MULTI-REGULARIZATION PARAMETERS ESTIMATION FOR GAUSSIAN MIXTURE CLASSIFIER BASED ON MDL PRINCIPLE

Xiuling Zhou[1,2], Ping Guo[1] and C. L. Philip Chen[3]
[1]*The Laboratory of Image Processing and Pattern Recognition, Beijing Normal University, Beijing, 100875, China*
[2]*Artificial Intelligence Institute, Beijing City University, Beijing, China*
[3]*The Faculty of Science & Technology, University of Macau, Macau SAR, China*

Keywords: Gaussian classifier, Covariance matrix estimation, Multi-regularization parameters selection, Minimum description length.

Abstract: Regularization is a solution to solve the problem of unstable estimation of covariance matrix with a small sample set in Gaussian classifier. And multi-regularization parameters estimation is more difficult than single parameter estimation. In this paper, KLIM_L covariance matrix estimation is derived theoretically based on MDL (minimum description length) principle for the small sample problem with high dimension. KLIM_L is a generalization of KLIM (Kullback-Leibler information measure) which considers the local difference in each dimension. Under the framework of MDL principle, multi-regularization parameters are selected by the criterion of minimization the KL divergence and estimated simply and directly by point estimation which is approximated by two-order Taylor expansion. It costs less computation time to estimate the multi-regularization parameters in KLIM_L than in RDA (regularized discriminant analysis) and in LOOC (leave-one-out covariance matrix estimate) where cross validation technique is adopted. And higher classification accuracy is achieved by the proposed KLIM_L estimator in experiment.

1 INTRODUCTION

Gaussian mixture model (GMM) has been widely used in real pattern recognition problem for clustering and classification, where the maximum likelihood criterion is adopted to estimate the model parameters with the training samples (Bishop, 2007) (Everitt, 1981). However, it often suffers from small sample size problem with high dimensional data. In this case, for d-dimensional data, if less than d+1 training samples from each class is available, the sample covariance matrix estimate in Gaussian classifier is singular. And this can lead to lower classification accuracy.

Regularization is a solution to this kind of problem. Shrinkage and regularized covariance estimators are examples of such techniques. Shrinkage estimators are a widely used class of estimators which regularize the covariance matrix by shrinking it toward some positive definite target structures, such as the identity matrix or the diagonal of the sample covariance (Friedman, 1989); (Hoffbeck, 1996); (Schafer, 2005); (Srivastava, 2007). More recently, a number of methods have been proposed for regularizing the covariance estimate by constraining the estimate of the covariance or its inverse to be sparse (Bickel, 2008); (Friedman, 2008); (Cao, 2011).

The above regularization methods mainly concern various mixture of sample covariance matrix, common covariance matrix and identity matrix or constraint the estimate of the covariance or its inverse to be sparse. In these methods, the regularization parameters are required to be determined by cross validation technique. Although the regularization methods have been successfully applied for classifying small-number data with some heuristic approximations (Friedman, 1989); (Hoffbeck, 1996), the selection of regularization parameters by cross validation technique is very computation-expensive. Moreover, cross-validation performance is not always well in the selection of linear models in some cases (Rivals, 1999).

Recently, a covariance matrix estimator called Kullback-Leibler information measure (KLIM) is developed based on minimum description length

Zhou X., Guo P. and Chen C..
MULTI-REGULARIZATION PARAMETERS ESTIMATION FOR GAUSSIAN MIXTURE CLASSIFIER BASED ON MDL PRINCIPLE.
DOI: 10.5220/0003669301120117
In *Proceedings of the International Conference on Neural Computation Theory and Applications* (NCTA-2011), pages 112-117
ISBN: 978-989-8425-84-3

(MDL) principle for small number samples with high dimension data (Guo, 2008). The KLIM estimator is derived theoretically by KL divergence. And a formula for fast estimation the regularization parameter is derived. However, since multi-parameters optimization is more difficult than single parameter optimization, only a special case that the regularization parameters are taken the same value for all dimensions is considered in KLIM. Though estimation of regularization parameter becomes simple, the accuracy of covariance matrix estimation is decreased by ignore the local difference in each dimension. This will result in decreasing the classification accuracy of Gaussian classifier finally.

In this paper, KLIM is generalized to KLIM_L which considers the local difference in each dimension. Based on MDL principle, the KLIM_L covariance matrix estimation is derived for the small sample problem with high dimension. Multi-regularization parameters in each dimension are selected by the criterion of minimization the KL divergence and estimated efficiently by two-order Taylor expansion. The feasibility and efficiency of KLIM_L are shown by the experiments.

2 THEORETICAL BACKGROUND

2.1 Gaussian Mixture Classifier

Given a data set $D = \{\mathbf{x}_i\}_{i=1}^{N}$ which will be classified. Assume that the data point in D is sampled from a Gaussian mixture model which has k component:

$$p(\mathbf{x}, \Theta) = \sum_{j=1}^{k} \alpha_j G(\mathbf{x}, \mathbf{m}_j, \mathbf{\Sigma}_j) \quad \text{with } \alpha_j \geq 0,\ \sum_{j=1}^{k} \alpha_j = 1, \tag{1}$$

where

$$G(\mathbf{x}, \mathbf{m}_j, \mathbf{\Sigma}_j) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}_j|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \mathbf{m}_j)^{\mathrm{T}} \mathbf{\Sigma}_j^{-1} (\mathbf{x} - \mathbf{m}_j)\} \tag{2}$$

is a general multivariate Gaussian density function. $\mathbf{x}$ is a random vector, the dimension of $\mathbf{x}$ is d and $\Theta = \{\alpha_j, \mathbf{m}_j, \mathbf{\Sigma}_j\}_{j=1}^{k}$ is a parameter vector set of Gaussian mixture model.

In the case of Gaussian mixture model, the Bayesian decision rule $j^* = \arg\max_j p(j \mid \mathbf{x}, \Theta)$ is adopted to classify the vector $\mathbf{x}$ into class j^* with the largest posterior probability $p(j \mid \mathbf{x}, \Theta)$. After model parameters Θ estimated by the maximum likelihood (ML) method with expectation-maximization (EM) algorithm (Redner, 1984), the posterior probability can be written in the form:

$$p(j \mid \mathbf{x}, \hat{\Theta}) = \frac{\hat{\alpha}_j G(\mathbf{x}, \hat{\mathbf{m}}_j, \hat{\mathbf{\Sigma}}_j)}{p(\mathbf{x}, \hat{\Theta})}, \quad j = 1, 2, \cdots, k. \tag{3}$$

And the classification rule becomes:

$$j^* = \arg\min_j d_j(\mathbf{x}), \quad j = 1, 2, \cdots, k, \tag{4}$$

where

$$d_j(\mathbf{x}) = (\mathbf{x} - \hat{\mathbf{m}}_j)^{\mathrm{T}} \hat{\mathbf{\Sigma}}_j^{-1} (\mathbf{x} - \hat{\mathbf{m}}_j) + \ln |\hat{\mathbf{\Sigma}}_j| - 2 \ln \hat{\alpha}_j. \tag{5}$$

This equation is often called the discriminant function for the class j (Aeberhard, 1994).

Since clustering is more general than classification in the mixture model analysis case, we consider the general case in the following.

2.2 Covariance Matrix Estimation

The central idea of the MDL principle is to represent an entire class of probability distributions as models by a single "universal" representative model, such that it would be able to imitate the behaviour of any model in the class. The best model class for a set of observed data is the one whose representative permits the shortest coding of the data. The MDL estimates of both the parameters and their total number are consistent; i.e., the estimates converge and the limit specifies the data generating model (Rissanen, 1978); (Barron, 1998). The codelength (probability distribution or a model) criterion of MDL involves in the KL divergence (Kullback, 1959).

Now considering a given sample data set $D = \{\mathbf{x}_i\}_{i=1}^{N}$ generated from an unknown density $p(\mathbf{X})$, it can be modelled by a finite Gaussian mixture density $p(\mathbf{x}, \Theta)$, where Θ is the parameter set. In the absence of knowledge of $p(\mathbf{x})$, it may be estimated by an empirical kernel density estimate $p_h(\mathbf{x})$ obtained from the data set. Because these two probability densities describe the same unknown density $p(\mathbf{x})$, they should be best matched with proper mixture parameters and smoothing parameters. According to MDL principle, the model parameters should be estimated with minimized KL divergence $KL(h, \Theta)$ based on the given data drawn from the unknown density $p(\mathbf{x})$ (Kullback, 1959),

$$KL(h,\Theta)=\int p_h(\mathbf{x})\ln\frac{p_h(\mathbf{x})}{p(\mathbf{x},\Theta)}\mathrm{d}\mathbf{x} \tag{6}$$

with

$$\begin{aligned}p_h(\mathbf{x})&=\frac{1}{N}\sum_{i=1}^{N}G(\mathbf{x},\mathbf{x}_i,\mathbf{W}_h)\\&=\frac{1}{N(2\pi)^{d/2}\mid\mathbf{W}_h\mid^{1/2}}\sum_{i=1}^{N}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T\mathbf{W}_h^{-1}(\mathbf{x}-\mathbf{x}_i)\right\}\end{aligned}. \tag{7}$$

Here $\mathbf{W}_h$ is a $d\times d$ dimensional diagonal matrix with a general form,

$$\mathbf{W}_h=diag(h_1,h_2,\cdots,h_d), \tag{8}$$

where h_i, $i=1,2,\cdots,d$ are smoothing parameters (or regularization parameters) in the nonparametric kernel density. In the following this set is denoted as $h=\{h_i\}_{i=1}^{d}$. The Eq. (6) equals to zero if and only if $p_h(\mathbf{x})=p(\mathbf{x},\Theta)$.

If the limit $h\to 0$, the kernel density function $p_h(\mathbf{x})$ becomes a δ function, then Eq. (6) reduces to the negative log likelihood function. So the ordinary EM algorithm can be re-derived based on the minimization of this KL divergence function with the limit $h\to 0$. The ML-estimated parameters are shown as follows:

$$\begin{aligned}\hat{\mathbf{m}}_j&=\frac{1}{n_j}\sum_{i=1}^{N}p(j\mid\mathbf{x}_i,\hat{\Theta})\mathbf{x}_i,\\\hat{\mathbf{\Sigma}}_j&=\frac{1}{n_j}\sum_{i=1}^{N}p(j\mid\mathbf{x}_i,\hat{\Theta})(\mathbf{x}_i-\hat{\mathbf{m}}_j)(\mathbf{x}_i-\hat{\mathbf{m}}_j)^{\mathrm{T}}.\end{aligned} \tag{9}$$

The covariance matrix estimation for the limit $h\searrow 0$ is shown as follows. By minimizing Eq. (6) with respect to $\mathbf{\Sigma}_j$, i.e., setting $\partial KL(h,\Theta)/\partial\mathbf{\Sigma}_j=0$, the following covariance matrix estimation formula can be obtained:

$$\hat{\mathbf{\Sigma}}_j=\frac{\int p_h(\mathbf{x})p(j\mid\mathbf{x},\hat{\Theta})(\mathbf{x}-\hat{\mathbf{m}}_j)(\mathbf{x}-\hat{\mathbf{m}}_j)^{\mathrm{T}}\mathrm{d}\mathbf{x}}{\int p_h(\mathbf{x})p(j\mid\mathbf{x},\hat{\Theta})\mathrm{d}\mathbf{x}}. \tag{10}$$

In this case, the Taylor expansion is used for $p(j\mid\mathbf{x},\Theta)$ at $\mathbf{x}=\mathbf{x}_i$ with respect to $\mathbf{x}$ and it is expanded to first order approximation:

$$p(j\mid\mathbf{x},\hat{\Theta})\approx p(j\mid\mathbf{x}_i,\hat{\Theta})+(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}}\nabla_x p(j\mid\mathbf{x}_i,\hat{\Theta}) \tag{11}$$

with $\nabla_x p(j\mid\mathbf{x}_i,\hat{\Theta})=\nabla_x p(j\mid\mathbf{x},\hat{\Theta})\mid_{\mathbf{x}=\mathbf{x}_i}$.

On substituting the above equation into Eq. (10) and according to the properties of probability density function, the following approximation is finally derived:

$$\mathbf{\Sigma}_j(h)\doteq\mathbf{W}_h+\hat{\mathbf{\Sigma}}_j. \tag{12}$$

The estimation in Eq. (12) is called as KLIM_L in the paper, where $\hat{\mathbf{\Sigma}}_j$ is the ML estimation when $h\to 0$, taking the form of Eq. (9).

3 MULTI- REGULARIZATION PARAMETERS ESTIMATION BASED ON MDL

3.1 Regularization Parameters Selection

The regularization parameter set h in the Gaussian kernel density plays an important role in estimating the mixture model parameter. Different h will generate different models. So selecting the regularization parameters is a model selection problem. In the paper the similar method as in (Guo, 2008) is adopted based on MDL principle to select the regularization parameters in KLIM_L.

According to the principle of MDL, it should be with the shortest codelength to select a model. When $h\neq 0$, the regularization parameters h can be estimated with the minimized KL divergence regarding h with ML estimated parameter $\hat{\Theta}$,

$$h^*=\arg\min J(h),\quad J(h)=KL(h,\hat{\Theta}). \tag{13}$$

Now a second order approximation for estimating the regularization parameter h is adopted here. Rewrite the $J(h)$ as:

$$J(h)=J_0(h)+J_e(h), \tag{14}$$

where $J_0(h)=-\int p_h(\mathbf{x})\ln p(\mathbf{x},\Theta)\mathrm{d}\mathbf{x}$,

$$J_e(h)=\int p_h(\mathbf{x})\ln p_h(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Replacing $\ln p(\mathbf{x},\Theta)$ with the second order term of Taylor expansion into the integral of $J_0(h)$ and resulting in the following approximation of $J_0(h)$,

$$\begin{aligned}J_0(h)\approx&-\frac{1}{N}\sum_{i=1}^{N}\ln p(\mathbf{x}_i,\Theta)\\&-\frac{1}{2N}\sum_{i=1}^{N}\mathrm{trace}(\mathbf{W}_h(\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}\ln p(\mathbf{x}_i,\Theta)))\end{aligned} \tag{15}$$

For very sparse data distribution, the following approximation can be used:

$$p_h(\mathbf{x})\ln p_h(\mathbf{x}) \approx \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{x},\mathbf{x}_i,\mathbf{W}_h)\ln\frac{1}{N}G(\mathbf{x},\mathbf{x}_i,\mathbf{W}_h)$$
$$= \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{x},\mathbf{x}_i,\mathbf{W}_h)[-\ln N - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{W}_h| - \frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}}\mathbf{W}_h^{-1}(\mathbf{x}-\mathbf{x}_i)] \quad (16)$$

Substituting the Eq. (16) into $J_e(h)$, it can be got:

$$J_e(h) = \int p_h(\mathbf{x})\ln p_h(\mathbf{x})d\mathbf{x} \approx -\ln N - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{W}_h| - \frac{d}{2} \quad (17)$$

So far, the approximation formula of $J(h)$ is obtained:

$$J(h) \approx -\frac{1}{2N}\sum_{i=1}^{N}\mathrm{trace}(\mathbf{W}_h(\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}\ln p(\mathbf{x}_i,\Theta))) - \frac{1}{2}\ln|\mathbf{W}_h| + C \quad (18)$$

where C is a constant irrelevant to h.

Let $H_{d\times d} = \nabla_{\mathbf{x}}\nabla_{\mathbf{x}}\ln p(\mathbf{x}_i,\Theta)$. Taking partial derivative of $J(h)$ to $\mathbf{W}_h$ and letting it be equal to zero, the rough approximation formula of h is obtained as follows:

$$\mathbf{W}_h^{-1} = -\frac{1}{N}\sum_{i=1}^{N} diag(H). \quad (19)$$

3.2 Approximation for Regularization Parameters

The Eq. (19) can be rewritten as follows:

$$\mathbf{W}_h^{-1} = -\frac{1}{N}\sum_{i=1}^{N} diag(H) = \frac{1}{N}diag(-\sum_{i=1}^{N}\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}\ln p(\mathbf{x}_i,\Theta))$$

with

$$-\sum_{i=1}^{N}\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}\ln p(\mathbf{x}_i,\Theta)$$
$$= \sum_{i=1}^{N}\{\sum_{j=1}^{k} p(j|\mathbf{x}_i,\Theta)[\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\mathbf{m}_j)(\mathbf{x}_i-\mathbf{m}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}]$$
$$+\sum_{j=1}^{k} p(j|\mathbf{x}_i,\Theta)\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\mathbf{m}_j)\sum_{j=1}^{k} p(j|\mathbf{x}_i,\Theta)(\mathbf{x}_i-\mathbf{m}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\}$$

Considering hard-cut case ($p(j|\mathbf{x}_i,\Theta) = 1$ or 0) and using the approximations $\sum_{i=1}^{N} p(j|\mathbf{x}_i,\Theta)(\mathbf{x}_i-\mathbf{m}_j)(\mathbf{x}_i-\mathbf{m}_j)^{\mathrm{T}} \approx n_j\hat{\boldsymbol{\Sigma}}_j$, $\sum_{i=1}^{N} p(j|\mathbf{x}_i,\Theta) \approx n_j$, $\boldsymbol{\Sigma}_j^{-1}\hat{\boldsymbol{\Sigma}}_j \approx I$ and $\sum_{j=1}^{k}\alpha_j\boldsymbol{\Sigma}_j^{-1} = \boldsymbol{\Sigma}^{-1}$, it can be obtained:

$$\mathbf{W}_h^{-1} \approx diag(\boldsymbol{\Sigma}^{-1}). \quad (20)$$

Suppose the eigenvalues and eigenvectors of the common covariance matrix $\boldsymbol{\Sigma}$ are λ_k and μ_k, $k = 1\cdots d$, where $\mu_k = (\mu_{k1},\mu_{k2},\cdots,\mu_{kd})^{\mathrm{T}}$, then there exists:

$$\boldsymbol{\Sigma} = \sum_{k=1}^{d}\lambda_k\mu_k\mu_k^{T},\ \boldsymbol{\Sigma}^{-1} = \sum_{k=1}^{d}\lambda_k^{-1}\mu_k\mu_k^{T}. \quad (21)$$

Substituting Eq. (21) into Eq. (20) and using the average eigenvalue $\bar{\lambda} = (\sum_{i=1}^{d}\lambda_i)/d = \mathrm{trace}(\boldsymbol{\Sigma})/d$ to substitute each eigenvalue of matrix $\boldsymbol{\Sigma}$ (only in the denominator), it can be obtained:

$$\mathbf{W}_h^{-1} \approx \frac{d^2}{\mathrm{trace}^2(\boldsymbol{\Sigma})}diag(\boldsymbol{\Sigma}) \text{ with } \boldsymbol{\Sigma} = [\sigma_{ij}]_{d\times d}.$$

Finally, the regularization parameters can be approximated by the following equation:

$$\mathbf{W}_h = \frac{\mathrm{trace}^2(\boldsymbol{\Sigma})}{d^2}diag(\frac{1}{\sigma_{11}},\cdots,\frac{1}{\sigma_{dd}}). \quad (22)$$

3.3 Comparison of KLIM_L with Regularization Methods

The four methods (KLIM_L, KLIM, RDA (Regularized Discriminant Analysis, Friedman, 1989) and LOOC (Leave-one-out covariance matrix estimate, Hoffbeck, 1996)) are all regularization methods to estimate the covariance matrix for small sample size problem. They all consider ML estimated covariance matrix with the additional extra matrices.

KLIM_L is derived by the similar way as KLIM under the framework of MDL principle. Meanwhile, the estimation of regularization parameters is similar to KLIM based on MDL principle. KLIM_L is a generalization of KLIM. Multi-regularization parameters are included and estimated in KLIM_L while one regularization parameter is estimated in KLIM. For every σ_{ii} ($i = 1\cdots d$), if it is taken by $\frac{1}{d}\mathrm{trace}(\boldsymbol{\Sigma})$, then $\mathbf{W}_h = (\mathrm{trace}(\boldsymbol{\Sigma})/d)\mathbf{I}_d$. It will reduce to the case of $\mathbf{W}_h = h\mathbf{I}_d$ in KLIM, where $h = \mathrm{trace}(\boldsymbol{\Sigma})/d$.

KLIM_L is derived based on MDL principle while RDA and LOOC are heuristically proposed. They differ in mixtures of covariance matrix considered and the criterion used to select the regularization parameters.

Different computation time costs are required in the four regularization discriminant methods. The time costs of them are sorted decreasing in the following order: KLIM, KLIM_L, LOOC and RDA. This will be validated by the following experiments.

4 EXPERIMENT RESULTS

In this section, the classification accuracy and time cost of KLIM_L are compared with LDA (Linear Discriminant Analysis, Aeberhard, 1994), RDA, LOOC and KLIM on COIL-20 object data (Nene, 1996).

COIL-20 is a database of gray-scale images of 20 objects. The objects were placed on a motorized turntable against a black background. The turntable was rotated through 360 degrees to vary object pose with respect to a fix camera. Images of the objects were taken at pose intervals of 5 degrees, which corresponds to 72 images per object. The total number of images is 1440 and the size of each image is 128×128.

In the experiment, the regularization parameter h of KLIM is estimated by $h = \text{trace}(\mathbf{\Sigma})/d$. The parameter matrix $\mathbf{W}_h$ of KLIM_L is estimated by Eq. (22). In RDA, the values of λ and γ are sampled in a coarse grid, (0.0, 0.25, 0.50, 0.75, 1.0), resulting in 25 data points. In LOOC, the four parameters are taken according to the table in (Hoffbeck, 1996). Six images are randomly selected as training samples from each class to estimate the mean and covariance matrix. And the remaining images are employed as testing samples to verify the classification accuracy. Since the dimension of image data is very high of 128×128, PCA is adopted here to reduce the data dimension. Experiments are performed with five different numbers of dimensions. Each experiment runs 25 times, and the mean and standard deviation of classification accuracy are reported as results. The results of experiment are shown in table 1 table 2.

In the experiment, the classification accuracy of KLIM_L is the best among the five compared methods, while the classification accuracy of KLIM is the second best. The classification accuracy of LOOC is the worst among the compared methods except in dimension 80, where the classification accuracy of LOOC is higher than that of LDA. Considering the time cost of regularization parameters estimating, KLIM_L needs a little more time to estimate the regularization parameters than KLIM needs, while RDA and LOOC need much more time than KLIM_L needs. The experimental results are consistent with the theoretical analysis.

5 CONCLUSIONS

In this paper, the KLIM_L covariance matrix estimation is derived based on MDL principle for the small sample problem with high dimension. Under the framework of MDL principle, multi-regularization parameters are estimated simply and directly by point estimation which is approximated by two-order Taylor expansion. KLIM_L is a generalization of KLIM. With the KL information measure, total samples can be used to estimate the regularization parameters in KLIM_L, making it less computation-expensive than using leave-one-out cross-validation method in RDA and LOOC.

Table 1: Mean classification accuracy on COIL-20 object database.

classifier	LDA	RDA	LOOC	KLIM	KLIM_L
80	81.6(2.6)	87.2(2.2)	82.2(1.8)	87.7(1.8)	90.0(2.2)
70	83.6(2.2)	86.4(1.8)	81.2(2.8)	87.0(1.7)	87.9(1.8)
60	85.0(2.2)	86.8(2.2)	82.8(2.3)	87.7(1.4)	88.1(1.6)
50	85.0(2.5)	86.3(2.4)	80.3(2.9)	87.5(2.0)	87.8(2.3)
40	86.9(1.7)	86.9(1.7)	80.6(3.2)	87.6(1.3)	87.6(1.6)

Table 2: Time cost (in seconds) of estimating regularization parameters on COIL-20 object database.

classifier	RDA	LOOC	KLIM	KLIM_L
80	62.0494	1.8286	3.8042e-005	8.1066e-005
70	48.8565	1.5147	3.4419e-005	7.8802e-005
60	34.2189	1.0275	2.9890e-005	5.2534e-005
50	23.9743	0.7324	2.9890e-005	4.9817e-005
40	16.2443	0.4987	3.0796e-005	4.9364e-005

KLIM_L estimator achieves higher classification accuracy than LDA, RDA, LOOC and KLIM estimators on COIL-20 data set. In the future work, the kernel method combined with these regularization discriminant methods will be studied for small sample problem with high dimension and the selection of kernel parameters will be investigated under some criterion.

ACKNOWLEDGEMENTS

The research work described in this paper was fully supported by the grants from the National Natural Science Foundation of China (Project No. 90820010, 60911130513). Prof. Guo is the author to whom the correspondence should be addressed, his e-mail address is pguo@ieee.org

REFERENCES

Bishop, C. M., 2007. Pattern recognition and machine learning, *Springer-Verlag New York, Inc. Secaucus*, NJ, USA.

Everitt, B. S., Hand, D., 1981. Finite Mixture Distributions, *Chapman and Hall*, London.

Friedman, J. H., 1989. Regularized discriminant analysis, *Journal of the American Statistical Association*, vol. 84, no. 405, 165–175.

Hoffbeck, J. P. and Landgrebe, D. A., 1996. Covariance matrix estimation and classification with limited training data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, 763–767.

Schafer, J. and Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1.

Srivastava, S., Gupta, M. R., Frigyik, B. A., 2007. Bayesian quadratic discriminant analysis, *J. Mach. Learning Res. 8*, 1277 – 1305.

Bickel, P. J. and Levina, E., 2008. Regularized estimation of large covariance matrices, *Annals of Statistics*, vol. 36, no. 1, 199–227.

Friedman, J., Hastie, T., and Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, vol. 9, no. 3, 432–441.

Cao, G., Bachega, L. R., Bouman, C. A., 2011. The Sparse Matrix Transform for Covariance Estimation and Analysis of High Dimensional Signals. *IEEE Transactions on Image Processing*, Volume 20, Issue 3, 625 – 640.

Rivals, I., Personnaz, L., 1999. On cross validation for model selection, *Neural Comput. 11*,863 – 870.

Guo, P., Jia, Y., and Lyu, M. R., 2008. A study of regularized Gaussian classifier in high-dimension small sample set case based on MDL principle with application to spectrum recognition, *Pattern Recognition, Vol. 41*, 2842~2854.

Redner, R. A., Walker, H. F., 1984. Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev. 26*, 195 – 239.

Aeberhard, S., Coomans, de Vel, D., O., 1994. Comparative analysis of statistical pattern recognition methods in high dimensional settings, *Pattern Recognition 27* (8), 1065 – 1077.

Rissanen, J., 1978. Modeling by shortest data description, *Automatica 14*, 465 – 471.

Barron, A., Rissanen, J., Yu, B., 1998. The minimum description length principle in coding and modeling, *IEEE Trans. Inform*. Theory 44 (6), 2743 – 2760.

Kullback, S., 1959. Information Theory and Statistics, *Wiley*, New York.

Nene, S. A., Nayar, S. K. and Murase, H., 1996. Columbia Object Image library(COIL-20). *Technical report CUCS-005-96*.