular EM takes much less time than ours. We shall use the regular EM if the assumption of mixture Gaussian is reasonable. However, in the outlier contamination case, we may want to pay some computational cost for good performance, that is when we should apply robust EM methods. Comparing with other robust EM methods, our proposed algorithm have advantages in low computational burden, high robustness and statistical efficiency. They are widely suitable for elliptical mixtures. In the next section, we will use the same simulation study to compare performance of the spatial-EM and the regular one under novelty detection problem.

In pattern recognition, finite mixtures are able to represent arbitrarily complex structure of data and have been successfully applied to unsupervised learning as well as supervised learning. Here we focus on applications of the Spatial-EM to supervised novelty detection and unsupervised clustering problem.

5 Novelty Detection

5.1 Outlyingness and Two-type Errors

Usually, an outlier region is associated with an outlyingness measure. For a finite mixture model, we use

$$H(\boldsymbol{x}) = \sum_{j=1}^{K} \tau_j G(\xi_j(\boldsymbol{x}))$$

as the outlyingness function to define outliers, where $\xi_j(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)$ and G is the cumulative distribution function (cdf) of $\chi^2(d)$ distribution. The reason behind it is from a well-known result. For a d-variate random vector $\boldsymbol{X}$ distributed as $N(\boldsymbol{\mu}, \Sigma)$, its Mahalanobis distance $(\boldsymbol{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$ follows a $\chi^2(d)$ distribution. Then the corresponding outlier region is

$$\{\boldsymbol{x} \in \mathrm{R}^d : H(\boldsymbol{x}) > 1-\varepsilon\}. \tag{5.1}$$

There are two-type errors for outlier detection: Type-I error and Type-II error.

$$P_{err1} = P(\text{identified as outlier}|\text{non-outlier}),$$
$$P_{err2} = P(\text{identified as non-outlier}|\text{outlier}).$$

Under a Gaussion mixture model, (5.1) has a Type-I error of ε. For a given data, $\boldsymbol{\theta} = \{\tau_j, \boldsymbol{\mu}_j, \Sigma_j\}_{i=1}^K$ are estimated and both types of errors can be estimated to evaluate performance of outlier detection methods given that those methods have the same number of parameters. $\hat{P}_{err1}$ is also called the false positive (alarm) rate. $\hat{P}_{err2}$ is the false negative rate and $1-\hat{P}_{err2}$ is the detection rate.

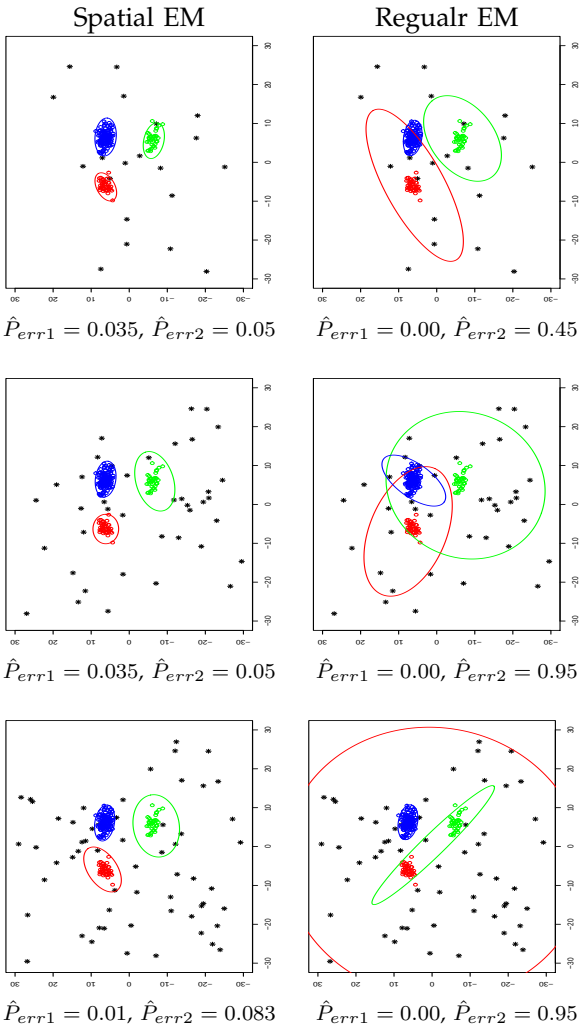


Fig. 1. Comparison between the Spatial-EM and regular EM on mixture of 3-component Gaussian distributions with outlier contamination. Ellipses in each plot represent the estimated 95% probability density contours of each component.

5.2 Estimating the Number of Components

We deal with supervised novelty detection as one-class learning problem. That is, the training sample contain "normal" observations, and the outlier detection is formulated as finding observations that significantly deviate from the training data. An important issue for this approach is the selection of the number of components K. We use a cross-validation approach and a so-called "one-standard-error" rule to choose the number of components [18]. The method starts by obtaining a set of candidate models for a range of values of K (from k_{min} to k_{max}) using cross-validation training data and estimating the average and standard deviation (sd) of Type-I errors using validation data. The number of components is then

$$\hat{K} = \arg\min_k \{\hat{P}_{err1}(k) \le \hat{P}_{err1}(\text{best } k) + sd\}. \tag{5.2}$$

That is, we choose the most parsimonious model whose mean $\hat{P}_{err1}$ is no more than on one standard deviation above the mean $\hat{P}_{err1}$ of the best model. In this way, the mixture model avoids the over-fitting problem and in the meantime preserves good performance.

5.3 Synthetic Data

We consider the same setup as the simulation did in Section 4.4. Increasing proportions 10%, 20%, 30% contamination points are added to 200 simulated points from a 3-component bivariate Gaussian mixture model. We set the parameter ε of the outlier detector (5.1) to be 0.05, which means that the probability of Type-I error is controlled to be 5%. Figure 1 compares performance of Spatial-EM based outlier detection and the regular-EM one. The Type-I errors are well-kept below 5% for both methods for all cases, however, Spatial-EM method achieves a much smaller Type-II error than the regular one. Our method is able to detect 95% outliers in the 10% contamination case, while the detection rate for the regular-EM is only 55%. When the contamination level increases to 20%, the regular-EM method completely fails with the detection rate 5%, while the proposed method has the detection rate 95%. Even in the case with 30% contamination, our method still maintains a 91.7% detection level.

5.4 New Species Discovery in Taxonomic Research

It is estimated that more than 90 percent of the world's species have not been described, yet species are being lost daily due to human destruction of natural habitats. The job of describing the earth's remaining species is exacerbated by the shrinking number of practicing taxonomists and the very slow pace of traditional taxonomic research. We believe that the pace of data gathering and analysis in taxonomy can be greatly increased through the integration of machine learning and data mining techniques into taxonomic research.

Here, we formulate new species discovery as an outlier detection problem. We apply the proposed Spatial-EM based novelty detection method to a small group of cypriniform fishes, comprising five species of suckers of the family *Catostomidae* and five species of minnows of the family *Cyprinidae*.

5.4.1 Data Set

The data set consists of 989 specimens from Tulane University Museum of Natural History(TUMNH). There are 10 species that include 128 *Carpiodes carpio*, 297 *Carpiodes cyprinus*, 172 *Carpiodes velifer*, 42 *Hypenteilum nigricans*, 36 *Pantosteus discobolus*, 53 *Campostoma olibolepis*, 39 *Cyprinus carpio*, 60 *Hybopsis storeriana*, 76 *Notropis petersoni*, and 86 *Luxilus zonatus*. We assign numbers 1 to 10 to the above species. The first five species belong to the family *Catostomidae* (suckers). The next five species belong to *Cyprinidae* (minnows). Both families are classified in the order *Cypriniformes*. For each species, 12 features are generated from 15 landmarks, which are biologically definable points along the body outline. In order to remove non-shape related variation in landmark coordinates, those 12 features have been translated to the origin and scaled to a common unit size. See [7] and [8] for a detailed description of the feature extraction process.

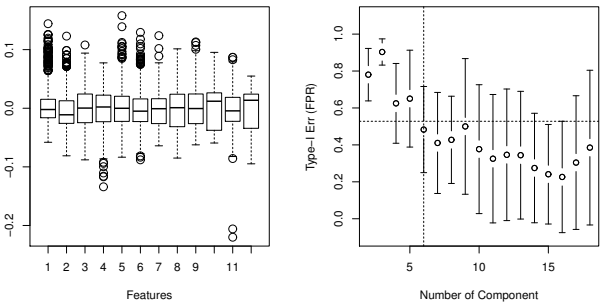


Fig. 2. (a) Box-plot of data for species 2 to 10; (b) One-standard-error rule for choosing the number of components.

5.4.2 Results

In this experiment, we treated specimens from one of the 10 species as a "undiscovered" specimens and specimens of the other 9 species as known. The nine known species are modeled as a finite mixture distribution. Figure 2 (a) is the box-plot of each feature for fishes species 2 to 10. The plot shows that the data set has a complex and heterogeneous structure (Plots on the other nine species are similar) and contains a considerable number of extreme values, which calls for a robust finite mixture modeling.

We first determine the number of components $\hat{K}$ using the "one-standard-error" rule by a 10-fold cross validation. As demonstrated in Figure 2 (b), the number of components is chosen to be 6. We then use the whole data to estimate the mixture parameters.

Two criteria are used for assessing performance. One is P_{err1} or P_{err2}. It evaluates the behavior of methods under the balanced error rate case, which is probably the most practical scenario. The other is the area under the ROC curve (AUC) that measures the overall performance. Note that the balanced error case is represented by the intersection of the ROC curve and the downward diagonal line.

For the first performance metric, the parameter ε of the outlier detector (5.1) is chosen such that

$$\hat{P}_{err1} \approx \hat{P}_{err2},$$

i.e., equal error rates. To demonstrate that our method is also robust to initial values, we repeat the procedure

Unknown Species	$\hat{P}_{err2}$ (also $\hat{P}_{err1}$)					AUC		
	Spatial-EM	Regular-EM	KSD	Gaussian	SVM	Spatial-EM	Regular-EM	Kotz-EM
Carpiodes carpio	[6] 0.260 (.040)	[9] 0.303 (.289)	0.234	0.408	0.156	0.821 (.014)	0.763 (.145)	0.802 (.088)
Carpiodes cyprinus	[8] 0.181 (.114)	[11] 0.212 (.230)	0.209	0.245	0.171	0.919 (.027)	0.932 (.043)	0.914 (.023)
Carpiodes velifer	[5] 0.110 (.009)	[9] 0.095 (.131)	0.180	0.144	0.094	0.930 (.003)	0.945 (.025)	0.938 (.006)
Hypentelium nigricans	[5] 0.007 (.011)	[11] 0.006 (.011)	0.071	0.054	0.017	0.995 (.004)	0.993 (.008)	0.990 (.003)
Pantosteus discobolus	[5] 0.042 (.065)	[9] 0.083 (.091)	0.056	0.091	0.029	0.991 (.006)	0.976 (.001)	0.991 (.001)
Campostoma oligolepis	[8] 0.151 (.065)	[12] 0.138 (.289)	0.208	0.385	0.158	0.908 (.026)	0.872 (.125)	0.792 (.085)
Cyprinus carpio	[7] 0.001 (.001)	[12] 0.019 (.034)	0.051	0.047	0.026	0.998 (.001)	0.998 (.003)	0.990 (.002)
Hybopsis storeriana	[7] 0.294 (.033)	[14] 0.371 (.403)	0.367	0.320	0.267	0.795 (.029)	0.834 (.174)	0.817 (.010)
Notropis petersoni	[7] 0.138 (.154)	[10] 0.181 (.159)	0.487	0.355	0.355	0.824 (.044)	0.780 (.155)	0.788 (.027)
Luxilus zonatus	[6] 0.324 (.086)	[5] 0.388 (.427)	0.512	0.460	0.344	0.776 (.012)	0.786 (.208)	0.659 (.070)

TABLE 2
Performance comparison of each method for fish species novelty discovery. A small $\hat{P}_{err2}$ (also $\hat{P}_{err1}$) and large AUC value indicate better performance. Standard deviations are included in parentheses and the number of components in brackets.

20 times with random initial location parameters. The average $\hat{P}_{err2}$'s ($\approx \hat{P}_{err1}$) are reported in Table 2 along with the standard deviation in parentheses and the number of components in brackets. The error rates of KSD and single Gaussian model are obtained from [8]. As expected, one single Gaussian distribution is not sufficient to model this complex data. Its average error rate is the highest. Two EM methods outperform non-parametric KSD because of the flexibility of mixture models. They identify most of undiscovered species as outliers with high detection rate and low false alarm rate. For example, the detection rates of *Hypentelium nigricans* and *Cyprinus carpio* are higher than 0.99 and the false alarm rates are less than 0.01. It is clear that the Spatial-EM based outlier detection yields the most favorable result in terms of error rates; it outperforms the regular-EM method in three aspects: (1) It has a higher novelty detection rate than the regular-EM in 7 out of 10 species; (2) It consistently has a much smaller standard deviation, indicating stability of our approach. The regular-EM is highly dependent on initialization. For the worst three species *Carpiodes carpio*, *Hybopsis storeriana*, and *Luxilus zonatus*, the regular-EM is not statistically better than a random guess, while the proposed method produces a detection rate higher than 0.740, 0.706, and 0.676, respectively. Spatial-EM significantly improves the sensitivity on initialization of the regular EM. (3) It has a much smaller number of components than the regular EM method for all but one species. On average, the regular-EM uses 4 more components than Spatial-EM to model outliers. Usually, overly complicated models tend to over fit data resulting poor generalization performance, which explains large variances of the regular EM. Spatial-EM handles outliers very well. It yields simple models with good performance.

One-class SVM method [39] is also included for comparison. We implemented one-class ν-SVM method using R function 'ksvm' in the package 'kernlab'. The bandwidth parameter σ of Gaussian kernel and the parameter ν are selected such that the type-I error and type-II error are approximately equal in each fish species. Since ν is an upper bound on the fraction of outliers in the training data, we search ν in a small interval around the type-I errors of the methods considered. Comparing with the Spatial-EM method, one-class ν-SVM has a lower false alarm rate in 5 species out of 10. Especially, the error rate 0.156 for *Capiodes carpio* is much lower than 0.260 of the Spatial EM. However, in the *Notropis petersoni* case, the Spatial-EM yields a small error rate 0.138 comparing to 0.352 of KSVM.

For the AUC criterion, the Spatial-EM, regular EM and Kotz-EM are compared. Function roc in R package 'pROC' is implemented to compute AUC. Kotz-EM seems to be inferior to the other two methods. One of reasons is probably its slow convergence. From the experiment in Section 4.4, Kotz-EM suffers slow convergence problem that downgrades performance and increases the computation burden. However, it is better than spatial-EM for *Hybopsis storeriana* species with a larger AUC and a smaller standard deviation. Regular EM outperforms spatial EM in cases of *Carpiodes cyprinus*, *Hybopsis storeriana* and *Luxilus zonatus*, where the conclusion conflicts if the first metric is used. It seems that the performance depends more on the evaluation metric than the methods. But if we look at the standard deviations of AUC for regular EM, we should say that the results of regular EM are very unstable and unreliable. There are 5 out of 10 cases in which the standard deviation for the regular EM method excesses 0.125. Particularly, it is 0.174

in *Hybopsis storeriana* and 0.208 in *Luxilus zonatus*. That makes the performance differences of the two methods in those two cases insignificant. Kotz-EM is relatively robust and Spatial-EM is the most robust one.

6 CLUSTERING

Mixture model-based clustering is one of the most popular and successful unsupervised learning approaches. It provides a probabilistic (soft) clustering of the data in terms of the fitted posterior probabilities (T_{ji}'s) of membership of the mixture components with respect to the clusters. An outright (hard) clustering can be subsequently obtained by assigning each observation to the component to which it has the highest fitted posterior probability of belonging. That is, $\boldsymbol{x}_i$ is assigned to the cluster $\arg\max_j T_{ji}$.

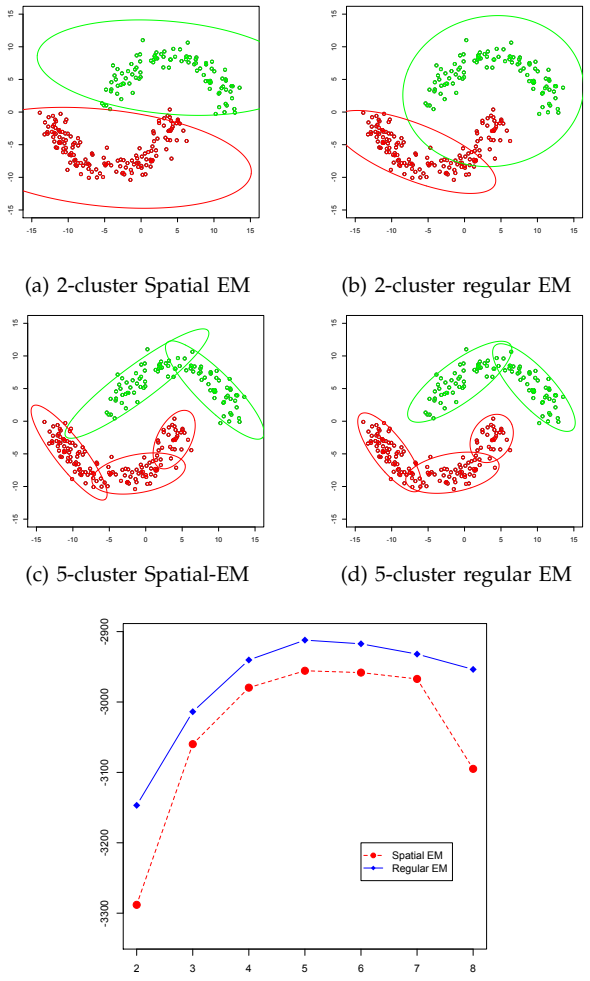
Model-based clustering approaches have a natural way to select the number of clusters based on some criteria, which have the common form of log-likelihood augmented by a model complexity penalty term. For example, Bayesian inference criterion (BIC) [14], [15], [42], the minimum message length (MML) [32], [49], the normalized entropy criterion (NEC) [6], [2] etc. have yielded good results for model choice in a range of applications. In this paper, we deal with robustness of model-based clustering. We assume that the number of clusters is known, otherwise, BIC is used. BIC is defined as twice of the log-likelihood minus $p \log N$, where the likelihood is the Gaussion based, N is the sample size and p is the number of independent parameters. For a K component mixture model, $p = K - 1 + K(d + d(d+1)/2)$ with d being the dimension.

For performance assessment, the class labels (ground truth) of training data or testing data are used to construct a confusion matrix (matching matrix). The false positive (rate) (FP/FPR), false negative (rate) (FN/FNR), true positive (TP) and true negative (TN) are computed from the confusion matrix to evaluate the accuracy of clustering methods.

We present evaluations of Spatial-EM clustering on the synthetic data and two real data sets. In the simulation experiment, we will see how the number of clusters impacts performance of the EM-based methods. Two real data sets, UCI Wisconsin diagnostic breast cancer data and yeast cell cycle data, are used for comparison of our method and some existing clustering methods.

6.1 Synthetic Data

Two samples of sizes 100 and 150 with different orientations and locations are generated from a half-moon shaped distribution. In this experiment, if we assume the number of clusters known to be 2, we will see both regular EM and Spatial-EM fail. Ellipses in Figures 3 (a-b) represent the estimated 95% probability density



(a) 2-cluster Spatial EM (b) 2-cluster regular EM

(c) 5-cluster Spatial-EM (d) 5-cluster regular EM

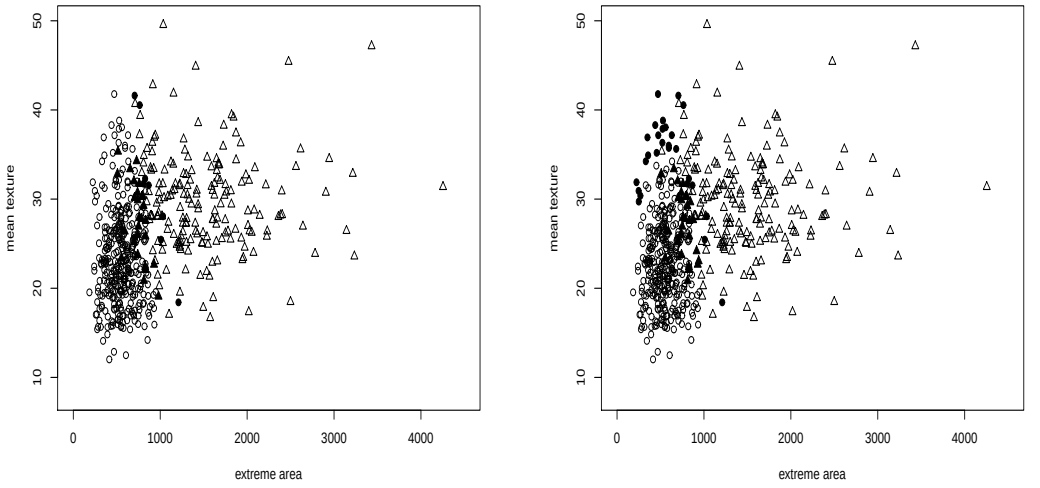
(e) BIC for choosing the number of clusters

Fig. 3. (a)-(d): Performance of 2-cluster and 5-cluster Spatial-EM and regular EM on two half-moon data. Ellipses represent represent the estimated 95% density contours of each cluster. (e): BIC criterion for choosing the number of clusters to be 5.

contours of two clusters, which do not quite follow the shape of the data cloud. Clearly, EM clustering methods based on Gaussian mixture yield clusters with density contours constrained to be elliptical. One cluster is not sufficient to model a half-moon shaped distribution, which has a concave support region.

We apply BIC criterion to choose the number of clusters. For a range of values of K (from 2 to 8), we obtain BIC for both methods, see Figure 3 (e). The optimal number of clusters is 5 with the largest BIC values in both methods. With 5 components, both methods perform much better than 2-cluster EM methods. Figures 3 (c-d) provide the density contours of 5 clusters. 2 or 3 clusters of Gaussian mixture capture the half-moon shape well. Regular EM performs better than the Spatial-EM in terms of a higher BIC value and tighter clusters. This is because

tb]



(a) Spatial-EM
FPR = 0.1320, FNR = 0.0224

(b) Regular EM
FPR = 0.1368, FNR = 0.0756

Fig. 4. A projection of the UCI Wisconsin diagnostic breast cancer data. ○ and △ represent a patient being benign and malignant respectively. Filled symbols represent misclassified observations. Our method outperforms the regular-EM one in terms of both errors.

if each component follows the Gaussian distribution, the sample mean and sample covariance matrix are most efficient estimators that makes the regular EM most statistically efficient.

6.2 UCI Wisconsin Diagnostic Breast Cancer Data

The Breast Cancer Wisconsin (Diagnostic) data set in the UCI Machine Learning Repository is available from *http://archive.ics.uci.edu/ml/datasets*. There are 569 observations from 357 patients with benign tumors and 212 patients with malignant tumors. For graphical illustration purpose, only two features *mean texture* and *extreme area* are used for clustering analysis, which is also the same setting as the experiment done by [15]. The scatter plot of these two features in Figure 4 shows a considerable overlap between benign and malignant patients.

In health care application, a malignant patient should get the most attention in the clinical practice. As usual, we define the malignant as a positive effect and benign as a negative effect. The model-based clustering from the Spatial-EM and the regular-EM algorithms yield results shown in Figure 4. Filled symbols represent misclassified observations. The Spatial-EM based clustering achieves lower error rates in both types of errors comparing with the regular EM method. More specifically, the resulting spatial-EM method has a FNR of 0.1320 slightly smaller than 0.1368 of the regular EM. The FPR of 0.0224 of Spatial EM is just around 1/3 of that of regular EM. In fact, medical screening tests that maintain a similar level of FNR but much smaller FPR can save time, money and clinic resource on the follow-up diagnostic procedures and more importantly, relieve unnecessary worries of those false positive diagnostic patients.

6.3 Yeast Cell Cycle Data

The yeast cell cycle data available from *http://faculty.washington.edu/kayee/model* contain expression levels of 384 genes over two cycles (17 time points). The expression levels peaked at different time periods correspond to the five phases of cell cycles. Group 1 has 67 genes whose expression levels reached peaks at early G1. Group 2 has 135 genes whose expression levels peaked at late G1 phase. Group 3 has 75 genes whose expression levels reach peak at S stage. Group 4 has 52 genes whose expression levels reached peaks at late G2 period. Group 5 has 55 genes whose expression levels peaked at M stage. In each group, a number of outliers are present. We expect our robust clustering results having a good approximation of this five class partition. We compare the proposed Spatial-EM clustering method with five mixture model-based clustering methods. Four of them are unsupervised regular-EM [53], X-EM [55], robust K-medoid and robust K-median method. The fifth one is supervised SCA [36]. Also, the popular supervised support vector machines (SVM) [4], the linear one as well as the kernel one, are included in the study.

X-EM estimates the mixture model parameters by maximizing a weighted likelihood. The weights are specifically designed for automatic model selection by introducing an additional parameter.

The model performance are measured based on four indices (Table 3): false positive (FP), false negative (FN), true positive (TP), true negative (TN). The total errors defined as FP+FN and error rate are shown in Table 4. The results of X-EM, Regular-EM, SCA and linear SVM are reproduced from [36], [55]. We implemented KSVM by R function 'ksvm' in

Cell division phase	Mehtods	FP	FN	TP	TN
Early G1 (67 genes)	**Spatial-EM**	20	17	50	297
	X-EM	11	24	43	306
	Reg EM	50	12	55	267
	K-medoid	26	19	48	291
	K-median	23	18	49	299
	SCA	21	21	46	296
	LSVM	38	10	57	279
	KSVM	7	9	58	310
Late G1 (135 genes)	**Spatial-EM**	32	18	117	217
	X-EM	13	54	81	236
	Reg EM	28	40	95	221
	K-medoid	39	22	113	207
	K-median	36	22	113	227
	SCA	24	35	100	225
	LSVM	43	10	125	206
	KSVM	26	5	130	223
S (75 genes)	**Spatial-EM**	13	42	33	296
	X-EM	10	47	28	299
	Reg EM	33	49	26	276
	K-medoid	37	42	33	272
	K-median	35	36	39	273
	SCA	37	36	39	272
	LSVM	72	18	57	237
	KSVM	7	26	49	302
G2 (52 genes)	**Spatial-EM**	17	17	35	315
	X-EM	13	22	30	319
	Reg EM	28	41	11	304
	K-medoid	35	36	16	297
	K-median	33	33	19	299
	SCA	18	29	23	314
	LSVM	46	5	47	286
	KSVM	6	10	42	326
M (55 genes)	**Spatial-EM**	19	7	48	310
	X-EM	12	26	29	317
	Reg EM	38	42	13	291
	K-medoid	15	33	22	314
	K-median	15	31	24	298
	SCA	19	8	47	310
	LSVM	47	2	53	282
	KSVM	8	4	51	321

TABLE 3
Performance comparison at each phase on yeast cell cycle data.

the package 'kernlab'. We used two-dimensional grid and refined grid searching for optimal parameters in order to balance between model complexity and better generalization. The optimal value for the slack parameter is chosen to be 5 and the hyper-parameter σ of Gaussian kernel is 0.075. The KSVM model contains 281 support vectors out of 384 observations. The training error rate of KSVM is 14.06%, which is much better than 26.30% of ours. However, if we look at the 3-fold cross-validation (testing) error, 25.26% of KSVM, is comparable to 24.14% of spatial EM when the same cross-validation data sets are applied. KSVM seems suffering the over-fitting problem, while our algorithm is an unsupervised method having a good generalization power.

The K-medoid method is implemented using R function 'pam' in 'cluster' package and the K-median method is implemented with 'kcca' in 'flexclust' package. Table 4 shows that the Spatial-EM outperforms all the other 6 methods in terms of the total error. The regular EM has high FPR and FNR with the poor recognition of the last two groups. We expect a poor performance of K-median and K-medoid because they ignore the covariance structure for each component, but they are still better than non-robust regular EM. K-median performs slightly better than K-medoid since it doesn't have any constraints on the center of each cluster. It is amazing to see that the two supervised learning methods that use the label information can not win against unsupervised X-EM and Spatial-EM. It can be seen that the X-EM has a relatively high FNR. This is probably because the weight scheme changes the fitted posterior probabilities and hence underestimate the covariance matrix in each component to produce a high false negative rate. The Spatial-EM correctly estimates the model parameters in the presence of outliers, hence yields the best result with a well-balanced FP and FN.

Methods	FP	FN	FP+FN	Error Rate
Spatial-EM	101	101	202	26.30%
X-EM	59	173	232	30.21%
Reg EM	177	184	361	47.01%
K-medoid	152	152	304	39.58%
K-median	142	140	282	36.72%
SCA	119	129	248	32.28%
SVM	246	45	291	37.89%
KSVM (Training)	54	54	108	14.06%

TABLE 4
Total error rate comparison on yeast cell cycle data.

7 Conclusions and Future Work

We proposed a new robust EM algorithm so called Spatial-EM for finite elliptical mixture learning. It replaces the sample mean and sample covariance with the spatial median and the modified spatial rank covariance matrix. It is robust not only to outliers but also to initial values in EM learning. Compared with many robust mixture learning procedures, the Spatial-EM has the advantages of computation ease and statistical efficiency. It has been applied to supervised learning with emphasis on outlier detection and unsupervised clustering. We adopt the outlier detection to taxonomic research on fish species novelty discovery, in which the known species are modeled as a finite mixture distribution, a new species is identified using an outlyingness function that measures distance to the underlying model (i.e., known species). The key ingredient of this application is to correctly estimate the mixture model given that the data of the known species are heterogeneous with a number of atypical observations. UCI Wisconsin diagnostic breast cancer data and yeast cell cycle data are used for clustering analysis. Our method shows superior performance

comparing with existing methods and demonstrates competitive classification power and high robustness.

The proposed method has some limitations. The first limitation is indeed for all EM methods. That is EM methods only suitable for the numerical vector type of data. They can't be directly applied to other types of data such as documents or graphs. The graph data contain pairwise similarity or dissimilarity relationships between objects, which are difficult to convert back to a vector representation of each object and hence the EM methods are not appropriate. The document file usually contains numerical, categorical and symbolic variables. As currently formulated, EM methods cannot handle categorical and symbolic features. EM methods also encounter difficulties to deal with image data directly. Because low rank structure in high dimension is a typical feature of image data, a direct application of EM methods will lead to the singularity problem of estimated covariance matrices. Hence the dimension reduction has to be done before applying EM methods.

The second limitation is the computation speed. Although our method is faster than most other robust procedures, its computational complexity is $O(n^2 + d^3)$, which may not be feasible for large-scale applications, especially in high dimensions. We will continue the work to parallelize or approximate to speed up the computation. Also other modifications of RCM definitely deserve exploration in our future work.

In the current work, we assumed the number of components known or simply applied BIC for choosing the number of clusters in unsupervised learning procedures. For supervised learning, we used the heuristic based "one-standard-error" rule to determine the number of components. In both cases, systematical and theoretical developments on the selection of robust models are needed and will be continuations of this work.

ACKNOWLEDGMENTS

Partial financial support for this research was provided by NSF under award numbers MCB-1027989 and MCB-1027830.

REFERENCES

[1] Bashir, S. and Carter, E.M. (2005). High breakdown mixture discriminant analysis. *Journal of Multivariate Analysis*, **93**(1), 102-111.

[2] Biernacki, C., Celeux, G. and Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, **20**, 267-272.

[3] Brown, B. (1983). Statistical uses of the spatial median. *Journal of the Royal Statistical Society, B*, **45**, 25-30.

[4] Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, **97**(1), 262-267.

[5] Campbell, N.A. (1984). Mixture models and atypical values. *Mathematical Geology*, **16**, 465-477.

[6] Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal*, **13**, 195-212.

[7] Chen, Y., Bart Jr, H., Dang, X. and Peng, H. (2007). Depth-based novelty detection and its application to taxonomic research. *The Seventh IEEE International Conference on Data Mining (ICDM)*, 113-122, Omaha, Nebraska.

[8] Chen, Y., Dang, X., Peng, H. and Bart Jr, H., (2009). Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 288-305.

[9] Chueng, Y. (2005). Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection. *IEEE Transactions on Knowledge and Data Engineering*, **17**(6), 750-761.

[10] Dang, X. and Serfling, R. (2010). Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Inference and Planning*, **140**, 198-213.

[11] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society, B* **39**, 1-38.

[12] Donoho, D. and Huber, P. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. Bickel, K. Doksum and J. Hodges, eds.) 157-184. Wadsworth, Belmont, CA.

[13] Figueiredo, M. and Jain, A.K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(3), 381-396.

[14] Fraley, C. and Raftery, A. (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification*, **16**, 297–306.

[15] Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

[16] Fujisawa, H. and Eguchi, S. (2006). Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference*, **136**(11), 3989-4011.

[17] Hardin, J. and Rocke, D.M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics and Data Analysis*, **44**, 625-638.

[18] Hastie, T., Tibshirani, R. and Friedman, J., (2001) *The Elements of Statistical Learning- Data Mining, Inference and Prediction*. Springer, New York.

[19] Huber, P.J. (1982), *Robust Statistics*. Wiley, New York.

[20] Kent, J.T., Tyler, D.E. and Vardi,Y. (1994). A curious likelihood identity for the multivariate t distribution. *Communication in Statistics - Simulation and Computations*, **23**, 441-453.

[21] Liu, C. and Rubin, D. B. (1995). ML estimation of the t distribution using EM and its extension. *Statistica Sinia*, **5**, 19-39.

[22] Marden, J. (1999). Some robust estimates of principal components, *Statistics & Probability Letters*, **43**, 349–359.

[23] Markaton, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, **56**, 483-486.

[24] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. Wiley.

[25] Melnykov, V. and Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, **4**, 80-116.

[26] Neykov, N., Filzmoser, P., Dimova, R. and Neytchev, P. (2007). Robust fitting of mixture using the trimmed likelihood estimator. *Computational Statistics and Data Analysis*, **52**, 299-308.

[27] McLachlan, G.J. and Krishan, T. (1997). *The EM Algorithm and Extensions*. Wiley.

[28] McLachlan, G.J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t distributions.

[29] McLachlan, G.J and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

[30] Miller, D.J. and Browning, J. (2003). A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabelled data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(11), 1468-1483.

[31] Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Springer.

[32] Oliver, J., Baxter, R. and Wallace, C. (1996). Unsupervised learning using MML. *Proceedings of 13th International Conference in Machine Learning*, 364-372.

[33] Peel, D. and McLachlan, G.J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**, 339-348.

[34] Plungpongpun, K. and Naik, D. (2008). Multivariate analysis of variance using a Kotz type distribution. In *Proceeding of the World Congress on Engineering 2008 Vol II*. WCE 2008, July 2-4 2008, London, U.K.

[35] Qin, Y. and Priebe, C.E. (2012). Maximum L_q-likelihood estimation via the expectation maximization algorithm: a robust estimation of mixture models. Submitted.

[36] Qu, Y. and Xu, S. (2004). Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. *Bioinformatics*, **20**(12), 19051913.

[37] Rao, C. R. (1988). Methodology based on the L_1-Norm in statistical inference. *Sankhya*, Series A, **50**, 289-313.

[38] Roberts, S. and Tarassenko, L. (1994). A probability resource allocating network for novelty detection. *Neural Computation*, **6**(2), 270-284.

[39] Schölkopf, B., Platt, J.C., Shawe-Taylor, J. and Smola, A. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, **13**(7), 1443-1471.

[40] Serfling, R. (2010). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. *Journal of Nonparametric Statistics*, **22**, 915–936.

[41] Shoham, S. (2002). Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. *Pattern Recognition*, **35**, 1127-1142

[42] Stanford, D. and Raftery, A.E. (2000). Principle curve clustering with noise. *IEEE Transactions on Pattern Analysis and machine Intelligence*, **22**, 601-609.

[43] Tadjudin, S. and Landgrebe, D.A. (2000). Robust parameter estimation for mixture model. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 439-445.

[44] Titterington, D.M., Smith, A.F.M. and Markov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

[45] Tyler, D. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, **15**, 234–251.

[46] Vardi, Y. and Zhang, C. (2000). The multivariate L_1-median and associated data depth. *Proceedings of National Academy of Sciences USA* 97, 1423-1436.

[47] Visuri, S., Koivunen, V., and Oja, H. (2000). Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, **91**, 557-575.

[48] Vlassis, N. and Likas, A. (2002). A greedy EM algorithm for Gaussian mixture learning. *Neural Processing Letters*, **15**, 77-87.

[49] Wallace, C. and Dowe, D. (1999). Minimum message length and Kolmogorov complexity. *The Computer Journal*, **42**(4), 270-283.

[50] Weber, A. (1909). *Theory of the Location of Industries* (translated by C. J. Friedrich from Weber's 1909 book). University of Chicago Press, 1929.

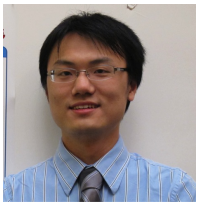
[51] Wu, C.F. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**, 95-103.

[52] Yamanishi, K., Takeuchi, J. I., Williams, G. and Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery* **8**, 275-300.

[53] Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. (2001). Model-based clustering and data transformations for gene expression data, *Bioinformatics*, **17**, 977-987.

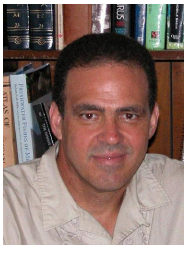
[54] Yu, K., Dang, X. and Chen, Y. (2013). Robustness of the affine equivariant scatter estimator based on the spatial rank covariance matrix. *Communications in Statistics - Theory and Method*, to appear.

[55] Zhang, Z. and Cheung, Y. (2006). On weight design of maximum weighted likelihood and an extended EM algorithm. *IEEE Transactions on Knowledge and Data Engineering*, **18**(10), 1429-1434.

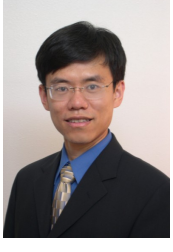


Kai Yu received the PhD degree from Department of Mathematics, University of Mississippi in 2012. Currently, he is Business Intelligence Engineer in Amazon Web Service to help building the fraud prevention and detection system using cutting-edge technology. His research interests include machine learning, data mining and pattern recognition.

Xin Dang received the PhD degree in statistics from the University of Texas at Dallas in 2005. Currently she is an assoicate professor of the department of mathematics at the University of Mississippi. Her research interests include robust and nonparametric statistics, statistical and numerical computing, and multivariate data analysis. In particular, she has focused on data depth and application, machine learning, and robust procedure computation. Dr. Dang is a member of the ASA and IMS.



Henry L. Bart Jr. is Professor of Ecology and Evolutionary Biology at Tulane University, and Director of the Tulane University Biodiversity Research Institute (TUBRI). He holds Bachelor of Science and Master of Science degrees in Biological Sciences from the University of New Orleans, and Ph.D. in Zoology from the University of Oklahoma. Prior to joining the Tulane University faculty in 1992, he held faculty positions at the University of Illinois and Auburn University. Barts research specialty is ecology and taxonomy of freshwater fishes. He is Curator of the Royal D. Suttkus Fish Collection at TUBRI - the largest research collection of post-larval fishes in the world. Bart leads a number of biodiversity informatics projects at TUBRI, including the GEOLocate georeferencing software development projects and the Fishnet2 network of fish collection databases. He is a member of a number of national boards and steering committees serving the U.S. natural history collections community. Bart teaches classes in Ichthyology, Stream Ecology, Natural Resource Conservation and Biodiversity Informatics to Tulane undergraduates and graduate students.



Yixin Chen received B.S.degree (1995) from the Department of Automation, Beijing Polytechnic University, the M.S. degree (1998) in control theory and application from Tsinghua University,and the M.S. (1999) and Ph.D. (2001) degrees in electrical engineering from the University of Wyoming. In 2003, he received the Ph.D. degree in computer science from The Pennsylvania State University. He had been an Assistant Professor of computer science at University of New Orleans. He is now an Associate Professor at the Department of Computer and Information Science, the University of Mississippi. His research interests include machine learning, data mining, computer vision, bioinformatics, and robotics and control. Dr. Chen is a member of the ACM, the IEEE, the IEEE Computer Society, the IEEE Neural Networks Society,and the IEEE Robotics and Automation Society.