# Unsupervised Selection and Discriminative Estimation of Orthogonal Gaussian Mixture Models for Handwritten Digit Recognition

Xuefeng Chen, Xiabi Liu*, and Yunde Jia
*Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology*
*{crocodel, liuxiabi, jiayunde}@bit.edu.cn*

## Abstract

*The problem of determining the appropriate number of components is important in finite mixture modeling for pattern classification. This paper considers the application of an unsupervised clustering method called AutoClass to training of Orthogonal Gaussian Mixture Models (OGMM). Actually, the number of components in OGMM of each class is selected based on AutoClass. In this way, the structures of OGMM for difference classes are not necessarily be the same as those in usual modeling scheme, so that the dissimilarity between the data distributions of different classes can be described more exactly. After the model selection is completed, a discriminative learning framework of Bayesian classifiers called Max-Min posterior pseudo-probabilities (MMP) is employed to estimate component parameters in OGMM of each class. We apply the proposed learning approach of OGMM to handwritten digit recognition. The experimental results on the MNIST database show the effectiveness of our approach.*

## 1. Introduction

Finite Mixture is a powerful statistical modeling tool in pattern classification research. They are flexible enough to approximate any given density with high accuracy. Learning finite mixture models from a given data set involves two related issues: the selection of the number of components (usually known as model selection) and the estimation of component parameters (usually known as parameter estimation).

The conventional model selection solution for pattern classification is to consider the same number of components for all the classes and decide it by classification experiments [1-2]. This strategy exhibits two drawbacks. Firstly, it is troublesome and time-consuming to conduct experiments with varying numbers of components. Secondly, the difference between the data distributions of different classes is ignored somewhat by assuming the same number of components to all the class models. Recently, unsupervised learning has been explored to tackle the problem of automatic selection of finite mixture models in the fields of pattern classification and data analysis [3-4].

The standard approach to parameter estimation of finite mixture models is Expectation-Maximization (EM) algorithm for maximum likelihood. However, the EM algorithm converges to the local minimum. Thus it is sensitive to the initial parameter values. In recent years, the researchers have resorted to discriminative learning for solving parameter estimation problem in pattern classification community. The proposed discriminative learning methods of parameter estimation include Minimum Classification Error (MCE) [5], Maximum Mutual Information [6], Max-Min Posterior Pseudo-Probabilities [7], etc. Previous works demonstrate that the discriminative learning algorithms outperform the traditional EM counterpart.

The Orthogonal Gaussian Mixtures Model (OGMM), a kind of finite mixture models, has been shown to be suitable for handwritten digit recognition [8]. In this paper, we combine unsupervised learning and discriminative learning to obtain the appropriate OGMM of digit classes for handwritten digit recognition. In the proposed method, (1) the Automatic Classification (AutoClass) [9], an unsupervised clustering algorithm with automatic determination of cluster number, is used to solve model selection. To our best knowledge, this is the first work on the

[*] Corresponding author. Tel.: +86-10-68913447; Fax: +86-10-86343 158.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.44

application of AutoClass to GMM learning in document analysis and recognition; and (2) the Max-Min posterior Pesudo-probabilities (MMP) [7], a discriminative learning framework of Bayesian classifiers, is employed to estimate component parameters. We evaluated the proposed learning approach to OGMM through handwritten digit recognition experiments on the MNIST database [10]. In the paper of Liu et al. [11], state-of-the-art techniques of handwritten digit recognition, including features and classifiers, are thoroughly investigated on MNIST database. They reported the best recognition rate of 99.58% for 8-direction gradient features (abbreviated to e-grg there), which came from SVM with RBF kernel. Using e-grg features extracted from MNIST database by courtesy of Liu, we achieved the comparable recognition rate of 99.31% on the test set. We further experientially compare our approach with the conventional strategy of model selection by experiments, as well as the EM and the MCE for parameter estimation. Compared with the model selection by experiments, the application of AutoClass brought the average 7.4% reduction in error rate on the test set. Compared with the EM and the MCE, the MMP brought the average 19.20% and 3.36% reduction in error rate on the test set, respectively. In detail, the error rate on the test set is reduced from 0.92% to 0.85% (EM), 0.77% to 0.71% (MCE), 0.74% to 0.69% (MMP), without or with AutoClass based model selection, respectively.

The rest of this paper is organized as follows. Section 2 briefly introduces the OGMM. Section 3 presents AutoClass based model selection method. Section 4 describes the parameter estimation using MMP method. Section 5 discusses the application of our approach to handwritten digit recognition and the corresponding experimental results. We give our conclusions in Section 6.

## 2. Orthogonal GMM

Gaussian Mixture Model (GMM) is by far the most commonly used finite mixtures. In the applications of GMM to statistical classification, the diagonal covariance matrices are often assumed for practical computation. Because the feature vectors are always correlated statistically in practice, using GMM with diagonal covariance matrix demands a large number of components to approximate the given density with enough accuracy. A feasible solution to this problem is to relieve the correlation among the elements in feature vectors by orthogonal transformation, in which the feature vectors are transformed to the space spanned by the eigenvectors of the covariance matrix. The corresponding GMM is called Orthogonal Gaussian Mixture Model (OGMM) [12].

The OGMM for modeling class conditional probability density function is described more formally in the following. Let $\boldsymbol{x}$ be a $D$ -dimensional feature vector, $C_i$ be the $i$ -th class, then we have

$$p(\boldsymbol{x}|C_i)=\sum_{k=1}^{K} w_k O_k\left(\boldsymbol{\Omega}_i^T \boldsymbol{x}\right) \tag{1}$$

where $k$ is the number of Gaussian components in the OGMM, $w_k$ is the weight of the $k$ -th Gaussian component, and $\boldsymbol{\Omega}_i$ is the transformation matrix in which the columns are the eigenvectors of within-class scatter matrix of $C_i$ .

Let $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ be the mean and covariance matrix of the $k$ -th Gaussian component, respectively. $\boldsymbol{\Sigma_k}$ is diagonal:

$$\boldsymbol{\Sigma_k} = \left[\sigma_{kj}\right]_{j=1}^{D} . \tag{2}$$

Then the $k$ -th Gaussian component in Eq. 1 is:

$$O_k\left(\boldsymbol{\Omega}_i^T \boldsymbol{x}\right) = (2\pi)^{-\frac{D}{2}} \left|\boldsymbol{\Sigma}_k\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\Omega}_i^T \boldsymbol{x}-\boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{\Omega}_i^T \boldsymbol{x}-\boldsymbol{\mu}_k\right)\right). \tag{3}$$

## 3. Autoclass based Model Selection

In this section, we present the OGMM selection method based on AutoClass [9], an unsupervised clustering method that use Naïve Bayesian classifier in combination with EM algorithm to find the most probable set of class descriptions.

Each class description consists of two sets of parameters: a set of discrete parameters $\boldsymbol{T}$ which describes the form of probabilistic distribution function, and a set of continuous parameters $\boldsymbol{V}$ that specifies values for the parameter appearing in $\boldsymbol{T}$ . Instead of hard-assigning the data to a class, AutoClass uses the weighted assignment, weighting on the probability of class membership.

The algorithm starts by randomly assigning cluster weights to data, and then searches for the most probable pair of $\boldsymbol{V}$ and $\boldsymbol{T}$ to classify data set. The search is performed by iterating the following two steps until convergence: 1) the maximum posterior (MAP) parameter values $\boldsymbol{V}$ are sought for a given $\boldsymbol{T}$ ; 2) Irrespective of any $\boldsymbol{V}$ , the most probable $\boldsymbol{T}$ is found out. To resolve the local optimum, AutoClass repeats the progress above from randomly generated initializations to search as many local optimums as possible. Finally, the local optimums are ranked according to MAP criterion.

After obtaining the ranked result of AutoClass on the data set of each class, we select the corresponding number of components based on the rank and clustering stability. The process of AutoClass based model selection for OGMM is summarized in Table 1.

**Table 1. Model selection of OGMM based on AutoClass**

```
Input
  The orthogonally transformed features for all the positive samples of a class.
Repeat
  Randomly initialize the parameters.
  Repeat
    1. Estimate the parameter set $\boldsymbol{V}$ by MAP criterion based on the given weighted assignment.
    2. Search for the best probable $\boldsymbol{T}$ by reassigning the cluster weights to data using the Naïve Bayesian Classifier based on the new parameter set $\boldsymbol{V}$ .
  Until convergence or reach the predefined maximum iterative times.
Until reach the predefined maximum local search times.
Output
  The optimum number of clusters
```

# 4. Parameter Estimation Using MMP

## 4.1. MMP

Max-Min posterior Pseudo-probabilities (MMP) is a new kind of discriminative learning for Bayesian classifiers, which is based on the notion of posterior pseudo-probability [7]. The posterior pseudo-probability of being the class $C_i$ for the data $\boldsymbol{x}$ is computed as

$$f\left(p\left(\boldsymbol{x}\middle|C_i\right)\right)=1-\exp\left(-\lambda p^t\left(\boldsymbol{x}\middle|C_i\right)\right), \qquad (4)$$

where $\lambda$ and $t$ are positive numbers. For any input pattern, we compute the corresponding posterior pseudo-probabilities of all the classes under consideration. Then the input pattern is classified into the class with the maximum posterior pseudo-probability.

Since $f\left(p\left(\boldsymbol{x}\middle|C_i\right)\right)$ is a smooth, monotonically increasing function of $p\left(\boldsymbol{x}\middle|C_i\right)$, the posterior pseudo-probabilities based classifier is consistent with traditional Bayesian classifier. However, $f\left(p\left(\boldsymbol{x}\middle|C_i\right)\right)$ takes values in [0, 1], so it is a natural similarity measure and is useful for (1) making rejection decision, (2) combining classifiers, and (3) assessing the performance of a classifier in a much more accurate way than that of counting the number of patterns classified correctly. Furthermore, the new discriminative learning models of Bayesian classifiers such as MMP can be realized by introducing posterior pseudo-probabilities.

The main idea behind the MMP learning is to optimize the classifier performance through maximizing posterior pseudo-probabilities towards 1 for each class and its positive samples, while minimizing those towards 0 for each class and its negative samples. More formally, let $f(\boldsymbol{x};\boldsymbol{\Lambda})$ be the posterior pseudo-probability measure function of a class, where $\boldsymbol{\Lambda}$ denote the set of unknown parameters in it. Let $\hat{\boldsymbol{x}}_i$ be the feature vector of arbitrary positive sample of the class, $\bar{\boldsymbol{x}}_i$ be the feature vector of arbitrary negative sample of the class, $m$ and $n$ be the number of positive and negative samples of the class, respectively. According to the idea above of the MMP learning, the objective function for estimating parameters is designed as

$$F(\boldsymbol{\Lambda})=\frac{1}{m}\sum_{i=1}^{m}\left[f\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\Lambda}\right)-1\right]^2+\frac{1}{n}\sum_{i=1}^{n}\left[f\left(\bar{\boldsymbol{x}}_i;\boldsymbol{\Lambda}\right)\right]^2 \qquad (5)$$

$F(\boldsymbol{\Lambda})=0$ means the perfect classification performance on the training data. Consequently, we can obtain the optimum parameter set $\boldsymbol{\Lambda}^*$ of the posterior pseudo-probability measure function by minimizing $F(\boldsymbol{\Lambda})$:

$$\boldsymbol{\Lambda}^*=\arg\min_{\boldsymbol{\Lambda}} F(\boldsymbol{\Lambda}). \qquad (6)$$

The gradient descent algorithm is employed to obtain the parameter set $\boldsymbol{\Lambda}^*$.

## 4.2. MMP tailored to OGMM estimation

By using OGMM to model class-conditional probability density in Eq. 4, we get

$$f(\boldsymbol{x};\boldsymbol{\Lambda})=1-\exp\left(-\lambda\left(\sum_{k=1}^{K}w_k O_k\left(\boldsymbol{\Omega}_i^T\boldsymbol{x}\right)\right)^t\right). \qquad (7)$$

Consequently, a posterior pseudo-probabilities based classifier with OGMM is established, in which the unknown parameters are

$$\boldsymbol{\Lambda}=\{\lambda,t,w_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\},k=1,\cdots,K. \qquad (8)$$

Some parameters in Eq. 8 must satisfy certain constrains. They are transformed to unconstrained domain for easier implementation. The constraints and transformation of parameters are listed in Table 2, where $\varphi$ is the preset minimum variance value in the covariance matrices for avoiding the estimation error caused by too small variance values. To sum up, the transformed parameter set is

$$\widetilde{\Lambda} = \{\widetilde{\lambda}, \widetilde{t}, \widetilde{w}_k, \boldsymbol{\mu}_k, \widetilde{\boldsymbol{\Sigma}}_k\}, k = 1, \cdots, K \ . \qquad (9)$$

We use the MMP learning algorithm to estimate these parameters, and then transform them into the original ones.

**Table 2. The constrains and transformation of parameters**

| Original parameters and constrains | Transformation of parameters |
|---|---|
| $\lambda > 0$ ; $t > 0$ | $\lambda = \exp(\widetilde{\lambda})$ ; $t = \exp(\widetilde{t})$ |
| $\sigma_{kj} > \varphi$ | $\sigma_{kj} = \exp(\widetilde{\sigma}_{kj}) + \varphi$ |
| $\sum w_k = 1$ | $w_k = \dfrac{e^{\widetilde{w}_k}}{\sum e^{\widetilde{w}_k}}$ |

## 5. Experimental Results

We apply the proposed learning method of OGMM to handwritten digit recognition. The 8-direction gradient features of Liu et al [11] (abbreviated to e-grg there) are used to represent digits in the experiments. In this paper, the original 200-D e-grg is compressed to 120-D by using the Principal Component Analysis (PCA) technique for reducing the computation cost.

We conducted the experiments of handwritten digit recognition on the well-known MNIST database [10], which includes 60,000 training samples and 10,000 test samples. In the experiments, we tested two training schemes, with or without AutoClass based model selection. Furthermore, the EM, the MCE, and the MMP are compared for parameter estimation in each training scheme.

### 5.1. Training results with AutoClass

The training process with AutoClass based model selection includes three stages. In the first stage, the number of components in the OGMM for each digit class was selected using the method described in Section 3. The resultant numbers of components for each digit class are listed in Table 3.

**Table 3. The selected number of components in the OGMM for each digit class**

| Class | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Num. Comp. | 4 | 8 | 5 | 4 | 5 | 6 | 6 | 9 | 4 | 8 |

In the second stage, we used the EM algorithm on positive samples of each digit class to get parameters in the corresponding OGMM. We further set $\lambda = 10$ and $t = 0.02$ in Eq. 7 by careful experiments. In the third stage, MCE and MMP were used on all the positive and negative samples of each class to revise the corresponding initial parameters obtained by the EM algorithm, respectively. Totally, we get three parameter sets for our handwritten digit classifier, which are corresponding with EM, MCE and MMP, respectively. Using each of these three classifiers, closed and open tests of handwritten digit recognition were implemented. The corresponding recognition results are listed in Table 4, where 'Train' denote the error rates on the training set, 'Test ' denote the error rates on the test set, and 'RT' denote the reduction in error rate for the test set, which is brought by the MMP compared with the MCE and the EM, respectively.

**Table 4. Error rates achieved with AutoClass**

| Learning methods | Train(%) | Test(%) | RT(%) |
|---|---|---|---|
| EM | 0.79 | 0.85 | 18.82 |
| MCE | 0.16 | 0.71 | 2.82 |
| MMP | 0.18 | 0.69 | - |

### 5.2. Training results without AutoClass

In order to show the importance of automatic model selection, we also implemented the experiments without AutoClass based model selection. In this scheme, the first stage in Section 5.1, i.e. selecting numbers of components by AutoClass based method, is canceled. Instead, we repeated EM training with varying number of components to find out the number with optimal classification result. Then the trained result was passed to the MMP or the MCE for discriminative learning as in Section 5.1. Consequently, another three handwritten digit classifiers based on posterior pseudo-probabilities were obtained and tested by the same experiments. The results are list in Table 5. Here 'RT_AC' means the reduction in error rate for the test set, which is brought by the training scheme with AutoClass, compared with the scheme without AutoClass.

**Table 5. Error rates achieved without AutoClass**

| Learning methods | Train(%) | Test(%) | RT_AC(%) |
|---|---|---|---|
| EM | 0.82 | 0.92 | 7.61 |
| MCE | 0.19 | 0.77 | 7.79 |
| MMP | 0.20 | 0.74 | 6.76 |

### 5.3. Result analysis

The data in Table 4-5 show that 1) automatic model selection using AutoClass is important to the training of OGMM. Without the model selection, the error rates on both training set and test set are increased no matter what parameter estimation method is used; 2) the performance of parameter estimation by MMP is better than other methods in comparison, no matter the automatic model selection is used or not.

In the paper of Liu et al [11], state-of-the-art techniques of handwritten digit recognition, including features and classifiers, are thoroughly investigated on MNIST database. They reported the best recognition rate of 99.58% on the test set for e-grg features by using SVM with RBF kernel. Our classifier achieved the comparable result, i.e. the recognition rate of 99.31% on the test set, for the same features.

## 6. Conclusion

In this paper, an approach combining unsupervised learning and discriminative learning has been proposed to learn Orthogonal Gaussian Mixture Models (OGMM) for handwritten digit recognition. Our main contributions are summarized as follows.

(1) The unsupervised clustering method of Automatic Classification (AutoClass) has been applied to automatic model selection of finite mixtures. To our best knowledge, it is the first work of using AutoClass for finite mixture model selection in document analysis and recognition. Compared with traditional model selection by experiments, AutoClass brings easier implementation and more reliability, which has been shown in the experiments.

(2) The OGMM, AutoClass, and the Max-Min posterior Pseudo-probabilities (MMP) for learning Bayesian classifiers are integrated to construct a powerful digit classifier which is comparable to previous best counterpart.

It should be noted that the proposed learning approach can be applied to other finite mixture models besides OGMM. We will investigate more applications of our approach in the future. Furthermore, the tasks of model selection and parameter estimation were separately completed in this work. The better learning result is expected to be obtained by simultaneously adjusting the number of components and estimating the component parameters. To this end, inserting MMP criterion into the model selection procedure is another future research direction of ours.

## Acknowledges

The authors would like to thank Dr. Cheng-Lin Liu for providing the e-grg features of handwritten digits. This research was partially supported by 973 Program of China (No. 2006CB303103), Excellent Young Scholars Research Fund of Beijing Institute of Technology (No. 2008YS1203), Open Fund of the National Laboratory of Pattern Recognition, and MCE Theme Project Fund of Microsoft Research Asia (FY08-RES-THEME-158).

## References

[1] S. Axelrod, V. Goel, et al., “Discriminative Estimation of Subspace Constrained Gaussian Mixture Models for Speech Recognition”, *IEEE. Trans. Audio, Speech , and Language Process.*, vol. 15, no. 1, 2007, pp. 172-189.

[2] J. Zeng, Z.Q. Liu, “Markov Random Field-Based Statistical Character Structure Modeling for Handwritten Chinese Character Recognition”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, 2008, pp. 767-780.

[3] C. Constantinos, et al., “Bayesian Feature and Model Selection for Gaussian Mixture Models”, *IEEE Trans. Patten Anal. Mach. Intell.*, vol. 28, no. 6, 2006, pp. 1013-1018.

[4] A.T.F. Mario, et al., “Unsupervised Learning of Finite Mixture Models”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, 2002, pp. 381-396.

[5] B.H. Juang, W. Chou, C.H. Lee, “Minimum classification error rate methods for speech recognition”, *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, May 1997, pp. 257-265.

[6] R. Nopsuwanchai, et al., “Maximization of Mutual Information for Offline Thai Handwriting Recognition”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, August 2006, pp. 1347-1351.

[7] X.B. Liu, Y.D. Jia, X.F. Chen, Y. Deng, and H. Fu , “Image Classification Using the Max-Min Posterior Pseudo-Probabilities Method”, Beijing Institute of Technology, Technical Report BIT-CS-20080001 , 2008, http://www.mcislab.org.cn/member/~xiabi/papers/2008_1.PDF

[8] R. Zhang, X.D. Ding, “Offline Handwritten Numeral Recognition Using Orthogonal Gaussian Mixture Model”, *Proc. 6th Int. Conf. Document Analysis and Recognition*, 2001, pp.1126-1129.

[9] P. Cheeseman, J. Stutz, “Bayesian classification (AutoClass): theory and results”, *Advances in knowledge discovery and data mining*, AAAI Press, 1996, pp.153-180.

[10] Y. LeCun, et al., “Comparison of Learning Algorithms for Handwritten Digit Recognition”, *Proc. Int. Conf. Artificial Neural Networks*, Nanterre, France, 1995, pp.53-60.

[11] C.L. Liu, K. Nakashima, H. Sako, H. Fujisawa, “Handwritten digit recognition: benchmarking of state-of-art techniques”, *Pattern Recognition,* vol.36,2003,pp.2271-2285.

[12] L. Liu, J.L. He, “On the use of Orthogonal GMM in Speaker Recognition”, *Proc. Int. Conf. Acoust., Speech., Signal Process.*, 1999, pp. 845-848.