Robust Finite Mixture Learning and its Application to Taxonomic Research

Kai Yu
Amazon Web Service
1918 8th Ave
Seattle, WA 98101
yukai@amazon.com

Xin Dang
Department of Mathematics
University of Mississippi
315 Hume Hall
University, 38677
xdang@olemiss.edu

Henry Bart, Jr.*
Department of Ecology and
Evolutionary Biology
Tulane University
New Orleans, LA 70118, USA
hbartjr@tulane.edu

Yixin Chen †
Department of Computer and
Information Science
University of Mississippi
University, MS 38677, USA
ychen@cs.olemiss.edu

ABSTRACT

This paper presents a new robust EM algorithm for the finite mixture learning procedures. The proposed 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 procedure. It is applied to new species identification problem in taxonomic research, in which the known species are modeled as a finite mixture distribution and an outlyingness function is used to measure distance to the underlying model. A species that lies in the region where the outlyingness is greater than a threshold is identified as a new 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. Our method shows a superior performance compared to existing methods and demonstrates excellent potential in new species discovery.

1. INTRODUCTION

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. In this paper, we tackle one of the most important and challenging research objectives in taxonomy - new species discovery.

From a machine learning perspective, new species discovery is closely related to *novelty detection* (or named *outlier detection*) [4, 5]. Novelty detection is one of the most challenging problems in data mining. When "normal" observations are given as a training data set, novelty detection can be formulated as finding observations that significantly deviate from the training data, which is essentially a one-class learning problem. There is an abundance of prior work that applied standard supervised learning techniques to outlier detection. We focus on finite mixture learning procedures.

Finite mixtures are powerful and flexible to represent arbitrarily complex probabilistic distribution of data. They have successfully been used for unsupervised learning as well as supervised learning [22, 11, 10]. In the outlier detection context, Roberts and Tarassenko [28] approximated the distribution of the training data by a Gaussian mixture model. A novelty score is defined as the maximum of the likelihood that the observation is generated by each Gaussian component. An observation is identified as novel if the score is less than a threshold. Miller and Browning [23] proposed a mixture model for a set of labeled and unlabeled samples. The mixture model includes two types of mixture components: predefined components and non-predefined components. The former generate data from known classes and assume that class labels are missing at random. The latter only generate unlabeled data, corresponding to the novelty in the unlabeled samples. Yamanishi *et al.* [37] detected outliers in an on-line process through a finite mixture model. A score is given to the datum on the learned model, with a high score indicating a high possibility of being an outlier.

Usually parameters of a mixture model are estimated by

*The author thanks support from the National Science Foundation under Grant No. MCB-1027830
†The author thanks support from the National Science Foundation under Grant No. MCB-1027989

the maximum likelihood estimate (MLE) via the expectation maximization (EM) algorithm [7, 21]. 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, weighted likelihood [19, 39], weighted trimmed likelihood [20], β-likelihood [12], L_q-likelihood [27] and likelihood of t mixtures [25, 30] are proposed to reduce the effects of outliers. 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 such as M-estimator [3, 32], MCD estimator [13] and S estimator [1]. In this paper, we propose to apply spatial rank based location and scatter estimators. They are highly robust. They are also computationally and statistically more efficient than the above robust estimators [38]. We develop a Spatial-EM algorithm for robust finite mixture learning. An outlyingness function is defined to measure deviation from the learned model. Samples lying in the region where the outlyingness is greater than a threshold are viewed as outliers (new species).

The remaining 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 formulates mixture model based novelty detection. In Section 6, we apply the proposed approach to new species discovery in taxonomy research and compare its performance to other existing methods. 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

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

for some $\boldsymbol{\mu}_j \in \mathbb{R}^d$, a positive definite symmetric $d \times d$ matrix Σ_j, and a nonnegative function h_j with

$$\int_0^\infty t^{d/2-1} h_j(t) dt < \infty$$

independent to $\boldsymbol{\mu}_j$ and Σ_j. The parameter $\boldsymbol{\mu}_j$ is the symmetric center of the j^{th} component and it equals the first moment if it exists. The scatter parameter Σ_j is proportional to the covariance matrix when it exists.

The family of mixture elliptical distributions contains a quite rich collection of models. Perhaps the most widely used one is the mixture of Gaussian distributions, in which

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

Other than that, the mixture of t distributions and Laplace distributions are commonly used in modeling data with heavy tails. In the case of 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. The degree freedom ν_j is an additional parameter that determines the fatness of the tail distribution. For $\nu_j = 1$, it is called d-variate Cauchy distribution which has very heavy tails where even the first moment doesn't exist. When $\nu_j \to \infty$, it yields the Gaussian distribution. As a generalization of multivariate mixture Laplace distribution, the mixture of Kotz type distribution [26] has the density

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

For detailed and comprehensive accounts on mixture models, see Fang and Anderson [9], McLachlan and Peel [22].

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, and 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 $\boldsymbol{x}_i$ comes from component j, zero otherwise. Then the log-likelihood of $\mathcal{Z}$ is defined by

$$L_c(\boldsymbol{\theta}|\mathcal{Z}) = \log \prod_{i=1}^{n} f((\boldsymbol{x}_i, \boldsymbol{y}_i)|\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{K} y_{ji} \log[\tau_j f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)]. \quad (2.3)$$

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, $E(Y_{ji}|\boldsymbol{\theta}^{(t)}, \boldsymbol{x}_i) = \Pr(Y_{ji} = 1|\boldsymbol{\theta}^{(t)}, \boldsymbol{x}_i)$ denoted as $T_{ji}^{(t)}$. By the Bayes rule, we have

$$T_{ji}^{(t)} = \frac{\tau_j^{(t)} f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_i^{(t)})}{\sum_{j=1}^{K} \tau_j^{(t)} f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j^{(t)})}. \quad (2.4)$$

$T_{ji}^{(t)}$'s can be interpreted as soft labels at the t^{th} iteration. Replacing y_{ji} with T_{ji} in (2.3), we have $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

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

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

For convenience, we define

$$w_{ji}^{(t)} = \frac{T_{ji}^{(t)}}{\sum_{i=1}^{n} T_{ji}^{(t)}}. \quad (2.5)$$

$w_{ji}^{(t)}$ can be viewed as a current weight of $\boldsymbol{x}_i$ contributed 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 :

$$\tau_j^{(t+1)} = \frac{\sum_{i=1}^n T_{ji}^{(t)}}{\sum_{j=1}^K \sum_{i=1}^n T_{ji}^{(t)}} = \frac{1}{n}\sum_{i=1}^n T_{ji}^{(t)}, \quad (2.6)$$

$$\boldsymbol{\mu}_j^{(t+1)} = \sum_{i=1}^n w_{ji}^{(t)} \boldsymbol{x}_i, \quad (2.7)$$

$$\Sigma_j^{(t+1)} = \sum_{i=1}^n w_{ji}^{(t)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})^T. \quad (2.8)$$

EM estimation has been proved to converge to maximum likelihood estimation (MLE) of the mixture parameters under mild conditions [7, 36, 21]. And such a 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 (2.7) and sample covariance (2.8) of each component in the M-step causes the sensitivity problem because they have the lowest possible breakdown point. Here the breakdown point is the prevailing quantitative measure of robustness proposed by Donoho and Huber [8]. 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 $1/n$ which goes to 0 when $n \to \infty$.

As an robust alternative, mixtures of t-distributions have been used for modeling data that have wider tails than Gaussian's observations [25, 30]. 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 non-linear 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 [16, 17] 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 [32] and Shoham [30]. Huber [15] 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

Let us start with the case in one dimension. 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

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

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 the center-oriented transformation is of vital importance for a rank concept in high dimensions where the natural ordering in 1D no longer exists.

Replacing the absolute value $|\cdot|$ in (3.1) by Euclidean norm of vector $\|\cdot\|$, we are ready to obtain a multivariate version of median and rank function.

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

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*, which minimizes ave$\|\boldsymbol{x} - \boldsymbol{x}_i\|$.

The spatial median, also termed as geometric median, L_1 median, has a century-long history dating back to Weber [35]. Brown [2] 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. For more comprehensive descriptions on the spatial median, refer to Small [31].

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 [29]. 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 [6] studied properties of depth-based outlier identifiers. Chen *et al.* [5] 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 6.

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 univariate counterpart, the spatial rank is not distribution-free; it characterizes the distribution of $\mathcal{X}$, especially directional information of the

distribution. To better understand their relationship, 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 [18] 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.* [34] 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 a 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.* [38] 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*

1 {*Initialization*} $\boldsymbol{\mu}_j^{(0)}$, $\Sigma_j^{(0)} = I_{d\times d}$, $\tau_j^{(0)} = 1/K$ *for all* j, $t = 0$
2 `Do Until` $\tau_j^{(t)}$*'s converge for all* j
3 `For` $j = 1$ `To` K
E-Step:
4 *Calculate* $T_{ji}^{(t)}$ *by Equations* (2.4), (2.1), (2.2)
M-Step:
5 *Update* $\tau_j^{(t+1)}$ *by Equation* (2.6)
6 *Define* $w_{ji}^{(t)}$ *as Equation* (2.5)
7 *Find* $\boldsymbol{\mu}_j^{(t+1)}$ *by Algorithm 2*
8 *Find* $(\tilde{\Sigma}_j^{(t+1)})^{-1}$ *and* $|\tilde{\Sigma}_j^{(t+1)}|^{-1/2}$ *by Algorithm 3*
9 `End`
10 $t = t + 1$
11 `End`

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

ALGORITHM 2. *Compute weighted spatial median* $\boldsymbol{\mu}_j^{(t+1)}$

1 `Input` $\{\boldsymbol{x}_i\}_{i=1}^n$, $\{w_{ji}^{(t)}\}_{i=1}^n$
2 `For` $\ell = 1$ `To` n
3 $\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_\ell) = \sum_{i=1}^n w_{ji}^{(t)}\boldsymbol{s}(\boldsymbol{x}_\ell - \boldsymbol{x}_i)$
4 `End`
5 $\boldsymbol{\mu}_j^{(t+1)} = \arg\min_{\boldsymbol{x}_\ell} \|\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_\ell)\|$
6 `Output` $\{\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_\ell)\}_{\ell=1}^n$, $\boldsymbol{\mu}_j^{(t+1)}$

ALGORITHM 3. *Compute the inverse of weighted MRCM* $\tilde{\boldsymbol{\Sigma}}_j^{(t+1)}$ *and its determinant*

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 $\boldsymbol{a}_m = \{T_{ji}^{(t)}\boldsymbol{u}_{j,m}^T(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})\}_{i=1,..,n}$
6 *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 $\quad \hat{\lambda}_{jm} = MAD\ (\{T_{ji_k}^{(t)} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_{i_k} - \boldsymbol{\mu}_j^{(t+1)})\}_{i_k})$
8 End
9 $\hat{\Lambda}_j = diag(\hat{\lambda}_{j1}^2, ..., \hat{\lambda}_{jd}^2)$
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. K-mean 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 to be noted on M-step. The first one is the way to update $\boldsymbol{\mu}_j^{(t+1)}$. Rather than using a modified Weiszfeld algorithm [33] to coordinate component weights w_{ji}, we confine our search of the solution in the pool of sample points. That is,

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

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

Secondly, 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 showed that this approaches performs very well.

5. NOVELTY DETECTION

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 novelty detection problem.

5.1 Outlyingness and Two-type Errors

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

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

as the outlyingness function to define outliers, where $\xi_j = (\boldsymbol{x} - \boldsymbol{\mu}_j)^T \Sigma_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)$ 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

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

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 model complexity. $\hat{P}_{err1}$ 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.

5.2 Estimating the Number of Components

An important issue in mixture modeling is the selecting 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 [14]. 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

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

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 mean time preserves good performance.

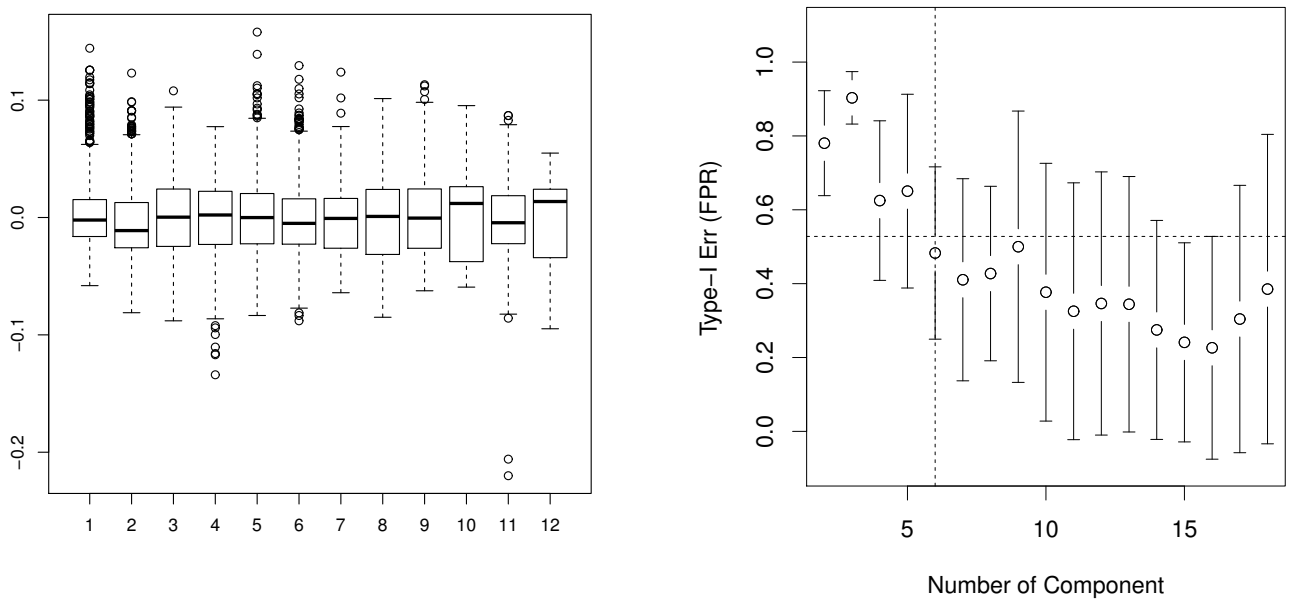
6. NEW SPECIES DISCOVERY IN TAXONOMIC RESEARCH

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*.

6.1 Data Set

The data set consists of 989 specimens from Tulane University Biodiversity Research Institute (TUBRI). 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 under 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 [4, 5] for a detailed description of the feature extraction process.

Figure 1 (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 and contains a considerable number of outliers, which calls for a robust finite mixture modeling.



(a) Box-Plot (b) One-Standard-Error Rule

Figure 1: (a) Box-Plot of data for species 2 to 10; (b) One-Standard-Error rule for choosing the number of components.

6.2 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.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 1 (b), the number of components is chosen to be 6. We then use the whole data to estimate the mixture parameters. 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 20 times with random initial location parameters. The average $\hat{P}_{err2}$'s ($\approx \hat{P}_{err1}$) are reported in Table 1 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 [5].

As expected, one single Gaussian distribution is not sufficient to model this complex data. Its average error rate is the highest among all four methods. Two EM methods outperform nonparametric 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; 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, too complicated models over-fit data with poor generalization performance, which explains large variances of the regular EM. Spatial-EM handles outliers very well. It yields simple models with good performance.

7. CONCLUSIONS AND FUTURE WORK

We proposed a new EM algorithm for finite elliptical mixture learning. It replaces the sample mean and sample covariance with the spatial median and the modified 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, it has the advantage of computation ease and statistical efficiency. It has been applied to new species identification problem in taxonomic research, 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 num-

Unknown Species	Spatial-EM	Regular-EM	KSD	Single Gaussian
Carpiodes carpio	[6] 0.260 (0.040)	[9] 0.303 (0.289)	0.234	0.408
Carpiodes cyprinus	[8] 0.181 (0.114)	[11] 0.212 (0.230)	0.209	0.245
Carpiodes velifer	[5] 0.110 (0.009)	[9] 0.095 (0.131)	0.180	0.144
Hypentelium nigricans	[5] 0.007 (0.011)	[11] 0.006 (0.011)	0.071	0.054
Pantosteus discobolus	[5] 0.042 (0.065)	[9] 0.083 (0.091)	0.056	0.091
Campostoma oligolepis	[8] 0.151 (0.065)	[12] 0.138 (0.289)	0.208	0.385
Cyprinus carpio	[7] 0.001 (0.001)	[12] 0.019 (0.034)	0.051	0.047
Hybopsis storeriana	[7] 0.294 (0.033)	[14] 0.371 (0.403)	0.367	0.320
Notropis petersoni	[7] 0.138 (0.154)	[10] 0.181 (0.159)	0.487	0.355
Luxilus zonatus	[6] 0.324 (0.086)	[5] 0.388 (0.427)	0.512	0.460

Table 1: $\hat{P}_{err2}$ (also $\hat{P}_{err1}$) of each method for fish species novelty discovery. Standard deviations are included in parentheses and the number of components in brackets.

ber of atypical observations. Our method shows a superior performance comparing with existing methods and demonstrates high potential in new species discovery.

The proposed method has some limitations. 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. Other modifications of RCM with a faster implementation definitely deserve exploration in our future work. In the current work, we used the "one-standard-error" rule to determine the number of components. It seems effective but is heuristic based. Systematical and theoretical developments on the selection of robust models are continuations of this work.

8. REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Fang, K. T., and Anderson, T. W. (1990). *Statistical Inference in Elliptically Contoured and Related Distributions*, Allerton Press, NewYork.

[10] 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.

[11] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association.* **97**, 458, 611-631.

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

[13] 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.

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

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

[16] 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.

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

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

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

[20] 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.

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

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

[23] 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/unlabled data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(11), 1468-1483.

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

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

[26] 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.

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

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

[29] Serfling, R., 2002, A depth function and a scale curve based on spatial quantiles. *Statisical Data Analysis Based on the L_1-Norm and Related Methods,* Dodge, Y., ed., 25-38

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

[31] Small, C. G. (1990) A survey of multidimensional medians. *International Statistical Review*, **58**(3), 263-277.

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

[33] 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.

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

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

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

[37] 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.

[38] 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.

[39] 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.