

# GMM Discriminant Analysis with Noisy Label for Each Class

Jian-wei Liu<sup>a,\*</sup>, Zheng-ping Ren<sup>a</sup>, Run-kun Lu<sup>a</sup>, Xiong-lin Luo<sup>a</sup>

<sup>a</sup>*Department of Automation, College of Information Science and Engineering, China University of Petroleum, Beijing, Beijing, China*

---

## Abstract

Real world datasets often contain noisy labels, and learning from such datasets using standard classification approaches may not produce the desired performance. In this paper, we propose a Gaussian Mixture Discriminant Analysis (GMDA) with noisy label for each class. We introduce flipping probability and class probability and use EM algorithms to solve the discriminant problem with label noise. We also provide the detail proofs of convergence. Experimental results on synthetic and real-world datasets show that the proposed approach notably outperforms other four state-of-art methods.

*Keywords:* Gaussian mixture models, label noise, discriminant analysis, maximum likelihood estimate

---

## 1. INTRODUCTION

Noisy label problem have been investigated for a long time in the machine learning literature and label noise-robust algorithms have numerous applications in medical image processing, spam filtering [1, 2, 3], Alzheimer disease prediction[1], gene expression classification [4], text processing [5, 6, 7, 8, 9, 10, 11], image recognition [12, 13, 14, 15, 16, 17]. Noisy labels are introduced by expert error and other unknown and unexpected factors. Mislabeled instances may lead to various potential negative consequences: bias the learning process, debase the prediction accuracy, and increase algorithm complexity of inferred models [3, 4] and the number of necessary labeling training samples, which is often produced by an expensive and time-consuming hand-annotation process or inefficient automatic annotation [1, 2], and increase difficulties in feature selection [18, 19]. The methods to deal with label noise can be classified into three [1]: 1) the label noise is ignored, and approaches that are robust to the presence of label noise, such as ensemble AdaBoost [20] and decision trees [5], are searched; 2) mislabeled instances are detected and removed, and then cleaned training samples [21, 22] are used to learn; and 3) models considering label noise are designed, and label noise-tolerant methods are determined. Label noise-tolerant methods enable researchers to take advantage of noise knowledge and use more sample information than noise-cleansing methods. The disadvantages are the increment

---

\*Corresponding author  
Email address: liujw@cup.edu.cn (Jian-wei Liu)

in algorithm complexity and the increase in the number of parameters to estimate.

Bootkrajang presented a robust normal discriminant analysis (rNDA) algorithm [23]. The algorithm solves the maximum likelihood estimate problem by employing the EM (Expectation Maximization) algorithm [24, 25, 26]. The rNDA model assumes that the examples in each class obey single Gaussian distribution; it is scarcely to verify. Thus, its performance on datasets that are not strictly Gaussian in each class seems insufficient.

Numerous studies on GMM have appeared in many fields, such as outlier mining [27], image processing [28, 29], clustering [30] and community detection[6]. [27] devised a approach to adapt to a continuously evolving outlier distribution. [31] proposed initializing mean vectors by choosing points with higher concentrations of neighbors, and using a truncated normal distribution for the preliminary estimation of dispersion matrices. DivideMix models the per-sample loss distribution with a mixture model to dynamically divide the training data into a labeled set with clean samples and an unlabeled set with noisy samples [28]. [29] addressed noisy labels issue and proposed selective negative learning and positive learning approach trained using a complementary label. [30] constructed a kernel Fisher discriminant (KFD) from training examples with noisy labels. [6] presented a procedure for community detection using GMMs that incorporates certain truncation and shrinkage effects that arise in the non-vanishing noise regime.

To solve this problem, we propose a new scheme to carry out the discriminant analysis with Gaussian mixture models (GMM) which has the ability to handle the non-Gaussian distributions. We employ a linear combination of Gaussian distributions to approximate the probabilistic distributions in each class and use the EM algorithm to solve the maximum likelihood estimate [32, 25, 26]. In the last several decades, researchers in the fields of statistics and computer vision have been interested in GMM.

The discriminant analysis discussed in this paper uses GMM to approximate data distributions and is applied to classification in the case of label noise. Maximum likelihood estimate method is used to determine the parameters. Moreover, this study derives the updating formulas of the parameters of the proposed Gaussian Mixture Discriminant Analysis (GMDA). The performance of GMDA is then compared with that of AdaBoost, rNDA, rLR, and rmLR [6] on two synthetic and six real-world datasets. Results show that our method can effectively and correctly estimate the parameters of both distribution and noise.

Our main contributions are as follows:

- 1) We propose a general discriminant analysis framework for attacking the noisy label problem. Dif-

ferent from previous approach, in this framework, the probabilistic distribution of each class on real data is captured by GMM instead of the single Gaussian distribution, this single Gaussian assumption for each class is apparently too harsh to be verified, and can scarcely reflect the actual scenarios.

2) We show that when flipping probability and class probability is introduced, the parameters of GMDA model and posterior probability to predict an unlabeled instance can be computed by using EM algorithm.

3) We provide the detail proofs of convergence for general situation, e.g., Gaussian classes with noisy labels and class-conditional Gaussian mixtures with noisy labels.

4) We have conducted extensive experiments on two synthetic datasets and six real-life datasets, which have different properties and scales, to demonstrate the effectiveness and efficiency of our proposed formulation.

The rest of this paper is organized as follows: the proposed GMMs discriminant analysis with noisy label for each class is formally introduced in Section II, convergence analysis for general situation, Gaussian classes with noisy labels and class-conditional Gaussian mixtures with noisy labels is presented in Section III, and related work in Section IV. Experimental results using synthetic and real-world datasets are discussed in Section V. Finally, the conclusion and future work are summarized in Section VI.

## 2. DISCRIMINANT ANALYSIS BASED ON GAUSSIAN MIXTURE MODELS

### 2.1. Description of the problem with the noise labels

Considering a statistical decision problem (pattern recognition, classification, and discrimination), we assume that some real data vectors

$$\mathbf{x} = (x_1, \dots, x_d) \in \mathcal{X}, \quad (\mathcal{X} = \mathcal{R}^d)$$

have to be classified with respect to a finite set of classes  $\Omega = \{\omega_1, \omega_2, \dots, \omega_K\}$ . The data vectors  $\mathbf{x} \in \mathcal{X}$  are supposed to occur randomly according to some unknown class-conditional pdfs  $p(\mathbf{x}|\omega)$  and the respective priori class probabilities  $p(\omega)$ ,  $\omega \in \Omega$ .

In case of supervised learning we are given a training set  $\mathcal{S}_\omega$  for each class  $\omega \in \Omega$ :

$$\mathcal{S}_\omega = \{\mathbf{x} \in \mathcal{X}\}, \quad \omega \in \Omega; \quad \mathcal{S} = \bigcup_{\omega \in \Omega} \mathcal{S}_\omega, \quad |\mathcal{S}| = \sum_{\omega \in \Omega} |\mathcal{S}_\omega|$$

where  $|\mathcal{S}|$  and  $|\mathcal{S}_\omega|$  denote the number of elements in set  $\mathcal{S}_\omega$ . The decision problem can be solved by means of Bayes decision function by computing the maximum-likelihood estimates of the conditional densities  $p(\mathbf{x}|\omega)$ . The related log-likelihood criterion  $L_\omega$  is given by

$$L_\omega = \frac{1}{|\mathcal{S}_\omega|} \sum_{\mathbf{x} \in \mathcal{S}_\omega} \log p(\mathbf{x}|\omega) p(\omega), \quad (\omega \in \Omega)$$

with normalization coefficient  $1/|\mathcal{S}_\omega|$  included for convenience.

In case of noisy labels we assume that the true label  $\omega \in \Omega$  of a given observation  $\mathbf{x} \in \mathcal{S}$  may be randomly interchanged (flipped, corrupted, substituted), i.e. for each data vector  $\mathbf{x} \in \mathcal{S}$  we are given a unique observed label  $\tilde{\omega} \in \Omega$ , which may differ from the true label  $\omega$ . Obviously, the probabilities  $p(\tilde{\omega})$  of the observed labels  $\tilde{\omega}$  may differ from the true probabilities  $p(\omega)$  of the true labels  $\omega$ . Assuming randomly substituted labels we denote  $p(\omega|\tilde{\omega})$  is the probability of the true label  $\omega$  given the observed label  $\tilde{\omega}$ . In this sense the conditional probability density  $\tilde{p}(\mathbf{x}|\tilde{\omega})$  of  $\mathbf{x} \in \mathcal{X}$  given an observed label  $\tilde{\omega}$  is a mixture

$$\tilde{p}(\mathbf{x}|\tilde{\omega}) = \sum_{\omega \in \Omega} p(\mathbf{x}|\omega)p(\omega|\tilde{\omega}), \mathbf{x} \in \mathcal{X}, \tilde{\omega} \in \Omega$$

consequently, with the probability  $p(\omega|\tilde{\omega})$ , any of the classes  $\omega \in \Omega$  can be the true source of the observation  $\mathbf{x} \in \mathcal{X}$ . Note that the probabilities  $p(\omega|\tilde{\omega})$ ,  $\omega \in \Omega$ ,  $p(\tilde{\omega})$ ,  $\tilde{\omega} \in \Omega$  are generally unknown and have to be estimated from data.

## 2.2. Gaussian mixture model

Lawrence and Schölkopf [19] proposed a probabilistic approach to label noise, and Bootkrajang [22], [8] extended the same model to multi-class case, assuming a Gaussian density for each class. We propose discriminant analysis based on GMM where GMM is used to approximate the probabilistic distribution of each class on real data. Supposed that the data log-likelihood is:

$$\begin{aligned} L &= \log \prod_{\mathbf{x} \in \mathcal{S}} p(\mathbf{x}|\tilde{\omega}, \theta_{\tilde{\omega}}) p(\tilde{\omega}) \\ &= \sum_{\mathbf{x} \in \mathcal{S}} \log p(\mathbf{x}|\tilde{\omega}, \theta_{\tilde{\omega}}) p(\tilde{\omega}) \end{aligned}$$

under our assumption conditions, the given model may be no longer valid. Thus, a hidden variable  $\omega \in \Omega$  is introduced to address the problem. The hidden variable  $\omega$  is considered as the real class label:

$$\begin{aligned} L &= \sum_{\mathbf{x} \in \mathcal{S}} \log \sum_{\omega \in \Omega} p(\mathbf{x}, \omega|\tilde{\omega}, \theta_{\tilde{\omega}}) p(\tilde{\omega}) \\ &= \log \prod_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} p(\mathbf{x}, \omega|\tilde{\omega}, \theta_{\tilde{\omega}}) p(\tilde{\omega}) \end{aligned} \quad (1)$$

The observed labels are generated from true labels and random noise. The joint probability  $\omega$  and  $\tilde{\omega}$  can be expressed as  $p(\omega, \tilde{\omega}_n) = p(\tilde{\omega}_n|\omega) p(\omega)$ . Input  $\mathbf{x}$  is conditionally independent from observed label  $\tilde{\omega}_n$  after knowing true label  $\omega$ , because the label noise is random. Thus, we change Eq.(1) into

$$\begin{aligned} L(\Theta) &= \sum_{\mathbf{x} \in \mathcal{S}} \log \sum_{\omega \in \Omega} p(\mathbf{x}|\tilde{\omega}, \omega, \theta_{\tilde{\omega}_n}) p(\tilde{\omega}_n, \omega) \\ &= \sum_{\mathbf{x} \in \mathcal{S}} \log \sum_{\omega \in \Omega} p(\mathbf{x}|\omega, \theta_\omega) p(\tilde{\omega}|\omega) p(\omega) \\ &= \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} \log \sum_{\omega \in \Omega} p(\mathbf{x}|\omega, \theta_\omega) p(\tilde{\omega}|\omega) p(\omega) \end{aligned} \quad (2)$$

where  $I(\cdot)$  is an indicator function. Flipping probability is defined as  $\gamma_{\tilde{\omega}, \omega} \stackrel{def}{=} p(\tilde{\omega}|\omega)$ , which is similar to reference [11], indicating the probability of true label  $\omega$  flipped to the observed label  $\tilde{\omega}$ . Class

probability is defined as  $\pi_\omega \stackrel{def}{=} p(\omega)$ , and the constraint conditions are  $\sum_{\tilde{\omega} \in \Omega} \gamma_{\tilde{\omega}, \omega} = 1$  and  $\sum_{\omega \in \Omega} \pi_\omega = 1$ . The set of flipping probability and class probability are denoted as  $\Gamma = \{\gamma_{\tilde{\omega}, \omega}\}_{\tilde{\omega} \in \Omega}$  and  $\Pi = \{\pi_\omega\}_{\omega \in \Omega}$ , respectively. The Gaussian mixture density function is

$$p(\mathbf{x} | \omega, \theta_\omega) = \sum_{m \in \mathcal{M}} w_m g(\mathbf{x} | \omega, \mu_m, \Sigma_m)$$

where  $\theta_\omega = \{w_m, \mu_m, \Sigma_m\}_{m \in \mathcal{M}}$  is the parameter set of the class  $k$ , The elements are weight, mean vector, and covariance matrix of the component  $m$ , and  $M$  is the number of components. The logarithm of a sum in (2) can be rewritten as:

$$\begin{aligned} & \log \sum_{\omega \in \Omega} p(\mathbf{x} | \omega; \theta_\omega) p(\tilde{\omega} | \omega) p(\omega) \\ &= \log \sum_{\omega \in \Omega} \gamma_{\tilde{\omega}, \omega} \pi_\omega \sum_{m \in \mathcal{M}} w_m g(\mathbf{x} | \omega, \mu_m, \Sigma_m) \\ &= \log \sum_{\omega \in \Omega} q(\omega) \frac{\gamma_{\tilde{\omega}, \omega} \pi_\omega \sum_{m \in \mathcal{M}} w_m g(\mathbf{x} | \omega, \mu_m, \Sigma_m)}{q(\omega)} \end{aligned} \quad (3)$$

where  $q(\cdot)$  is an arbitrary distribution of  $\omega$ , and the constraint condition is  $\sum_{\omega \in \Omega} q(\omega) = 1$ .

According to Jensen's inequality [33], if  $Z$  is a random variable, and  $g(\cdot)$  is a concave function, then

$$g(E(Z)) \geq E(g(Z))$$

Thus the lower bound of (3) is derived. Since  $g(\cdot)$  is a concave function, and again by Jensen's inequality we have:

$$\begin{aligned} & f\left(E_{\omega \sim q(\omega)}\left(\frac{\gamma_{\tilde{\omega}, \omega} \pi_\omega p(\mathbf{x} | \omega, \theta_\omega)}{q(\omega)}\right)\right) \\ & \geq E_{\omega \sim q(\omega)}\left(f\left(\frac{\gamma_{\tilde{\omega}, \omega} \pi_\omega p(\mathbf{x} | \omega, \theta_\omega)}{q(\omega)}\right)\right) \end{aligned}$$

i.e.

$$\begin{aligned} & \log \sum_{\omega \in \Omega} q(\omega) \frac{\gamma_{\tilde{\omega}, \omega} \pi_\omega p(\mathbf{x} | \omega, \theta_\omega)}{q(\omega)} \\ & \geq \sum_{\omega \in \Omega} q(\omega) \log \frac{\gamma_{\tilde{\omega}, \omega} \pi_\omega p(\mathbf{x} | \omega, \theta_\omega)}{q(\omega)} \end{aligned}$$

A Lagrange multiplier  $\lambda_{G_1}$  is introduced to find a local optimum of the objective function subject to  $\sum_{\omega \in \Omega} q(\omega) = 1$ , and the corresponding Lagrange function is:

$$\begin{aligned} G_1 &= \sum_{\omega \in \Omega} q(\omega) [\log \gamma_{\tilde{\omega}, \omega} \pi_\omega p(\mathbf{x} | \omega, \theta_\omega) - \log q(\omega)] \\ & \quad + \lambda_{G_1} \left[1 - \sum_{\omega \in \Omega} q(\omega)\right] \end{aligned}$$

By setting the derivative w.r.t  $q(\omega)$  equals to zero, we obtain the formula as follow:

$$\begin{aligned} & \log p(\mathbf{x} | \omega, \theta_\omega) \gamma_{\tilde{\omega}, \omega} \pi_\omega - \log q(\omega) - 1 - \lambda_{G_1} = 0 \\ & \Rightarrow \log q(\omega) = \log p(\mathbf{x} | \omega, \theta_\omega) \gamma_{\tilde{\omega}, \omega} \pi_\omega - 1 - \lambda_{G_1} \\ & \Rightarrow q(\omega) = p(\mathbf{x} | \omega, \theta_\omega) \gamma_{\tilde{\omega}, \omega} \pi_\omega \cdot e^{-(1 + \lambda_{G_1})} \end{aligned} \quad (4)$$

Further, we integrate  $q(\omega)$  in the field of  $\Omega$ , and the optimal solution w.r.t  $\lambda_{G_1}$  is:

$$\begin{aligned} \sum_{\omega \in \Omega} q(\omega) &= e^{-(1+\lambda_{G_1})} \sum_{\omega \in \Omega} p(\mathbf{x}|\omega, \theta_\omega) \gamma_{\tilde{\omega}, \omega} \pi_\omega = 1 \\ \Rightarrow \lambda_{G_1} &= \log \sum_{\omega \in \Omega} p(\mathbf{x}|\omega, \theta_\omega) \gamma_{\tilde{\omega}, \omega} \pi_\omega - 1 \end{aligned}$$

$\lambda_{G_1}$  is plugged back into (4), and  $q(\omega)$  is solved as follows:

$$\begin{aligned} q(\omega) &= \frac{p(\mathbf{x}|\omega, \theta_\omega) \gamma_{\tilde{\omega}, \omega} \pi_\omega}{\sum_{\omega \in \Omega} p(\mathbf{x}|\omega, \theta_\omega) \gamma_{\tilde{\omega}, \omega} \pi_\omega} \\ &= p(\omega | x, \tilde{\omega}) \end{aligned} \tag{5}$$

Jensen's inequality is employed and the lower bound of  $\log p(\mathbf{x}|\omega, \theta_\omega) = \log \sum_{m \in \mathcal{M}} w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m})$  is derived as

$$\begin{aligned} &\log p(\mathbf{x}|\omega, \theta_\omega) \\ &= \log \sum_{m \in \mathcal{M}} w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m}) \\ &= \log \sum_{m \in \mathcal{M}} h_{\omega, m} \frac{w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m})}{h_{\omega, m}} \\ &\geq \sum_{m \in \mathcal{M}} h_{\omega, m} \log \frac{w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m})}{h_{\omega, m}} \\ &= \sum_{m \in \mathcal{M}} h_{\omega, m} [\log w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m}) - \log h_{\omega, m}] \end{aligned}$$

where  $\sum_{m \in \mathcal{M}} h_{\omega, m} = 1$ . Lagrange multiplier  $\lambda_{G_2}$  is introduced again to solve for  $h_{\omega, m}$ , the corresponding Lagrange function is as follows:

$$\begin{aligned} G_2 &= \sum_{m \in \mathcal{M}} h_{\omega, m} [\log w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m}) - \log h_{\omega, m}] \\ &\quad + \lambda_{G_2} \left( 1 - \sum_{m \in \mathcal{M}} h_{\omega, m} \right) \end{aligned}$$

By setting the derivative w.r.t  $h_{\omega, m}$  equals to zero, we obtain the formula as follow:

$$\begin{aligned} &\log w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m}) - \log h_{\omega, m} - 1 - \lambda_{G_2} = 0 \\ \Rightarrow \log h_{\omega, m} &= \log w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m}) - 1 - \lambda_{G_2} \\ \Rightarrow h_{\omega, m} &= w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m}) \cdot e^{-(1+\lambda_{G_2})} \end{aligned}$$

Furthermore, we integrate  $h_{\omega, m}$  and the optimal solution w.r.t  $\lambda_{G_2}$  is:

$$\begin{aligned} \sum_{m \in \mathcal{M}} h_{\omega, m} &= \sum_{m \in \mathcal{M}} w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m}) \cdot e^{-(1+\lambda_{G_2})} \\ \Rightarrow \lambda_{G_2} &= \log \sum_{m \in \mathcal{M}} w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m}) - 1 \end{aligned}$$

We plug  $\lambda_{G_2}$  back and solve for  $h_{\omega, m}$  as follows:

$$\begin{aligned} h_{\omega, m} &= \frac{w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m})}{\sum_{m \in \mathcal{M}} w_{\omega, m} g(\mathbf{x}|\omega, \mu_{\omega, m}, \boldsymbol{\Sigma}_{\omega, m})} \\ &= \Pr(m | \mathbf{x}, \omega) \end{aligned} \tag{6}$$

As a result, equation (5) and (6) are plugged back to derive the objective function (7):

$$\begin{aligned}
Q = & \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} \left\{ p(\omega | \mathbf{x}, \tilde{\omega}) \right. \\
& \cdot \left. \sum_{m \in \mathcal{M}} \Pr(m | \mathbf{x}, \omega) \log \frac{w_{\omega, m} g(\mathbf{x} | \omega, \mu_{\omega, m}, \Sigma_{\omega, m})}{\Pr(m | \mathbf{x}, \omega)} \right\} \\
& + \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} p(\omega | \mathbf{x}, \tilde{\omega}) \log \gamma_{\tilde{\omega}, \omega} + \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} p(\omega | \mathbf{x}, \tilde{\omega}) \log \pi_{\omega}
\end{aligned} \tag{7}$$

The parameters to be estimated are  $\Theta = \{\theta_{\omega}\}_{\omega \in \Omega}$ ,  $\Gamma = \{\gamma_{\tilde{\omega}, \omega}\}_{\tilde{\omega} \in \Omega}$ , and  $\Pi = \{\pi_{\omega}\}_{\omega \in \Omega}$ , where  $\theta_{\omega} = \{w_{\omega, m}, \mu_{\omega, m}, \Sigma_{\omega, m}\}_{m \in \mathcal{M}}$ . The EM method is employed to solve for the parameters.

**E step:**  $p(\omega | \mathbf{x}, \tilde{\omega})$  and  $\Pr(m | \mathbf{x}, \omega)$  are calculated according to equations (5) and (6), which are listed as follows:

$$p(\omega | \mathbf{x}, \tilde{\omega}) = \frac{p(\mathbf{x} | \omega, \theta_{\omega}) \gamma_{\tilde{\omega}, \omega} \pi_{\omega}}{\sum_{\omega \in \Omega} p(\mathbf{x} | \omega, \theta_{\omega}) \gamma_{\tilde{\omega}, \omega} \pi_{\omega}} \tag{8}$$

$$\Pr(m | \mathbf{x}_n, \omega_n = k) = \frac{w_{\omega, m} g(\mathbf{x} | \omega, \mu_{\omega, m}, \Sigma_{\omega, m})}{\sum_{m \in \mathcal{M}} w_{\omega, m} g(\mathbf{x} | \omega, \mu_{\omega, m}, \Sigma_{\omega, m})} \tag{9}$$

**M step:** (7) is optimized to solve for the local optimum of  $\Theta = \{\theta_{\omega}\}_{\omega \in \Omega}$ ,  $\Gamma = \{\gamma_{\tilde{\omega}, \omega}\}_{\tilde{\omega} \in \Omega}$ , and  $\Pi = \{\pi_{\omega}\}_{\omega \in \Omega}$ .

### 2.3. Updating

**Updating rule of  $\mu_{\omega, m}$ :** by setting the derivative of equation (7) w.r.t  $\mu_{\omega, m}$  equals to zero, we derive the updating rule as follow:

$$\begin{aligned}
\frac{\partial Q}{\partial \mu_{\omega, m}} = & \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} \left\{ p(\omega | \mathbf{x}, \tilde{\omega}) \right. \\
& \cdot \left. \Pr(m | \mathbf{x}, \omega) \cdot \Sigma_{\omega, m}^{-1} \cdot (\mathbf{x} - \mu_{\omega, m}) \right\} = 0 \\
\Rightarrow \mu_{\omega, m} = & \frac{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \Pr(m | \mathbf{x}, \omega) \mathbf{x}}{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \Pr(m | \mathbf{x}, \omega)}
\end{aligned} \tag{10}$$

**Updating rule of  $\Sigma_{\omega, m}$ :** by setting the derivative of equation (7) w.r.t  $\Sigma_{\omega, m}$  equals to zero, we derive the updating rule as follow:

$$\begin{aligned}
\frac{\partial Q}{\partial \Sigma_{\omega, m}} = & - \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} \left\{ p(\omega | \mathbf{x}, \tilde{\omega}) \Pr(m | \mathbf{x}, \omega) \cdot \left[ \Sigma_{\omega, m}^{-1} \right. \right. \\
& \left. \left. - \Sigma_{\omega, m}^{-1} (\mathbf{x} - \mu_{\omega, m}) (\mathbf{x} - \mu_{\omega, m})^T \Sigma_{\omega, m}^{-1} \right] \right\} = 0 \\
\Rightarrow \Sigma_{\omega, m} = & \frac{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \Pr(m | \mathbf{x}, \omega) (\mathbf{x} - \mu_{\omega, m}) (\mathbf{x} - \mu_{\omega, m})^T}{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \Pr(m | \mathbf{x}, \omega)}
\end{aligned} \tag{11}$$

**Updating rule of  $w_{\omega, m}$ :** a Lagrange multiplier  $\lambda_{w_{\omega, m}}$  is introduced to guarantee the constraint condition  $\sum_{m \in \mathcal{M}} w_{\omega, m} = 1$ , the corresponding Lagrange function is designed as follows:

$$Q_{\lambda_{w_{\omega, m}}} = Q + \lambda_{w_{\omega, m}} \left( 1 - \sum_{m \in \mathcal{M}} w_{\omega, m} \right)$$

and by setting the derivative of  $Q_{\lambda_{w_{\omega,m}}}$  w.r.t  $w_{\omega,m}$  equals to zero, we derive the formula as follow:

$$\begin{aligned}
& \frac{\partial Q_{\lambda_{w_{\omega,m}}}}{\partial w_{\omega,m}} = 0 \\
& = \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \Pr(m | \mathbf{x}, \omega) \frac{1}{w_{\omega,m}} - \lambda_{w_{\omega,m}} \\
& \Rightarrow \lambda_{w_{\omega,m}} w_{\omega,m} = \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \Pr(m | \mathbf{x}, \omega) \\
& \stackrel{\text{intergrate } w_{\omega,m}}{\Rightarrow} \lambda_{w_{\omega,m}} \sum_{m \in \mathcal{M}} w_{\omega,m} = \lambda_{w_{\omega,m}} \\
& = \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \sum_{m \in \mathcal{M}} \Pr(m | \mathbf{x}, \omega)
\end{aligned}$$

and we plug  $\lambda_{w_{\omega,m}}$  back and solve for  $w_{\omega,m}$  as follows:

$$w_{\omega,m} = \frac{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \Pr(m | \mathbf{x}, \omega)}{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \sum_{m \in \mathcal{M}} \Pr(m | \mathbf{x}, \omega)} \quad (12)$$

**Updating the rule of  $\gamma_{\tilde{\omega},\omega}$  :** a Lagrange multiplier  $\lambda_{\gamma_{\tilde{\omega},\omega}}$  is introduced to guarantee the constraint condition , and the corresponding Lagrange function is designed as follows:

$$\begin{aligned}
& \frac{\partial Q_{\gamma_{\tilde{\omega},\omega}}}{\partial \gamma_{\tilde{\omega},\omega}} = \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \frac{1}{\gamma_{\tilde{\omega},\omega}} - \lambda_{\gamma_{\tilde{\omega},\omega}} = 0 \\
& \Rightarrow \lambda_{\gamma_{\tilde{\omega},\omega}} \gamma_{\tilde{\omega},\omega} = \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \\
& \stackrel{\text{intergrate } \gamma_{\tilde{\omega},\omega}}{\Rightarrow} \lambda_{\gamma_{\tilde{\omega},\omega}} \sum_{\tilde{\omega} \in \Omega} \gamma_{\tilde{\omega},\omega} = \lambda_{\gamma_{\tilde{\omega},\omega}} \\
& = \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\tilde{\omega} \in \Omega} p(\omega | \mathbf{x}, \tilde{\omega})
\end{aligned}$$

and we plug  $\lambda_{\gamma_{\tilde{\omega},\omega}}$  back and solve for  $\gamma_{\tilde{\omega},\omega}$  as follows:

$$\gamma_{\tilde{\omega},\omega} = \frac{\sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega})}{\sum_{\mathbf{x} \in \mathcal{S}} \sum_{\tilde{\omega} \in \Omega} p(\omega | \mathbf{x}, \tilde{\omega})} \quad (13)$$

**Updating class probability:** a Lagrange multiplier  $\lambda_{\pi_{\omega}}$  is introduced to guarantee the constraint condition  $\sum_{\omega \in \Omega} \pi_{\omega} = 1$  , a Lagrange function is designed as follows:

$$Q_{\lambda_{\pi_{\omega}}} = Q + \lambda_{\pi_{\omega}} \left( 1 - \sum_{\omega \in \Omega} \pi_{\omega} \right)$$

and by setting the derivative of  $Q_{\lambda_{\pi_{\omega}}}$  w.r.t  $\pi_{\omega}$  equals to zero, we derive the formula as follow:

$$\begin{aligned}
& \frac{\partial Q_{\lambda_{\pi_{\omega}}}}{\partial \pi_{\omega}} = \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \frac{1}{\pi_{\omega}} - \lambda_{\pi_{\omega}} = 0 \\
& \Rightarrow \lambda_{\pi_{\omega}} \pi_{\omega} = \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega}) \\
& \stackrel{\text{intergrate } \pi_{\omega}}{\Rightarrow} \lambda_{\pi_{\omega}} = \sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} p(\omega | \mathbf{x}, \tilde{\omega})
\end{aligned}$$

and we plug  $\lambda_{\pi_{\omega}}$  back and solve for  $\pi_{\omega}$  as follows:

$$\pi_{\omega} = \frac{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} p(\omega | \mathbf{x}, \tilde{\omega})}{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega} = \omega) \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} p(\omega | \mathbf{x}, \tilde{\omega})} \quad (14)$$

**Predicting posterior probability:** to predict an unlabeled instance  $\mathbf{x}_u$ , the posterior probability of each class is calculated as follows:

$$\begin{aligned} p(\omega | \mathbf{x}_u) &= \frac{p(\mathbf{x}_u | \omega, \theta_\omega) \pi_\omega}{\sum_{\omega \in \Omega} p(\mathbf{x}_u | \omega, \theta_\omega) \pi_\omega} \\ &= \frac{\pi_\omega \sum_{m \in \mathcal{M}} w_{\omega, m} g(\mathbf{x} | \omega, \mu_{\omega, m}, \Sigma_{\omega, m})}{\sum_{\omega \in \Omega} \pi_\omega \sum_{m \in \mathcal{M}} w_{\omega, m} g(\mathbf{x} | \omega, \mu_{\omega, m}, \Sigma_{\omega, m})} \end{aligned}$$

The maximum class posterior probability decides the class of this unlabeled instance.

For convenience, we summarize the overall updating process in algorithm 1.

---

**Algorithm 1** GMDA

---

**input:**  $\Theta = \{\theta_\omega\}_{\omega \in \Omega}$ ,  $\Gamma = \{\gamma_{\tilde{\omega}, \omega}\}_{\tilde{\omega} \in \Omega}$ , and  $\Pi = \{\pi_\omega\}_{\omega \in \Omega}$ .

**Initialize:**  $w_{\omega, m}, \mu_{\omega, m}, \Sigma_{\omega, m}$  is obtained by employing k-means method, and the initialization value of  $\pi_\omega$  is:

$$\pi_\omega = \frac{\sum_{\tilde{\omega} \in \Omega} I(\tilde{\omega})}{|\Omega|}$$

**if** iter  $\leq$  itermax **then**

$p(\omega = k | \mathbf{x}_n, \tilde{\omega} = j)$  and  $\Pr(m | \mathbf{x}_n, \omega_n = k)$  are calculated according to (8) and (9).

$\mu_{\omega, m}, \Sigma_{\omega, m}, w_{\omega, m}, \gamma_{\tilde{\omega}, \omega}, \pi_\omega$  is updated according to (10)–(14).

**end if**

**end**

Output:  $\Theta, \Gamma$ , and  $\Pi$ .

---

### 3. CONVERGENCE ANALYSIS

#### 3.1. General Solution

The log-likelihood function for the general decision problem with randomly substituted labels is given by

$$\begin{aligned} L_0 &= \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log \tilde{p}(\mathbf{x} | \psi(\mathbf{x})) \pi(\psi(\mathbf{x})) \\ &= \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log \sum_{\omega \in \Omega} [p(\mathbf{x} | \omega) p(\omega | \psi(\mathbf{x})) \pi(\psi(\mathbf{x}))] \end{aligned} \tag{15}$$

If we denote  $\mathcal{S}_{\tilde{\omega}}$  the training set of data vectors with the observed label  $\tilde{\omega}$ :

$$\mathcal{S}_{\tilde{\omega}} = \{\mathbf{x} \in \mathcal{S} : \psi(\mathbf{x}) = \tilde{\omega}\}, \tilde{\omega} \in \Omega, \mathcal{S} = \bigcup_{\tilde{\omega} \in \Omega} \mathcal{S}_{\tilde{\omega}}, |\mathcal{S}| = \sum_{\tilde{\omega} \in \Omega} |\mathcal{S}_{\tilde{\omega}}|$$

then, using the relation  $\tilde{\omega} = \psi(\mathbf{x})$ , we can express the log-likelihood function (15) equivalently in the form

$$L_0 = \frac{1}{|\mathcal{S}|} \sum_{\tilde{\omega} \in \Omega} \sum_{\mathbf{x} \in \mathcal{S}_{\tilde{\omega}}} \left[ \log \left( \sum_{\omega \in \Omega} p(\mathbf{x} | \omega) p(\omega | \tilde{\omega}) \right) + \log \pi(\tilde{\omega}) \right]$$

and further

$$L_0 = \sum_{\tilde{\omega} \in \Omega} \frac{|\mathcal{S}_{\tilde{\omega}}|}{|\mathcal{S}|} \log \pi(\tilde{\omega}) + \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log [p(\mathbf{x}|\omega)p(\omega|\psi(\mathbf{x}))] \quad (16)$$

We recall that for any two probability distributions  $p(\omega)$ ,  $\pi(\omega)$  the well-known Kullback-Leibler information divergence is non-negative

$$I(p, \pi) = \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{\pi(\omega)} \geq 0$$

and the inequality (11) can be rewritten in the form

$$\sum_{\omega \in \Omega} p(\omega) \log \pi(\omega) \leq \sum_{\omega \in \Omega} p(\omega) \log p(\omega) \quad (17)$$

Consequently, the left-hand part of the inequality (17) is maximized by setting  $\pi(\omega) = p(\omega)$ ,  $\omega \in \Omega$ . For the same reason the first sum in (16) is maximized by the maximum likelihood estimate  $\pi(\tilde{\omega}) = |\mathcal{S}_{\tilde{\omega}}| / |\mathcal{S}|$ ,  $\tilde{\omega} \in \Omega$  in terms of relative frequencies. In the following we repeatedly make use of this practically useful consequence of the inequality (17).

The second term in  $L_0$  can be maximized by EM algorithm. If we define the conditional weights

$$\begin{aligned} q(\omega|\mathbf{x}, \psi(\mathbf{x})) &= \frac{p(\mathbf{x}|\omega)p(\omega|\psi(\mathbf{x}))}{\sum_{\omega \in \Omega} p(\mathbf{x}|\omega)p(\omega|\psi(\mathbf{x}))} \\ \omega \in \Omega, \sum_{\omega \in \Omega} q(\omega|\mathbf{x}, \psi(\mathbf{x})) &= 1, \mathbf{x} \in \mathcal{S} \end{aligned} \quad (18)$$

then the second part of the log-likelihood function (16):

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log \left[ \sum_{\omega \in \Omega} p(\mathbf{x}|\omega) p(\omega|\psi(\mathbf{x})) \right] \quad (19)$$

can be expanded in the form [46, 47]:

$$\begin{aligned} L &= \sum_{\omega \in \Omega} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \{ q(\omega|\mathbf{x}, \psi(\mathbf{x})) \log [p(\mathbf{x}|\omega) p(\omega|\psi(\mathbf{x}))] \\ &\quad - q(\omega|\mathbf{x}, \psi(\mathbf{x})) \log q(\omega|\mathbf{x}, \psi(\mathbf{x})) \} \end{aligned} \quad (20)$$

Similarly, denoting by apostrophe the mixture components and mixture parameters in the next iteration of EM algorithm, we can write

$$\begin{aligned} L' &= \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log \left[ \sum_{\omega \in \Omega} p'(\mathbf{x}|\omega) p'(\omega|\psi(\mathbf{x})) \right] \\ &= \sum_{\omega \in \Omega} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \{ q(\omega|\mathbf{x}, \psi(\mathbf{x})) \log [p'(\mathbf{x}|\omega) p'(\omega|\psi(\mathbf{x}))] \\ &\quad - q(\omega|\mathbf{x}, \psi(\mathbf{x})) \log q'(\omega|\mathbf{x}, \psi(\mathbf{x})) \} \end{aligned} \quad (21)$$

where

$$\begin{aligned} q'(\omega|\mathbf{x}, \psi(\mathbf{x})) &= \frac{p'(\mathbf{x}|\omega)p'(\omega|\psi(\mathbf{x}))}{\sum_{\omega \in \Omega} p'(\mathbf{x}|\omega)p'(\omega|\psi(\mathbf{x}))}, \\ \omega \in \Omega, \mathbf{x} \in \mathcal{S} \end{aligned} \quad (22)$$

Now, the increment of EM algorithm in one iteration can be expressed as follows

$$\begin{aligned}
& L' - L \\
&= \sum_{\omega \in \Omega} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log \left[ \frac{p'(\mathbf{x} | \omega) p'(\omega | \psi(\mathbf{x}))}{p(\mathbf{x} | \omega) p(\omega | \psi(\mathbf{x}))} \right] \\
&+ \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log \frac{q(\omega | \mathbf{x}, \psi(\mathbf{x}))}{q'(\omega | \mathbf{x}, \psi(\mathbf{x}))}
\end{aligned} \tag{23}$$

The last sum in Eq. (23) is again the well-known non-negative Kullback-Leibler information divergence:

$$I(q, q') = \sum_{\omega \in \Omega} q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log \frac{q(\omega | \mathbf{x}, \psi(\mathbf{x}))}{q'(\omega | \mathbf{x}, \psi(\mathbf{x}))} \geq 0 \tag{24}$$

and therefore, we can write

$$\begin{aligned}
& L' - L \\
&\geq \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log \left[ \frac{p'(\mathbf{x} | \omega) p'(\omega | \psi(\mathbf{x}))}{p(\mathbf{x} | \omega) p(\omega | \psi(\mathbf{x}))} \right]
\end{aligned} \tag{25}$$

Thus, for the sake of the monotonic property of EM algorithm, we have to guarantee the inequality

$$\begin{aligned}
& L' - L \\
&\geq \sum_{\omega \in \Omega} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \left[ q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log \frac{p'(\mathbf{x} | \omega)}{p(\mathbf{x} | \omega)} \right] \\
&+ \sum_{\omega \in \Omega} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \left[ q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log \frac{p'(\omega | \psi(\mathbf{x}))}{p(\omega | \psi(\mathbf{x}))} \right] \geq 0
\end{aligned} \tag{26}$$

Here the sum over  $x \in \mathcal{S}$  in the second term can be decomposed into summing over  $\mathbf{x} \in \mathcal{S}_{\tilde{\omega}}$ :

$$\begin{aligned}
& L' - L \\
&= \sum_{\omega \in \Omega} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \left[ q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log \frac{p'(\mathbf{x} | \omega)}{p(\mathbf{x} | \omega)} \right] \\
&+ \sum_{\tilde{\omega} \in \Omega} \frac{|\mathcal{S}_{\tilde{\omega}}|}{|\mathcal{S}|} \sum_{\omega \in \Omega} \left[ \frac{1}{|\mathcal{S}_{\tilde{\omega}}|} \sum_{\mathbf{x} \in \mathcal{S}_{\tilde{\omega}}} q(\omega | \mathbf{x}, \tilde{\omega}) \right] \log \frac{p'(\omega | \tilde{\omega})}{p(\omega | \tilde{\omega})}
\end{aligned} \tag{27}$$

Now, if we define the EM iteration equations in the form

$$p'(\omega | \tilde{\omega}) = \frac{1}{|\mathcal{S}_{\tilde{\omega}}|} \sum_{\mathbf{x} \in \mathcal{S}_{\tilde{\omega}}} q(\omega | \mathbf{x}, \tilde{\omega}), \quad \omega \in \Omega, \tilde{\omega} \in \Omega \tag{28}$$

$$\begin{aligned}
& p'(\cdot | \omega) = \arg \max_{p(\cdot | \omega)} \left\{ \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log p(\mathbf{x} | \omega) \right\} \\
& \omega \in \Omega
\end{aligned} \tag{29}$$

then the monotonic property of EM algorithm is guaranteed because, by substitution (28), the second term in (26) is nonnegative as a sum of nonnegative Kullback-Leibler information divergences:

$$\sum_{\omega \in \Omega} p'(\omega | \tilde{\omega}) \log \frac{p'(\omega | \tilde{\omega})}{p(\omega | \tilde{\omega})} \geq 0, \tilde{\omega} \in \Omega \tag{30}$$

and the Eq. (29) implies the inequalities

$$\begin{aligned}
& \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log p'(\mathbf{x} | \omega) \\
&\geq \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(\omega | \mathbf{x}, \psi(\mathbf{x})) \log p(\mathbf{x} | \omega), \omega \in \Omega
\end{aligned}$$

which can be rewritten in the form:

$$\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(\omega|\mathbf{x}, \psi(\mathbf{x})) \log \frac{p'(\mathbf{x}|\omega)}{p(\mathbf{x}|\omega)} \geq 0, \omega \in \Omega \quad (31)$$

Consequently, the first term in (27) is nonnegative and in view of the above inequalities (30), (31), the EM iteration equations in the general form (23), (28) and (29) imply the basic monotonic property of EM algorithm [25, 26].

### 3.2. Gaussian Classes with Noisy Labels

Assuming a particular type, e.g. Gaussian class-conditional densities

$$p(\mathbf{x}|\omega) = f(\mathbf{x}|\mu_\omega, \Sigma_\omega), \omega \in \Omega \quad (32)$$

we can write Eq. (29) in a more specific form

$$\begin{aligned} \{\mu'_\omega, \Sigma'_\omega\} = \\ \arg \max_{\{\mu_\omega, \Sigma_\omega\}} \left[ \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(\omega|\mathbf{x}, \psi(\mathbf{x})) \log f(\mathbf{x}|\mu_\omega, \Sigma_\omega) \right] \\ \omega \in \Omega \end{aligned} \quad (33)$$

As the maximized expression in Eq. (33) is a weighted likelihood function, we can easily verify [26] that the explicit solution can be expressed as a weighted analogy of the standard maximum likelihood estimate. In particular, let  $F(\mathbf{x}|\mu)$  be a probability density with a parameter  $\mu$  having a standard maximum likelihood estimate  $\hat{\mu}$  :

$$L_\mu = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log f(\mathbf{x}|\mu) \rightarrow \max \Rightarrow \hat{\mu} = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \mathbf{x} \quad (34)$$

If  $N(\mathbf{x})$  is the number of repeated occurrences of  $\mathbf{x} \in \mathcal{X}$  in  $\mathcal{S}$  and  $q(\mathbf{x})$  denotes the relative frequency of  $\mathbf{x} \in \mathcal{S}$ :

$$q(\mathbf{x}) = \frac{N(\mathbf{x})}{|\mathcal{S}|}, \sum_{\mathbf{x} \in \mathcal{X}} q(\mathbf{x}) = 1, (\mathbf{x} \notin \mathcal{S} \Rightarrow q(\mathbf{x}) = 0)$$

then the Eq. (34) can be rewritten equivalently in the form

$$L_\mu = \sum_{\mathbf{x} \in \mathcal{X}} q(\mathbf{x}) \log F(\mathbf{x}|\mu) \rightarrow \max \Rightarrow \hat{\mu} = \sum_{\mathbf{x} \in \mathcal{X}} q(\mathbf{x}) \mathbf{x} \quad (35)$$

From the comparison of Eq. (34) and (35) it follows that the weighted likelihood (35) is maximized by the corresponding weighted maximum likelihood estimate (for a detailed proof in [47]). Consequently, in view of (33), we can write:

$$\begin{aligned} \mu'_\omega = \frac{1}{\sum_{\mathbf{x} \in \mathcal{S}} q(\omega|\mathbf{x}, \psi(\mathbf{x}))} \sum_{\mathbf{x} \in \mathcal{S}} q(\omega|\mathbf{x}, \psi(\mathbf{x})) \mathbf{x} \\ \omega \in \Omega \end{aligned} \quad (36)$$

$$\begin{aligned}\Sigma'_\omega &= \frac{1}{\sum_{\mathbf{x} \in \mathcal{S}} q(\omega|\mathbf{x}, \psi(\mathbf{x}))} \sum_{\mathbf{x} \in \mathcal{S}} q(\omega|\mathbf{x}, \psi(\mathbf{x})) \mathbf{x}\mathbf{x}^T \\ &\quad - \mu'_\omega \mu'^T_\omega, \quad \omega \in \Omega\end{aligned}\tag{37}$$

We can conclude that the problem of parameter estimation for the Gaussian classes with noisy labels can be solved by repeating the EM iteration equations (18), (28), (36) and (37).

### 3.3. Class-conditional Gaussian Mixtures with Noisy Labels

The Gaussian assumption (32) is well known to be rather restrictive and can be essentially relaxed by approximating the unknown class-conditional densities  $p(\mathbf{x}|\omega)$  by Gaussian mixtures. In particular, we assume

$$\begin{aligned}p(\mathbf{x}|\omega) &= \sum_{m \in \mathcal{M}_\omega} w_{m\omega} f(\mathbf{x}|\mu_{m\omega}, \Sigma_{m\omega}), \\ \sum_{m \in \mathcal{M}_\omega} w_{m\omega} &= 1, \quad \omega \in \Omega\end{aligned}\tag{38}$$

where  $m \in \mathcal{M}_\omega$  denotes the component's index set of the class-conditional mixture  $P(\mathbf{x}|\omega)$ . Making substitution (38) in (19) we obtain the log-likelihood function in the following more general form:

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log \left[ \sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} p(\omega|\psi(\mathbf{x})) w_{m\omega} f(\mathbf{x}|\mu_{m\omega}, \Sigma_{m\omega}) \right]$$

If we introduce the conditional component weights:

$$\begin{aligned}h(m, \omega|\mathbf{x}, \psi(\mathbf{x})) &= \frac{p(\omega|\psi(\mathbf{x})) w_{m\omega} f(\mathbf{x}|\mu_{m\omega}, \Sigma_{m\omega})}{\sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} p(\omega|\psi(\mathbf{x})) w_{m\omega} f(\mathbf{x}|\mu_{m\omega}, \Sigma_{m\omega})}, \\ m \in \mathcal{M}_\omega, \omega \in \Omega, \mathbf{x} \in \mathcal{S}\end{aligned}\tag{39}$$

then, in analogy with (20), we can expand the log-likelihood criterion (36) in the form:

$$\begin{aligned}L &= \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} h(m, \omega|\mathbf{x}, \psi(\mathbf{x})) \\ &\quad \cdot \log [p(\omega|\psi(\mathbf{x})) w_{m\omega} f(\mathbf{x}|\mu_{m\omega}, \Sigma_{m\omega})] - \\ &\quad - h(m, \omega|\mathbf{x}, \psi(\mathbf{x})) \log h(m, \omega|\mathbf{x}, \psi(\mathbf{x})) \\ \text{s.t. } &\sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} h(m, \omega|\mathbf{x}, \psi(\mathbf{x})) = 1\end{aligned}$$

Again, in analogy with (17) - (25), we come to the inequality

$$\begin{aligned}L' - L &\geq \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} h(m, \omega|\mathbf{x}, \psi(\mathbf{x})) \\ &\quad \log \left[ \frac{p'(\omega|\psi(\mathbf{x})) w_{m\omega} f(\mathbf{x}|\mu'_{m\omega}, \Sigma'_{m\omega})}{p(\omega|\psi(\mathbf{x})) w_{m\omega} f(\mathbf{x}|\mu_{m\omega}, \Sigma_{m\omega})} \right] \geq 0\end{aligned}\tag{40}$$

to be guaranteed for the sake of the monotonic property of EM algorithm. The right-hand side of (40) can be satisfied separately in two parts:

$$\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} h(m, \omega|\mathbf{x}, \psi(\mathbf{x})) \log \frac{f(\mathbf{x}|\mu'_{m\omega}, \Sigma'_{m\omega})}{f(\mathbf{x}|\mu_{m\omega}, \Sigma_{m\omega})} \geq 0\tag{41}$$

$$\sum_{\omega \in \Omega} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \sum_{m \in \mathcal{M}_\omega} h(m, \omega | \mathbf{x}, \psi(\mathbf{x})) \log \frac{p'(\omega | \psi(\mathbf{x})) w'_{m\omega}}{p(\omega | \psi(\mathbf{x})) w_{m\omega}} \geq 0 \quad (42)$$

The first inequality (41) is satisfied if we define the new parameters  $\mu'_{m\omega}, \Sigma'_{m\omega}$  by Eq. (33)

$$\begin{aligned} & \{\mu'_{m\omega}, \Sigma'_{m\omega}\} = \\ & \arg \max_{\{\mu_{m\omega}, \Sigma_{m\omega}\}} \left[ \sum_{\mathbf{x} \in \mathcal{S}} h(m, \omega | \mathbf{x}, \psi(\mathbf{x})) \log f(\mathbf{x} | \mu_{m\omega}, \Sigma_{m\omega}) \right], \\ & \omega \in \Omega \end{aligned} \quad (43)$$

Again, using the weighted likelihood analogy of the standard maximum likelihood estimate [46,47], we can write the following explicit solution of the Eq. (43) :

$$\begin{aligned} \mu'_{m\omega} &= \frac{1}{\sum_{\mathbf{x} \in \mathcal{S}} h(m, \omega | \mathbf{x}, \psi(\mathbf{x}))} \sum_{\mathbf{x} \in \mathcal{S}} h(m, \omega | \mathbf{x}, \psi(\mathbf{x})) \mathbf{x} \\ m &\in \mathcal{M}_\omega, \omega \in \Omega \end{aligned} \quad (44)$$

$$\begin{aligned} \Sigma'_{m\omega} &= \frac{1}{\sum_{\mathbf{x} \in \mathcal{S}} h(m, \omega | \mathbf{x}, \psi(\mathbf{x}))} \\ &\cdot \sum_{\mathbf{x} \in \mathcal{S}} h(m, \omega | \mathbf{x}, \psi(\mathbf{x})) \mathbf{x} \mathbf{x}^T - \mu'_{m\omega} \mu'_{m\omega}{}^T \\ m &\in \mathcal{M}_\omega, \omega \in \Omega \end{aligned} \quad (45)$$

The inequality (42) can be further decomposed as follows

$$\begin{aligned} & \sum_{\tilde{\omega} \in \Omega} \frac{|\mathcal{S}_{\tilde{\omega}}|}{|\mathcal{S}|} \sum_{\omega \in \Omega} \left[ \frac{1}{|\mathcal{S}_{\tilde{\omega}}|} \sum_{\mathbf{x} \in \mathcal{S}_{\tilde{\omega}}} \sum_{m \in \mathcal{M}_\omega} h(m, \omega | \mathbf{x}, \tilde{\omega}) \right], \\ & \log \frac{p'(\omega | \tilde{\omega})}{p(\omega | \tilde{\omega})} \geq 0 \end{aligned} \quad (46)$$

$$\sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} \left[ \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} h(m | \omega, \mathbf{x}, \psi(\mathbf{x})) \right] \log \frac{w'_{m\omega}}{w_{m\omega}} \geq 0 \quad (47)$$

where

$$h(m | \omega, \mathbf{x}, \psi(\mathbf{x})) = \frac{h(m, \omega | \mathbf{x}, \psi(\mathbf{x}))}{\sum_{m \in \mathcal{M}_\omega} h(m, \omega | \mathbf{x}, \psi(\mathbf{x}))}$$

Considering the inequality (17) we can write again the EM iteration equations for the parameters  $p'(\omega | \tilde{\omega}), w'_{m\omega}$  in explicit form. In particular, the inequality (45) is satisfied if we set

$$p'(\omega | \tilde{\omega}) = \frac{1}{|\mathcal{S}_{\tilde{\omega}}|} \sum_{\mathbf{x} \in \mathcal{S}_{\tilde{\omega}}} \sum_{m \in \mathcal{M}_\omega} h(m, \omega | \mathbf{x}, \tilde{\omega}), \omega \in \Omega, \tilde{\omega} \in \Omega \quad (48)$$

and the second inequality (47) is satisfied if we define the new component weights  $w'_{m\omega}$  by equation:

$$\begin{aligned} w'_{m\omega} &= \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} h(m | \omega, \mathbf{x}, \psi(\mathbf{x})) \\ &= \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \frac{h(m, \omega | \mathbf{x}, \psi(\mathbf{x}))}{\sum_{m \in \mathcal{M}_\omega} h(m, \omega | \mathbf{x}, \psi(\mathbf{x}))}, \\ m &\in \mathcal{M}_\omega, \omega \in \Omega \end{aligned} \quad (49)$$

The EM algorithm for the problem of estimating class-conditional Gaussian mixtures with noisy labels can be summarized in terms of the iteration equations (39), (44), (46), (48) and (49).

#### 4. RELATED WORK

The problem of discriminant analysis has been studied by researchers from many disciplines, such as physical, biological and social sciences, cognitive science, psychology, engineering, and medicine [34]. Recently, the discriminant analysis with label noise has gained substantial research attention. Various solution strategies have been proposed to prevent a learning algorithm from overfitting the noisy data, the robust classifiers with capability to diminish the effect of label noise to a certain extent have obtained varying levels of success. [24, 25, 26, 27, 28, 29, 30, 31, 32][35, 9, 36, 10, 37, 38, 11, 39, 40]. For instance, in [38], the emphasis functions that combine both sample errors and their proximity to the classification border are explored. Long and Servedio demonstrated in [11] that for a broad class of convex potential functions, any boosting algorithm was highly susceptible to random classification noise. They also emphasized that the result was unsuitable for non-convex potential function. In [39], a comprehensive empirical investigation using neural network algorithms to learn from imbalanced data with labeling errors was explored.

Lee and Liu transformed the learning problem with positive and unlabeled examples into a problem of learning with noise by labeling all unlabeled examples as negative and using logistic regression to learn from the weighting noisy examples [40]. In [41], based on consistency assurance that the label noise ultimately did not hinder the search for the optimal classifier of the noise-free sample, the study proved that any surrogate loss function could be used for classification with noisy labels by using importance reweighting. [41] also showed that the noise rate that could be estimated was upper bounded by the conditional probability of the noisy sample. Bootkrajang and Kabán built a discriminative model by modeling class noise distributions and reinterpreted existing discriminative models from the class noise perspective [8]. They proved that the error of label-noise robust logistic regression was bounded, and that label-noise robust logistic regression behaved in the same way as logistic regression when label noise did not exist or when the label flipping was symmetric. They also demonstrated that the weighting mechanism of label-noise robust logistic regression improved upon logistic regression with asymmetric label flipping. However, in [8], the loss function did not define the latent true label but defined the observed noisy label instead. Rooyen et al. proposed in [12] a convex classification calibrated unhinged loss and proved that it is robust under symmetric label noise. The loss further avoided minimization of any convex potential over a linear function class that could result in classification performance equivalent to random guessing. In [42], corruption problems that were classified as mutually contaminated distributions were considered, and authors argued that optimized balanced error on corrupted data was equivalently optimized as the binary label error on clean data.

Based on the boundary conditional class noise assumption, instead of modeling data generation or

conditional class probability both for symmetric and asymmetric cases, Jun and Cai assumed that the class noise was distributed as an unnormalized Gaussian and an unnormalized Laplace centered on the linear class boundary, and proposed Gaussian noise model and Laplace noise model, respectively [13]. They then further reinterpreted logistic regression and probit regression by using the proposed class noise probability.

Previous theoretical work on the label noise problem assumed that the two classes were separable, and the label noise was independent of the true class label or that the noise proportions for each class were known. [42, 14] introduced a general mixture proportion estimation framework for classification with label noise that eliminated these assumptions. When the class-conditional distributions overlapped and the label noise was not symmetric, [42, 14] presented assumptions ensuring identifiability and the existence of a consistent estimator of the optimal risk and given associated estimation strategies. For any arbitrary pair of contaminated distributions, a unique pair of non-contaminated distributions satisfied the proposed assumptions. Scott argued in [15] that a solution to mixture proportion estimation led to solutions to various weakly supervised learning problems, such as anomaly detection, learning from positive and unlabeled examples, domain adaptation, and classification with label noise. He established a rate of convergence for mixture proportion estimation under an appropriate distributional assumption based on surrogate risk minimization and showed that this rate of convergence can derive the consistency of the algorithm and provide a practical implementation of mixture proportion estimation and demonstrate its efficacy in classification with noisy labels [15].

By modeling the corruption process through a Markov kernel and defining the corrupted learning problem to be the corrupted experiment, Brendan and Williamson developed a general framework for tackling corrupted learning problems as well as introduced minimax upper and lower bounds on the risk for learning in the presence of corruption [43].

Manwani and Sastry studies in [44] noise tolerance under risk minimization. They assume that the actual training set given to the learning algorithm was obtained from the noise-free data set, the class label of each example is corrupted and that a learning method was noise tolerant if the classifiers learned with noise-free data and with noisy data, and both have the same classification accuracy on the noise-free data. They showed that risk minimization under 0-1 loss function was a promising approach for learning from noisy training data, and that Fisher linear discriminant and linear least squares under squared error loss were noise tolerant under uniform noise, but not under non-uniform noise. The risk minimization under other loss functions was not noise tolerant [16].

A great deal of research has been conducted on both theory and applications for such label noise

problem. Despite much attention paid to discriminant analysis for noisy data [17], the investigation focused on the instances of generating a single Gaussian model. Furthermore, symmetric and asymmetric label noise was introduced to describe the contaminated distribution of corrupted binary labels. However, the instances that belonged to the same class usually were ruled by multiple GMM because of the presence of non-Gaussian distribution, which is mixed proportionally by Gaussian distribution of different means and variances. However, to the best of our knowledge, the discriminant analysis with noisy labels based on GMM has received limited research attention mainly because of mathematical difficulties. In particular, the commonly used approaches, such as matrix analysis, are no longer directly applicable to deal with both symmetric and asymmetric label noise problem because the presence of asymmetric label noise cannot be expressed in the normal form. In this paper, therefore, we intend to tackle such an important yet challenging problem. [45] presents a novel deep self-learning framework, which does not rely on any assumption on the distribution of the noisy labels, and train a robust network on the real noisy datasets without extra supervision.

Similar to our approach, Bouveyron also proposed to use the explicit global mixture model of more than two classes [46], however, Bouveyron’s method is totally different from our approach. Bouveyron’s approach compare the supervised information given by the learning data with an unsupervised modelling based on the Gaussian mixture model, if some learning data have wrong labels, the comparison of the supervised information with an unsupervised modelling of the data allows to detect the inconsistent labels. Then it is possible afterward to build a supervised classifier by giving a low confidence to the learning observations with inconsistent labels.

## 5. EXPERIMENTS AND DISCUSSION

### 5.1. Datasets and Preprocessing

Synthetic datasets and real-world datasets are used in our experiments. Table I presents a summary of the datasets. Two synthetic datasets are created randomly by our Matlab code. We apply the following real-world datasets: Boston, Breast Issue, SPECT Heart, Waveform, Wine, and Iris. Real-world datasets are UCI datasets [47].

The datasets are equally divided into training data and test data. The original class labels are treated as true labels. Symmetric and asymmetric label errors are injected into the datasets artificially. As the label noise was generated randomly, the label noise rate and label error rate were not equal.

### 5.2. Results and Discussion

First, we illustrate the convergence of object function value of the proposed method. We use the synth1 dataset with 20% label errors as a representative. The objective function value is almost stable

Table 1: CHARACTERISTICS OF THE DATASETS

Dataset	Characteristics		
	Samples	Dimensionality	Classes
Synth1	2000	30	5
Synth2	1000	2	2
Breast	106	9	6
Iris	150	4	3
Wine	178	13	3
Heart	267	22	2
Boston	506	13	2
Waveform	5000	21	3

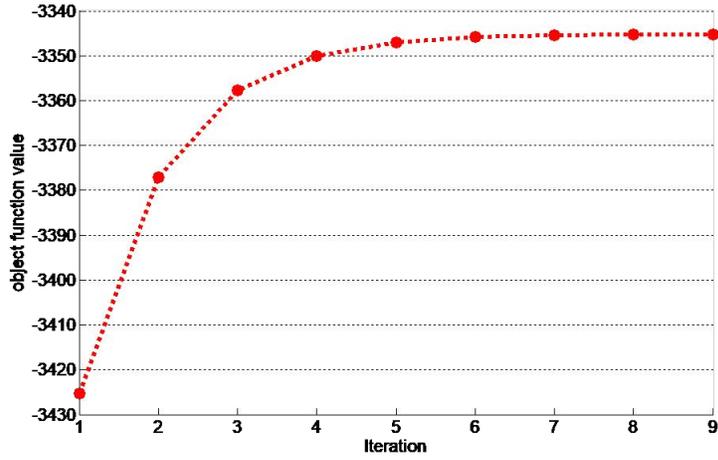


Figure 1: Convergence of the proposed method using EM algorithm

after 5 iterations in Fig. 1. As shown in Fig. 2, the error rate of prediction after 5 iterations is 12.80% and 12.80% after 10 iterations. The error rate changes rapidly during the first 4 iterations, and the error rate stabilizes after 5 iterations. The proposed method using the EM approach is convergent and effective. The change of the mixture models centers at different iterations is illustrated in Fig. 3. We

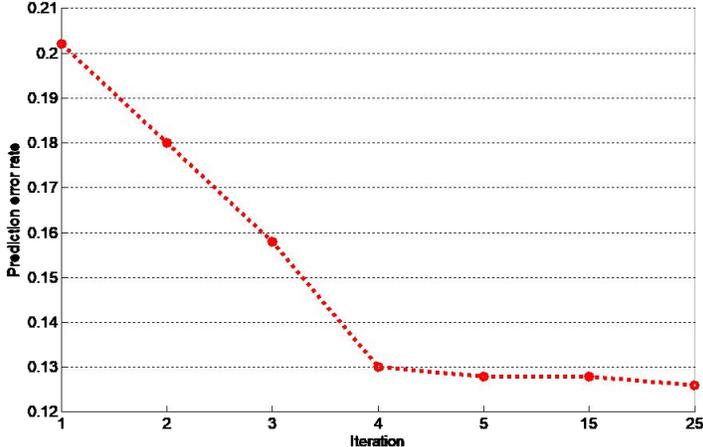


Figure 2: Error rates at different iterations

compare the proposed method to rNDA, AdaBoost, rLR, and rmLR. As the maximum likelihood used in our method is totally dependent on the training data, we change the training sample size from 20% to 80% to determine the effect of the size of the training data on the performance of classifiers. Fig. 4 shows the effect of the number of training samples on error rates of the predictions of the methods. All the methods mentioned are affected by the number of training samples, and all methods perform best with 50% samples. We run the following experiments with 50% samples.

Fig. 5 shows the original dataset synth2 and the estimated mixture centers obtained by our method in comparison with 20% label errors synth2 and its estimated mean vectors obtained by our method. Table 2 shows the parameters and the error rates of predictions obtained by our method using the original dataset synth2 and synth2 with 20% label errors. The diagonal elements  $\gamma_{ij}$  ( $i = j$ ) indicate the probability of labels that are not flipped. Table II shows that the unflipped probabilities of original dataset synth2 are extremely close to 1, whereas that of 20% label errors data are close to 0.8; and the real probability of each class is 0.5. The estimated class probabilities are all close to 0.5. The results confirm the reality. The error rates of predictions using the two datasets mentioned are 13.00% and 12.60%, respectively, which differ slightly. In the noisy case, prediction is better than the original case because our method has already considered flipping. Fig. 6 shows the results of two synthetic datasets and six real-world datasets using the proposed method and the other four methods mentioned. All results are presented in Tables III–V. The bold values of error in % shown in Table III and IV indicate

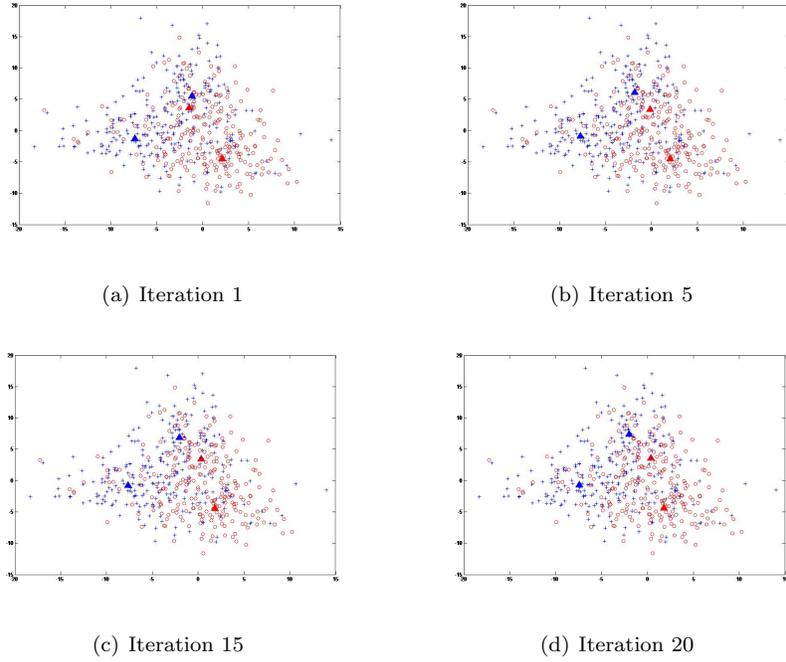


Figure 3: Model's mixture centers at different iterations

Table 2: RESULTS ON DATASET SYNTH2 USING THE PROPOSED METHOD

Dataset	Results			
	Flipping Probability		Class Probability	Error Rate
Original Synth2	1 3.03e - 04	2.49e - 05 1	[ 0.4994 0.5006 ]	13.00%
Synth2 with 20% Label Error	0.7825 0.1977	0.2175 0.8023	[ 0.5101 0.4899 ]	12.60%

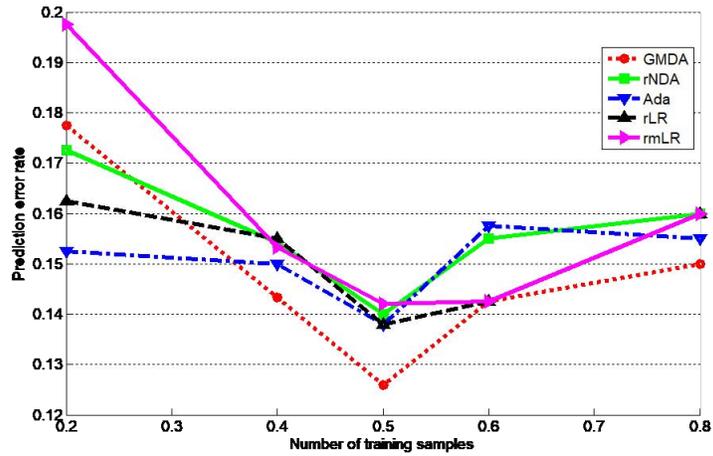


Figure 4: Effect of the number of samples

the winners. The trend is that at a higher label noise level, a higher prediction error rate is obtained. The AdaBoost method, which is a label noise-robust method, is affected most by the label errors and has mostly the largest error rate. The rLR and the rmLR methods have similar performance. The proposed method has a much better performance and the lowest error rate in most cases, and outperforms others mostly both in the symmetric and asymmetric noise cases. The AdaBoost, rLR, and rmLR methods cannot predict when samples of one class label are all flipped in binary cases. On

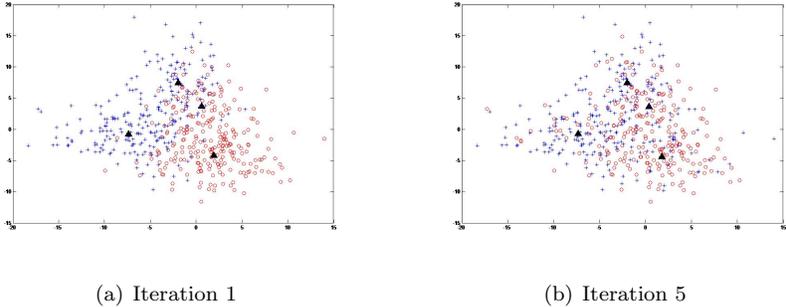
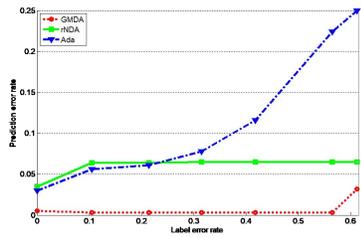


Figure 5: Dataset synth2 and estimated mixture centers obtained by using the proposed method

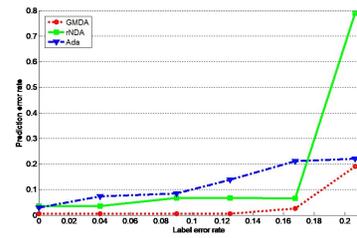
the other hand, Fig. 6 also shows that the proposed method significantly outperforms other methods on synthetic datasets synth1 with larger dimension. We run another experiment to determine the effect of the dimension of the datasets. We create a set of datasets consisting of 1000 samples, three classes, and dimension from 5 to 30. Then, 20% label errors are artificially injected into these datasets.

Fig. 7 shows the results on different datasets with different dimensions using different methods. The trend is that with a higher dimension, a lower error rate is obtained. The proposed method performs much better than the others in all dimensionality cases.

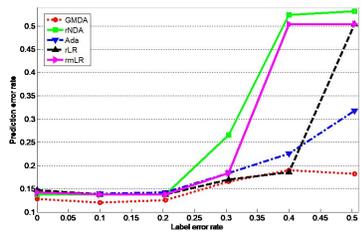
We employ a set of synthetic datasets to study the effect of the mixture number of the datasets and that of the component number of the model. Each dataset contains 1000 samples with 10 dimensions and comprises three classes, component numbers from 1 to 5 for each class. Fig. 7 shows that the more components in each class the dataset has, the larger the error rate is. In almost all cases, the proposed method notably outperforms the other methods. According to Fig. 6 and Tables III, IV, V, our method achieves better performance in a multi-class case than in a two-class case. We run one more experiment to investigate the effect of the number of classes. Datasets used are created artificially as well; they all contain 1000 samples with 10 dimensions. Fig. 8 shows that our proposed method is affected much less than the other methods. Our previous experiments on synthetic datasets have been conducted on the premise of taking the correct mixture number. In this paper, we tested the datasets using an invalid mixture number. We take a three-component dataset as a representative and set the component number of the model from 1 to 6. From Fig. 9, we can discern that an invalid mixture



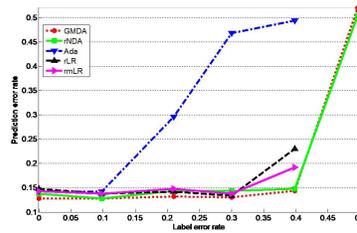
(a) Dataset synth1 with symmetric label errors



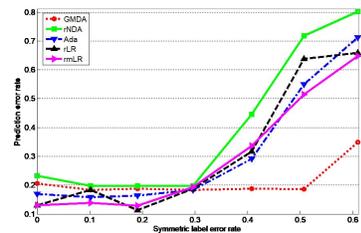
(b) Dataset synth1 with asymmetric label errors



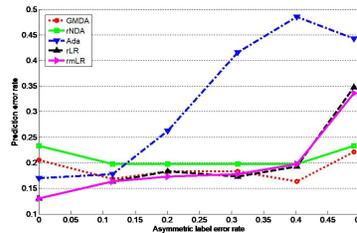
(c) Dataset synth2 with symmetric label errors



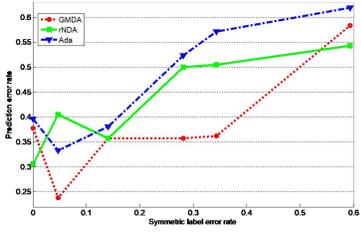
(d) Dataset synth2 with asymmetric label errors



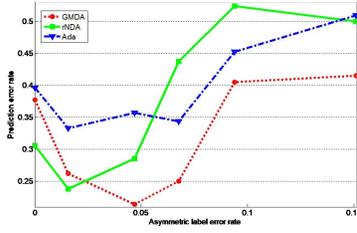
(e) Boston dataset with symmetric label errors



(f) Boston dataset with asymmetric label errors



(g) Breast Issue dataset with symmetric label errors



(h) Breast Issue dataset with asymmetric label errors

Figure 6: Results on different datasets with different noise levels using different methods

Table 3: PREDICTION ERROR RATE ON TWO SYNTHETIC DATASETS AND SIX REAL-WORLD DATASETS WITH DIFFERENT SYMMETRIC LABEL NOISE RATES

Dataset	Method	Symmetric label noise rate					
		0	0.1	0.2	0.3	0.4	0.5
Synth1	Ada	0.03	0.056	0.061	0.078	0.116	0.224
	rNDA	0.035	0.064	0.064	0.065	0.065	0.065
	GMDA	<b>0.005</b>	<b>0.003</b>	<b>0.003</b>	<b>0.003</b>	<b>0.030</b>	<b>0.003</b>
	rmLR	0.142	0.138	0.138	0.184	0.504	0.504
	rLR	0.147	0.138	0.138	0.17	0.186	0.504
Synth2	Ada	0.142	0.14	0.142	0.184	0.226	0.318
	rNDA	0.138	0.138	0.138	0.266	0.524	0.532
	GMDA	<b>0.128</b>	<b>0.120</b>	<b>0.126</b>	<b>0.166</b>	0.19	<b>0.182</b>
Breast Issue	Ada	0.396	0.333	0.381	0.523	0.571	0.619
	rNDA	<b>0.305</b>	0.404	<b>0.357</b>	0.5	0.504	0.542
	GMDA	0.377	<b>0.238</b>	<b>0.357</b>	<b>0.357</b>	<b>0.500</b>	<b>0.504</b>
Iris	Ada	0.026	0.05	0.183	0.200	0.333	0.233
	rNDA	0.026	0.033	<b>0.033</b>	0.05	0.100	<b>0.08</b>
	GMDA	<b>0.013</b>	<b>0.016</b>	<b>0.033</b>	<b>0.05</b>	<b>0.083</b>	<b>0.08</b>
Wine	Ada	0.078	0.123	0.112	0.157	0.236	0.606
	rNDA	0.044	<b>0.011</b>	<b>0.044</b>	0.044	<b>0.044</b>	0.078
	GMDA	<b>0.033</b>	0.022	<b>0.044</b>	<b>0.033</b>	0.045	<b>0.076</b>
	rmLR	0.548	0.654	0.691	0.737	0.6	0.5
	rLR	0.533	0.635	0.672	0.7	0.65	0.6
Heart	Ada	0.270	<b>0.261</b>	0.299	0.537	0.32	0.537
	rNDA	0.609	0.626	0.607	0.637	0.39	0.7
	GMDA	<b>0.248</b>	<b>0.261</b>	<b>0.261</b>	<b>0.289</b>	<b>0.287</b>	<b>0.271</b>
	rmLR	<b>0.130</b>	0.138	0.128	0.193	0.336	0.514
	rLR	<b>0.130</b>	<b>0.183</b>	<b>0.113</b>	0.188	0.316	0.638
Boston	Ada	0.17	0.158	0.163	<b>0.183</b>	0.292	0.549
	rNDA	0.233	0.198	0.198	0.198	0.445	0.717
	GMDA	0.205	<b>0.183</b>	0.188	<b>0.183</b>	<b>0.188</b>	<b>0.185</b>
Wave form	Ada	<b>0.188</b>	<b>0.201</b>	<b>0.223</b>	<b>0.223</b>	0.290	0.312
	rNDA	0.293	0.248	0.249	0.251	0.252	0.253
	GMDA	0.222	0.228	0.231	0.232	<b>0.244</b>	<b>0.244</b>

Table 4: PREDICTION ERROR RATE ON TWO SYNTHETIC DATASETS AND SIX REAL-WORLD DATASETS WITH DIFFERENT ASYMMETRIC LABEL NOISE RATES

Dataset	Method	Asymmetric label noise rate					
		0	0.1	0.2	0.3	0.4	0.5
Synth1	Ada	0.03	0.073	0.085	0.138	0.211	0.221
	rNDA	0.035	0.035	0.067	0.067	0.065	0.79
	GMDA	<b>0.005</b>	<b>0.005</b>	<b>0.005</b>	<b>0.005</b>	<b>0.026</b>	<b>0.191</b>
	rmLR	0.142	0.138	0.148	0.138	0.192	/
	rLR	0.147	0.138	0.142	0.134	0.23	/
Synth2	Ada	0.142	0.142	0.296	0.468	0.494	/
	rNDA	0.138	<b>0.128</b>	0.144	0.144	0.148	<b>0.504</b>
	GMDA	<b>0.128</b>	<b>0.128</b>	<b>0.132</b>	<b>0.13</b>	<b>0.144</b>	0.52
Breast Issue	Ada	0.396	0.333	0.357	0.343	0.452	0.509
	rNDA	<b>0.305</b>	<b>0.238</b>	0.285	0.437	0.523	0.5
	GMDA	0.377	0.261	<b>0.214</b>	<b>0.25</b>	<b>0.404</b>	<b>0.415</b>
Iris	Ada	0.026	0.016	0.183	0.177	0.433	/
	rNDA	0.026	<b>0.016</b>	<b>0.016</b>	<b>0.022</b>	<b>0.033</b>	/
	GMDA	<b>0.013</b>	<b>0.016</b>	<b>0.016</b>	<b>0.022</b>	<b>0.033</b>	/
Wine	Ada	0.078	0.140	0.112	0.197	0.408	/
	rNDA	0.044	0.056	0.056	<b>0.042</b>	<b>0.056</b>	/
	GMDA	<b>0.033</b>	<b>0.042</b>	<b>0.042</b>	<b>0.042</b>	<b>0.056</b>	/
	rmLR	0.548	0.448	0.336	0.355	0.700	/
	rLR	0.533	0.420	0.411	0.336	0.672	/
Heart	Ada	0.270	0.299	0.307	0.392	0.719	/
	rNDA	0.609	0.364	0.271	0.364	0.729	/
	GMDA	<b>0.248</b>	<b>0.261</b>	<b>0.261</b>	<b>0.261</b>	<b>0.327</b>	/
	rmLR	<b>0.130</b>	<b>0.163</b>	<b>0.173</b>	<b>0.178</b>	0.198	0.336
	rLR	<b>0.130</b>	<b>0.163</b>	0.183	0.173	0.193	0.347
Boston	Ada	0.17	0.178	0.262	0.415	0.485	0.442
	rNDA	0.233	0.198	0.198	0.198	0.445	0.233
	GMDA	0.205	0.168	0.183	0.183	<b>0.163</b>	<b>0.221</b>
Wave form	Ada	<b>0.188</b>	<b>0.196</b>	0.286	0.339	0.390	/
	rNDA	0.293	0.23	0.231	0.493	0.500	/
	GMDA	0.222	0.227	<b>0.226</b>	<b>0.280</b>	<b>0.296</b>	/

Table 5: WIN / DRAW / LOSE

Dataset	Method	Win/Draw/Lose	
		Symmetric	Asymmetric
Synth1	Ada	0/0/6	0/0/6
	rNDA	0/0/6	0/0/6
	GMDA	6/0/0	6/0/0
	rmLR	0/0/6	0/0/6
	rLR	1/0/5	0/0/6
Synth2	Ada	0/0/6	0/0/6
	rNDA	0/0/6	1/1/4
	GMDA	5/0/1	4/1/1
Breast Issue	Ada	0/0/6	0/0/6
	rNDA	1/1/4	2/0/4
	GMDA	4/1/1	4/0/2
Iris	Ada	0/0/6	0/0/5
	rNDA	0/3/3	0/4/1
	GMDA	3/3/0	1/4/0
Wine	Ada	0/0/6	0/0/5
	rNDA	1/2/3	0/2/3
	GMDA	3/2/1	3/2/0
	rmLR	0/0/6	0/0/5
	rLR	0/0/6	0/0/5
Heart	Ada	0/1/5	0/0/5
	rNDA	0/0/6	0/0/5
	GMDA	5/1/0	5/0/0
	rmLR	1/1/4	1/2/3
	rLR	1/1/4	1/2/3
Boston	Ada	0/1/5	0/0/6
	rNDA	0/0/6	0/0/6
	GMDA	2/1/3	2/0/4
Wave form	Ada	4/0/2	2/0/3
	rNDA	0/0/6	0/0/5
	GMDA	2/0/4	3/0/2

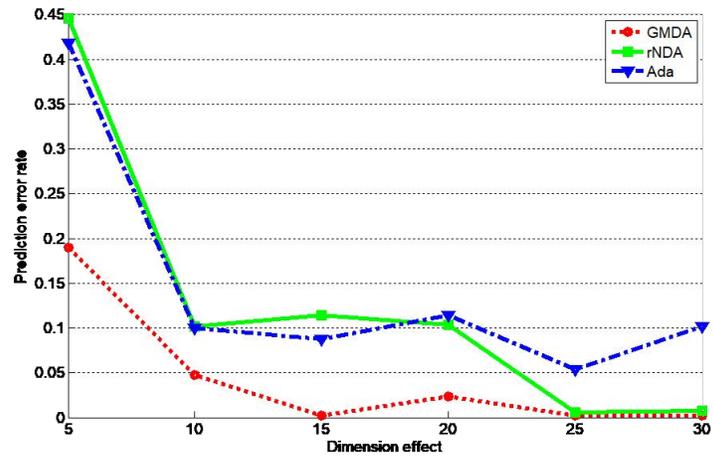


Figure 7: Effect of the dimension of datasets

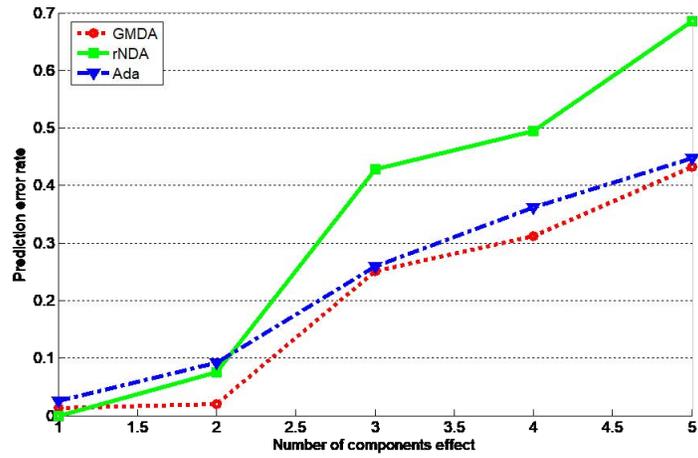


Figure 8: Effect of the number of components of the model (number of dataset mixtures is 3)

number makes an increment on error rate, whereas rNDA and AdaBoost have error rates of 0.5040 and 0.4660, respectively. Our method still performs remarkably better even when an invalid mixture number is used. Poor performance is expected when the number of the component of the number is 1.

The EM method we employed finds the local optimum each time; the initial cluster center is significant. The initial values used in this paper are obtained by k-means method, which is very sensitive to initial cluster centers. A bad initial cluster center leads to poor cluster performance and affects the error rate of the proposed method. Moreover, the difference between samples in the same class is much smaller than that between classes; thus, obtaining proper initial values for our method is difficult. The mixture model proposed in this paper is finite; the number of mixture components is provided in advance and cannot be changed to adapt to a simpler or more complicated situation. Thus, the estimation of the number of components and adapted mixture number remain to be studied.

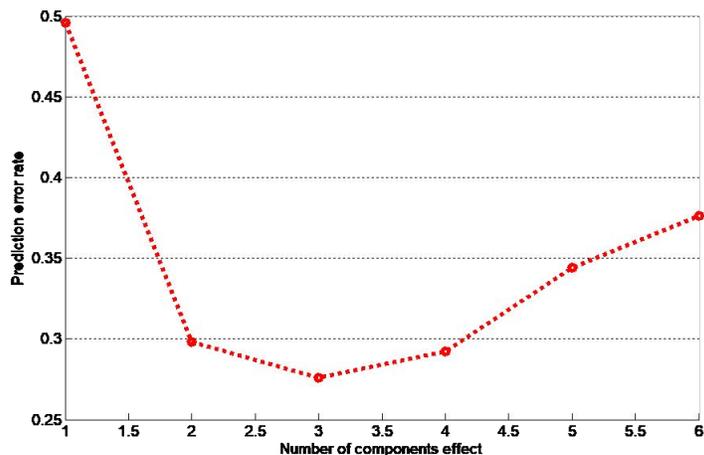


Figure 9: Effect of the number of classes

## 6. EXPERIMENTAL RESULTS ON LARGE SCALE DATA SETS

To verify the performance of our proposed approach on large scale data sets, we employ six synthetic datasets to study the effect of the mixture number of the datasets and that of the component number of the model. Each dataset contains 15000 samples with 200 dimensions and comprises two classes, component numbers from 1 to 5 for each class. The 6 datasets together with their sizes  $N$  and number of features  $D$  are listed in TABLE VI. More specifically, first, we randomly generate six synthetic datasets for verification goal. The data set is a mixture of two types of labels, with the covariance range from 0 to 250, which means that the correlation between these two types of labels is from

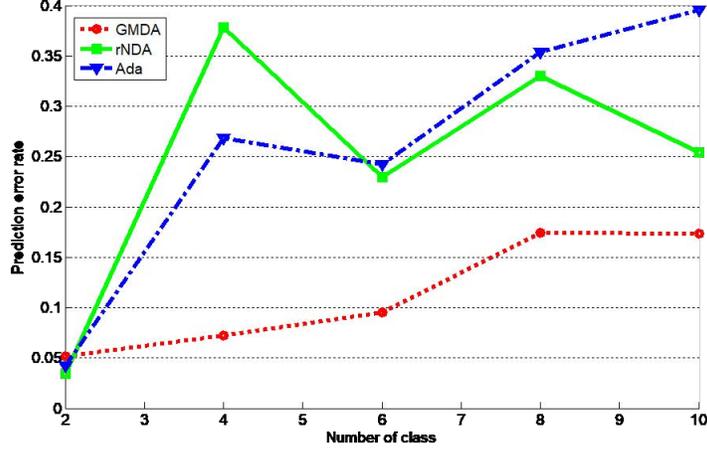


Figure 10: Effect of the number of classes

Table 6: CHARACTERISTICS OF THE LARGE SCALE DATASETS

Dataset	Characteristics					
	N	D	covariance	ross-validation	Number of samples in class 1	Number of samples in lass 2
Synth1	15000	200	250	5	7592	7408
Synth2	15000	200	0	5	6619	8381
Synth3	15000	200	250	10	8434	6566
Synth4	15000	200	0	10	6619	8381
Synth5	15000	200	250	3	7707	7293
Synth6	15000	200	0	3	7796	7204

uncorrelation to the maximum correlation. Similarly, for the division of data sets according to the TABLE VI, we have carried out 10 fold, 5 fold and 3 fold cross-validation respectively. In addition, we add noise labels based on 0%, 10%, 20%, 30%, 40% and 50% of the total number of labels. TABLE VII shows the error rate with 5, 10, and 3-cross-validation for all comparison methods. In the TABLE VII, (a1), (b1), and (c1) summarize the error rates on synthetic datasets with label correlation group, (a2), (b2), and (c2) summarize the error rates on synthetic datasets with label uncorrelation group. Fig. 11 shows the learning performances for all comparison methods, similarly, in the Fig. 11, (a1), (b1), and (c1) are the learning performances on synthetic datasets with label correlation group, (a2), (b2), and (c2) are the learning performances on synthetic datasets with label uncorrelation group.

From TABLE VII and Fig. 11, we can see that:

(a) The number of these two types of tags in synthetic data sets is comparable to each other. They belong to the synthetic data sets with relatively balanced class size and large sample size, it has certain

Table 7: EXPERIMENTAL RESULTS IN ERROR RATE ON SIX DIFFERENT CORRELATED SYNTHETIC DATASETS WITH DIFFERENT CROSS-VALIDATION PROCESSES

(a) error rate with 5-cross-validation on correlated Synth1 dataset						
Method	label noise rate					
	0.0	0.1	0.2	0.3	0.4	0.5
GMDA	0.5023	0.5023	0.5023	<b>0.5023</b>	<b>0.5023</b>	0.4977
rNDA	0.5063	0.5063	0.5063	0.5063	0.5063	<b>0.4937</b>
Ada	0.4953	<b>0.4853</b>	0.4997	0.5007	0.5090	0.5047
rLR	0.4927	0.4907	0.5057	0.5017	0.5110	0.5087
rmLR	<b>0.4890</b>	0.4877	<b>0.4953</b>	0.5080	0.5080	0.5090

(b) error rate with 10-cross-validation on correlated Synth3 dataset						
Method	label noise rate					
	0.0	0.1	0.2	0.3	0.4	0.5
GMDA	0.5073	0.5073	0.4927	0.4927	0.4927	<b>0.4927</b>
rNDA	0.5067	0.5067	0.4933	0.4933	0.4933	0.4933
Ada	<b>0.4520</b>	<b>0.4520</b>	<b>0.4520</b>	<b>0.4520</b>	0.5480	0.5480
rLR	0.5033	0.5087	0.4973	0.4807	0.4867	0.4947
rmLR	0.5120	0.4987	0.4973	0.4740	<b>0.4827</b>	0.4967

(c) error rate with 3-cross-validation on correlated Synth5 dataset						
Method	label noise rate					
	0.0	0.1	0.2	0.3	0.4	0.5
GMDA	0.5060	0.5060	0.5060	0.5060	0.5060	<b>0.4938</b>
rNDA	0.4944	0.5056	0.5056	0.5056	0.5056	0.5056
Ada	<b>0.4812</b>	<b>0.4812</b>	<b>0.4812</b>	<b>0.4812</b>	0.5188	0.5188
rLR	0.4838	0.4956	0.4884	0.4818	0.4934	0.5144
rmLR	0.4846	0.4908	0.4904	0.4950	<b>0.4904</b>	0.5152

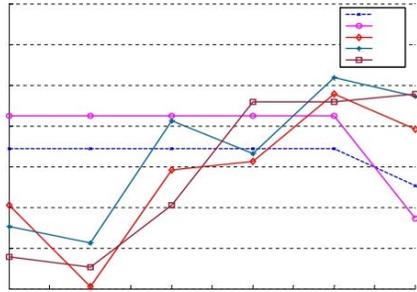
(d) error rate with 5-cross-validation on uncorrelated Synth2 dataset						
Method	label noise rate					
	0.0	0.1	0.2	0.3	0.4	0.5
GMDA	0.5570	0.5570	0.5570	0.5577	0.9640	0.5570
rNDA	0.4427	0.4427	0.4427	<b>0.4427</b>	<b>0.4427</b>	0.4427
Ada	0.4427	0.4427	0.4427	0.5573	0.5573	0.4427
rLR	<b>0.1653</b>	<b>0.1103</b>	<b>0.3853</b>	0.5573	0.9483	<b>0.1653</b>
rmLR	0.2730	0.2730	0.4303	0.5477	0.7920	0.2730

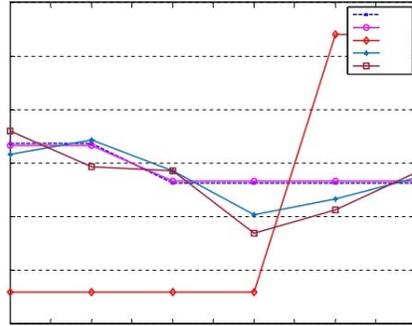
(e) error rate with 10-cross-validation on uncorrelated Synth4 dataset						
Method	label noise rate					
	0.0	0.1	0.2	0.3	0.4	0.5
GMDA	<b>0.0420</b>	0.5440	0.5453	0.5440	0.5473	0.9613
rNDA	0.4560	0.4560	0.4560	0.4560	0.4560	<b>0.4560</b>
Ada	0.4560	0.4560	0.4560	0.4560	0.5440	0.5440
rLR	0.0647	<b>0.2313</b>	<b>0.2320</b>	<b>0.2133</b>	<b>0.4267</b>	0.9353
rmLR	0.0600	0.2313	0.3453	0.3733	0.4473	0.9400

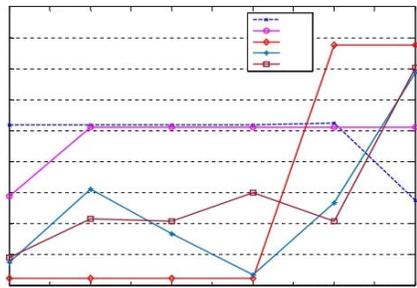
(f) error rate with 3-cross-validation on uncorrelated Synth6 dataset						
Method	label noise rate					
	0.0	0.1	0.2	0.3	0.4	0.5
GMDA	0.0424	0.5572	0.5572	0.5570	0.5673	0.9628
rNDA	0.4422	0.4422	0.4422	0.4422	0.4422	0.4422
Ada	0.4422	0.4422	0.4422	0.4422	0.5578	0.5578
rLR	0.0754	0.1754	0.1562	0.1902	0.4370	0.9264
rmLR	0.0708	0.2686	0.2376	0.3258	0.4577	0.9112



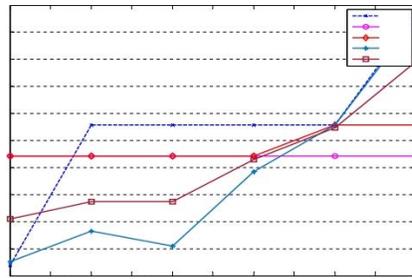
(a) correlated synth1 dataset



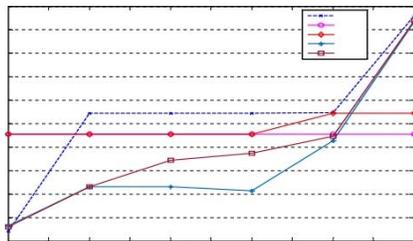
(b) correlated synth3 dataset



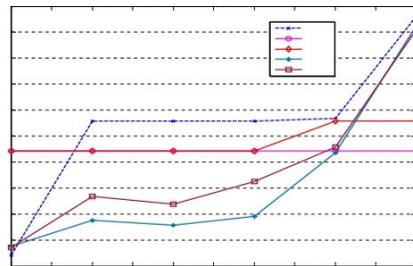
(c) correlated synth5 dataset



(d) uncorrelated synth2 dataset



(e) uncorrelated synth4 dataset



(f) uncorrelated synth6 dataset

representativeness.

(b) When the characteristics difference between the two classes is small, that is, the covariance between two classes is equal to 250. The effect of the increase of noise labels is not very obvious. Compared with other methods, the GMDA has a certain capability of noise resistance, and its error rate is up and down at 50%. Secondly, with the increase of noise tags, the error rate of the GMDA decreases by about 1%. Similar to the GMDA, under these synthetic datasets, the other comparison algorithms also fluctuate at specific values. We suspect that this is caused by the randomly generated synthetic datasets, and this small fluctuation does not affect the evaluation of the noise resistance performance of various models.

(c) When the characteristics difference between the two classes is large, that is, the covariance between two classes is equal to 0, the addition of noise tags have different degrees of impact on these algorithms. Generally speaking, the more noise tags are, the higher the classification error rate is. But when we look the label noise rate at the interval  $[0.2, 0.4]$ , these algorithms all are with good anti-noise performance and have ability of active noise cancelation. It is worthy of note that ADA algorithm is different from the trend curve of other algorithms. When the difference between the two kinds of tags is small, the fluctuation is large. When the difference between the two kinds of tags is large, the fluctuation is small. We can choose the appropriate algorithm according to the actual data set.

In sum up, when the characteristics difference between the two classes in mixture model is obvious, it is meaningful to analyze and compare the experimental results. Whereas the correlation between the two classes in mixture model is large, the effect of the increase of noise labels is vague and limited. As increase of noise labels, prediction error rate does not change much.

## 7. CONCLUSION

This paper presented a discriminant analysis based on Gaussian mixture models and applied to classification in the presence of label noise. We derived the updating formulas of the parameters. The experiments on two synthetic datasets and several real-world datasets showed that the proposed method was convergent and effective and mostly outperformed the other methods. Compared with the other methods, our method was less affected by the factors discussed in the preceding sections.

We found that the number of training samples affected the performance significantly, that is, the number of training samples is increased if necessary. If the samples were insufficient for maximum likelihood to estimate, Bayes estimation was used, where prior information was utilized, or domain adaptation learning was used, where a source dataset that was akin to a target dataset was used to

help.

The number of components of a model given in advance may not be adapted to all the classes; it might lead to further calculation on a simpler case or less approximation on a more complicated case. Therefore, we considered a more flexible and adaptable infinite mixture model that estimates the hidden number of components from the training datasets.

## 8. ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their valuable comments and suggestions.

## References

- [1] D. F. Nettleton, A. Orriols-Puig, A. Fornells, A study of the effect of different types of noise on the precision of supervised learning techniques, *Artificial intelligence review* 33 (4) (2010) 275–306.
- [2] B. Frénay, M. Verleysen, Classification in the presence of label noise: a survey, *IEEE transactions on neural networks and learning systems* 25 (5) (2013) 845–869.
- [3] C. E. Brodley, M. A. Friedl, Identifying mislabeled training data, *Journal of artificial intelligence research* 11 (1999) 131–167.
- [4] G. L. Libralon, A. C. P. de Leon Ferreira, A. C. Lorena, et al., Pre-processing for noise detection in gene expression classification data, *Journal of the Brazilian Computer Society* 15 (1) (2009) 3–11.
- [5] J. Abellán, A. R. Masegosa, Bagging decision trees on data sets with classification noise, in: *International Symposium on Foundations of Information and Knowledge Systems*, Springer, 2010, pp. 248–265.
- [6] H. Mathews, V. Mayya, A. Volfovsky, G. Reeves, Gaussian mixture models for stochastic block models with non-vanishing noise, in: *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, IEEE, 2019, pp. 699–703.
- [7] T. J. Hastie, R. Tibshirani, J. H. Friedman, *The elements of statistical learning*, 2001.
- [8] J. Bootkrajang, A. Kabán, Label-noise robust logistic regression and its applications, in: *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 2012, pp. 143–158.
- [9] M. Kearns, Efficient noise-tolerant learning from statistical queries, *Journal of the ACM (JACM)* 45 (6) (1998) 983–1006.

- [10] Y. Li, L. F. Wessels, D. de Ridder, M. J. Reinders, Classification in the presence of class noise using a probabilistic kernel fisher method, *Pattern Recognition* 40 (12) (2007) 3349–3357.
- [11] P. M. Long, R. A. Servedio, Random classification noise defeats all convex potential boosters, *Machine learning* 78 (3) (2010) 287–304.
- [12] B. Van Rooyen, A. K. Menon, R. C. Williamson, Learning with symmetric label noise: The importance of being unhinged, *arXiv preprint arXiv:1505.07634* (2015).
- [13] J. Du, Z. Cai, Modelling class noise with symmetric and asymmetric distributions, in: *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [14] C. Scott, G. Blanchard, G. Handy, Classification with asymmetric label noise: Consistency and maximal denoising, in: *Conference on learning theory*, PMLR, 2013, pp. 489–511.
- [15] C. Scott, G. Blanchard, G. Handy, Classification with asymmetric label noise: Consistency and maximal denoising, in: *Conference on learning theory*, PMLR, 2013, pp. 489–511.
- [16] N. Manwani, P. Sastry, Noise tolerance under risk minimization, *IEEE transactions on cybernetics* 43 (3) (2013) 1146–1151.
- [17] S. A. Goldman, R. H. Sloan, Can pac learning algorithms tolerate random attribute noise?, *Algorithmica* (1995).
- [18] W. Zhang, R. Rekaya, K. Bertrand, A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer, *Bioinformatics* 22 (3) (2006) 317–325.
- [19] A. A. Shanab, T. M. Khoshgoftaar, R. Wald, Robustness of threshold-based feature rankers with data sampling on noisy and imbalanced data, in: *Twenty-Fifth International FLAIRS Conference*, 2012.
- [20] G. Ratsch, Soft margins for adaboost, *Machine Learning* 42 (3) (2001) 287–320.
- [21] D. R. Wilson, T. R. Martinez, Reduction techniques for instance-based learning algorithms, *Machine learning* 38 (3) (2000) 257–286.
- [22] C. E. Brodley, M. A. Friedl, et al., Identifying and eliminating mislabeled training instances, in: *Proceedings of the National Conference on Artificial Intelligence*, Citeseer, 1996, pp. 799–805.
- [23] J. Bootkrajang, Supervised learning with random labelling errors, Ph.D. thesis, University of Birmingham (2013).

- [24] T. K. Moon, The expectation-maximization algorithm, *IEEE Signal processing magazine* 13 (6) (1996) 47–60.
- [25] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1) (1977) 1–22.
- [26] J. Grim, On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions, *Kybernetika -Praha-* 18 (3) (1982) 173–190.
- [27] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, R. Horaud, Deepgum: Learning deep robust regression with a gaussian-uniform mixture model, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 202–217.
- [28] J. Li, R. Socher, S. C. Hoi, Dividemix: Learning with noisy labels as semi-supervised learning, *arXiv preprint arXiv:2002.07394* (2020).
- [29] Y. Kim, J. Yim, J. Yun, J. Kim, Nlnl: Negative learning for noisy labels, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 101–110.
- [30] N. Lawrence, B. Schölkopf, Estimating a kernel fisher discriminant in the presence of label noise, in: *18th International Conference on Machine Learning (ICML 2001)*, Morgan Kaufmann, 2001, pp. 306–306.
- [31] V. Melnykov, I. Melnykov, Initializing the em algorithm in gaussian mixture models with an unknown number of components, *Computational Statistics & Data Analysis* 56 (6) (2012) 1381–1395.
- [32] P. Boykin, T. Mor, M. Pulver, V. Roychowdhury, F. Vatan, *Proceedings of the 40th annual symposium on foundations of computer science* (1999).
- [33] C. M. Bishop, et al., *Neural networks for pattern recognition*, Oxford university press, 1995.
- [34] S. Ji, J. Ye, Generalized linear discriminant analysis: a unified framework and efficient model selection, *IEEE Transactions on Neural Networks* 19 (10) (2008) 1768–1782.
- [35] R. Durrant, A. Kabán, Error bounds for kernel fisher linear discriminant in gaussian hilbert space, in: *Artificial Intelligence and Statistics*, PMLR, 2012, pp. 337–345.
- [36] D. Angluin, P. Laird, Learning from noisy examples, *Machine Learning* 2 (4) (1988) 343–370.
- [37] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, A. Tewari, Learning with noisy labels, *Advances in neural information processing systems* 26 (2013) 1196–1204.

- [38] V. Gómez-Verdejo, M. Ortega-Moral, J. Arenas-García, A. R. Figueiras-Vidal, Boosting by weighting critical and erroneous samples, *Neurocomputing* 69 (7-9) (2006) 679–685.
- [39] T. M. Khoshgoftaar, J. Van Hulse, A. Napolitano, Supervised neural network modeling: an empirical investigation into learning from imbalanced data with labeling errors, *IEEE Transactions on Neural Networks* 21 (5) (2010) 813–830.
- [40] W. S. Lee, B. Liu, Learning with positive and unlabeled examples using weighted logistic regression, in: *ICML*, Vol. 3, 2003, pp. 448–455.
- [41] T. Liu, D. Tao, Classification with noisy labels by importance reweighting, *IEEE Transactions on pattern analysis and machine intelligence* 38 (3) (2015) 447–461.
- [42] A. K. Menon, B. V. Rooyen, C. S. Ong, B. Williamson, Learning from corrupted binary labels via class-probability estimation (2015).
- [43] C. Scott, A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels (2015).
- [44] B. van Rooyen, R. C. Williamson, Learning in the presence of corruption, *arXiv preprint arXiv:1504.00091* (2015).
- [45] J. Han, P. Luo, X. Wang, Deep self-learning from noisy labels, 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019) 5137–5146.
- [46] C. Bouveyron, S. Girard, Robust supervised classification with mixture models: Learning from data with uncertain labels, *Pattern Recognition* 42 (11) (2009) 2649–2658.
- [47] M. Lichman, et al., *Uci machine learning repository* (2013).