Classification for High-Dimension Small-Sample data Sets Based on Kullback-Leibler Information Measure*

Ping Guo and Michael R. Lyu
Department of Computer Science & Engineering
The Chinese University of Hong Kong, Shatin, NT, Hong Kong

Abstract *In classifying samples by Gaussian classifier, the covariance matrix estimated with a small number sample set becomes unstable, which leads to degrading the classification accuracy. In this paper, we discuss the covariance matrix estimation problem for small number samples with high dimension setting based on Kullback-Leibler Information Measure. A new covariance matrix estimator is developed, and a fast, rough estimating regularization parameter formula is derived. Experiments are performed to investigate the classification accuracy with developed covariance matrix estimator and higher classification accuracy results are obtained.*

Keywords: Classification, Covariance matrix estimation, Small sample set with high dimension, Smoothing Parameter Selection, Kullback-Leibler Information Measure

1 Introduction

In classification, when a set of samples is given, the goal is to classify them into proper groups according to some criterion of class membership. In recent years, several classification algorithms have been developed to partition a data set into pre-defined classes. When the data are viewed as arising from two or more clusters mixed in varying proportions, we can use finite Gaussian mixture distribution to analyze the data set. The Gaussian mixture distribution analysis method has been used widely in a variety of important practical situations, and the likelihood approach to the fitting of Gaussian mixture models has been utilized extensively.

When classifying data with the Gaussian mixture model, the mean vector and covariance matrix of each component are not known in advance, and they must be estimated from the given data set. While a large size data set is desirable for estimating the parameters more accurately, in some real world situation, only a small-size data set can be obtained because of certain restriction, e.g, high cost in collecting data set. For a relatively small number sample data set, if the dimension d of variable $\mathbf{x}$ is comparable to the number of training samples n_j in class j, the problem becomes poorly-posed. Worse, if the number n_j of training samples is less than data dimensionality, the problem becomes ill-posed. In the later case, not all parameters can be properly estimated and classification accuracy is degraded.

There are two possible solutions for this kind of problems: one is dimensionality reduction, and the other is regularization[1]. Regularization is the procedure of biasing parameters towards what are thought to be more plausible values, which reduces the variance of the estimates at the cost of introducing additional bias. The regularization techniques have been highly successful in classifying small number data with some heuristic approximations[1, 2]. However, the heuristic method, for example RDA[2], requires to select regularization parameters (or called *model*) with some statis-

*The work described in this paper was supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region (Project No. CUHK4432/99E).

tical techniques such as leave-one-out cross-validation, which is computation-expensive. Furthermore, recent studies show that cross-validation does not always perform well in the selection of linear models[3], therefore it is worthy to develop new techniques to deal with this kind of problems.

Kullback–Leibler information measure[4, 5] can be considered as "distance" between two probability density models, whereas this measure is also called as Kullback-Leibler divergence. In this paper, based on the mixture model analysis with Kullback-Leibler information measure, we present the results of investigating covariance matrix estimation and smoothing parameter selection in Gaussian classifier for the classification problem of small sample sets with high dimension.

2 Classification

2.1 Classification with Gaussian Mixture Model

The data points to be classified are assumed to be samples from a mixture of k Gaussian densities, in which the joint probabilistic density is expressed as,

$$\begin{aligned} p(x,\Theta) &= \sum_{j=1}^{k} \alpha_j G(x, m_j, \Sigma_j), \\ \text{with} \quad \alpha_j &\geq 0, \text{ and } \sum_{j=1}^{k} \alpha_j = 1 \end{aligned} \tag{1}$$

where

$$G(x, m_j, \Sigma_j) = \frac{\exp[-\frac{1}{2}(x - m_j)^T \Sigma_j^{-1}(x - m_j)]}{(2\pi)^{d/2}|\Sigma_j|^{\frac{1}{2}}} \tag{2}$$

is the multivariate Gaussian density function, x denotes a random vector, d is the dimension of x, and parameter $\Theta = \{\alpha_j, m_j, \Sigma_j\}_{j=1}^{k}$ is the set of finite mixture model parameter vectors. Here α_j is the *prior* probability, m_j is the mean vector, and Σ_j is the covariance matrix of the j-th component. Based on a given data set, these parameters can be estimated by maximum likelihood(ML) learning with EM algorithm[6, 7].

The Bayesian decision rule is used to classify the x into class j with the largest *posterior* probability. The *posterior* probability $p(j|x)$ represents the probability that a sample point x belongs to class j. Now we use Bayesian decision $j^* = \arg\max_j p(j|x)$ to classify x into class j^*. The densities $p(j|x)$ are usually unknown and have to be estimated from the training samples. With maximum likelihood estimation, the *posterior* density can be written in the form,

$$p(j|x) = \frac{\alpha_j G(x, m_j, \Sigma_j)}{p(x, \Theta)}. \tag{3}$$

If taking the logarithm of the above equation and omitting the common factors of the classes, we obtain the following classification rule,

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

with

$$d_j(x) = (x - m_j)^T \Sigma_j^{-1}(x - m_j) + \ln|\Sigma_j| - 2\ln\alpha_j \tag{5}$$

This equation is often called the *discriminant score* for j-th class in the literature. Further more, if the *prior* probability α_j is the same for all classes, it becomes a discriminant function when omitting the $2\ln\alpha_j$ term.

2.2 Covariance Matrix Estimation based on Kullback-Leibler Information Measure

When the sample number N is small, the estimated covariance matrix becomes inaccurate, and hence the classification accuracy is reduced. To solve this problem, several techniques are proposed. In this paper, we address this problem by using Kullback-Leibler divergence.

We consider that the system can be described by a finite Gaussian mixture model, on the other hand, the data set can be considered as samples drawn from a nonparametric density distribution $p_h(x)$[8]. The "distance" of these two probability density distribution can be measured with the following Kullback-Leibler (KL) divergence[4, 5],

$$KL(h,k,\Theta) = \int p_h(x) \ln \frac{p_h(x)}{p(x,\Theta)} dx \qquad (6)$$

where $p_h(x)$ is assigned as Gaussian kernel density for given samples $D = \{x_i\}_{i=1}^N$,

$$p_h(x) = \frac{1}{N} \sum_{i=1}^{N} G(x, x_i, h^2 I_d). \qquad (7)$$

Here h is the smoothing parameter and I_d is a $d \times d$ dimensional identity matrix.

The ordinary EM algorithm[6, 7] can be re-derived based on the minimization of the Kullback–Leibler divergence function (6) with the limit $h \to 0$.

In the nonparametric kernel density function, the smoothing parameter h plays an important role in the estimating mixture model parameter. To avoid integration difficulty, when h is small, we can use Taylor expansion for $p(j|x)$ at $x = x_i$ and take up to the second order approximation, i.e.,

$$\begin{aligned} p(j|x) &\approx p(j|x_i) + (x - x_i)^T \nabla_x p(j|x_i) \quad (8) \\ &\quad + \frac{1}{2}(x - x_i)^T \nabla_x^2 p(j|x_i)(x - x_i) \end{aligned}$$

where the operator ∇_x and ∇_x^2 are referred to first and second order derivative, respectively.

With this approximation, the following covariance matrix estimation formula can be obtained when minimizing the cost function equation (6) for parameter learning. (It is called KLIM in this paper.)

In the second order approximation, the covariance matrix estimation formula is

$$\mathbf{\Sigma_j}(2,h) \approx h^2 I_d + \frac{\hat{\Sigma}_j}{(1+\eta)} + \frac{\Sigma_Q}{(1+\eta)}. \qquad (9)$$

When h is very small, it reduces into the first order approximation,

$$\mathbf{\Sigma_j}(1,h) = h^2 I_d + \hat{\Sigma}_j. \qquad (10)$$

The following notations are used:

$$\begin{aligned} \eta &= \frac{h^2}{2n_j} \mathrm{Trace}[\sum_{i=1}^{N} H_i(j)], \\ n_j &= \alpha_j N, \qquad H_i(j) = \nabla_x^2 p(j|x_i), \\ \alpha_j &= \frac{1}{N} \sum_{i=1}^{N} p(j|x_i), \qquad m_j = \frac{1}{n_j} \sum_{i=1}^{N} p(j|x_i) x_i, \end{aligned}$$

$$\Sigma_Q = \frac{h^2}{2N} \sum_{i=1}^{N} [\mathrm{Trace}[H_i(j)]](x_i - m_j)(x_i - m_j)^T,$$

$$\hat{\mathbf{\Sigma}}_\mathbf{j} = \frac{1}{n_j} \sum_{i=1}^{N} p(j|x_i)(x_i - m_j)(x_i - m_j)^T. \quad (11)$$

The Hessian matrix can be computed as the following,

$$\begin{aligned} H_i(j) &= p(j|x_i)\{\mathbf{\Sigma_j^{-1}}(x_i - m_j)(x_i - m_j)^T \mathbf{\Sigma_j^{-1}} \\ &\quad - \sum_{j=1}^{k} p(j|x_i)[\mathbf{\Sigma_j^{-1}}(x_i - m_j)(x_i - m_j)^T \mathbf{\Sigma_j^{-1}}]\} \\ &\quad + p(j|x_i)\{\sum_{j=1}^{k} p(j|x_i)\mathbf{\Sigma_j^{-1}} - \mathbf{\Sigma_j^{-1}}\} \\ &\quad + 2p(j|x_i)[\sum_{j=1}^{k} p(j|x_i)\mathbf{\Sigma_j^{-1}}(x_i - m_j) \qquad (12) \\ &\quad - \Sigma_j^{-1}(x_i - m_j)] \sum_{j=1}^{k} p(j|x_i)(x_i - m_j)^T \mathbf{\Sigma_j^{-1}} \end{aligned}$$

Since the quantity such as $\sum_{j=1}^{k} p(j|x_i)Q(j)$ represents the averaged value $Q(j)$ over all classes, the above regularization term reflects the difference between single class quantity and averaged quantity. If there is only one class or the classes are well separated, this Hessian matrix will be a null matrix and estimator the $\mathbf{\Sigma_j}(2,h)$ reduces into $\mathbf{\Sigma_j}(1,h)$.

From the above, we can see that the new kind of regularization form is obtained based on Kullback–Leibler information measure, where the sole parameter h controls the degree of regularization. Next we discuss how an optimal value of smoothing parameter h can be selected based on training samples.

2.3 Smoothing Parameter Selection

There are several ways to select smoothing parameter h, for example, with training samples we can use cross validation statistical technique to select the optimal smoothing parameter. As we know, the goal in selecting smoothing parameter is to produce a model for the probability density which is as close as possible to the unknown density $p(\mathbf{x}, \Theta)$[9]. According to the principle of KL information measure, when $h \neq 0$, the smooth parameter h can be estimated with minimized KL divergence,

$$h^* = \arg\min J(h), \qquad J(h) = KL(k^*, \Theta^*, h) \tag{13}$$

where the parameters with an asterisk represent learnt parameters.

The integration can be approximated by *Monte Carlo method* [10, 11]. For the sake of less computation expense, we use second order approximation for estimating the value of smoothing parameter h in this work.

Using Taylor expansion to logarithmic term in KL integration function, we can obtain,

$$J(h) = KL(h, k^*, \Theta^*) \approx J_0(h) + J_e(h) \tag{14}$$

where the approximations are

$$J_0(h) = -\frac{1}{N}\sum_{i=1}^{N} \ln p(x_i, \Theta) + h^2 J_r(x_i, \Theta) \tag{15}$$

$$J_r(x_i, \Theta) = -\frac{1}{2N}\sum_{i=1}^{N} \mathrm{Trace}[\nabla_x^2 \ln p(x_i, \Theta)] \tag{16}$$

$$J_e(h) = \frac{1}{N}\sum_{i=1}^{N} \ln p_h(x_i) + \frac{h^2}{2N}\sum_{i=1}^{N} \mathrm{Trace}[\nabla_x^2 \ln p_h(x_i)] \tag{17}$$

Now the function $J(h)$ can be computed based on the original samples with summation instead of integration.

For very sparse data distribution, we can use the following approximation to estimate the smoothing parameter.

$$p_h(x) \ln p_h(x) \approx \frac{1}{N}\sum_{i=1}^{N} G(x, x_i, h^2) \ln \frac{1}{N} G(x, x_i, h^2 I_d).$$

Under this approximation, the rough estimation formula is obtained as,

$$h^2 \approx \frac{d}{2J_r(x_i, \Theta)}. \tag{18}$$

2.4 Comparison of KLIM with Other Discriminant Analysis Methods

When the class membership of training samples is known, the hard-cut version of $p(j|x)$ is used in the mean vector and covariance matrix estimation,

$$p(j|x_i) = \begin{cases} 1, & \text{If } x_i \in \text{ class } j \\ 0, & \text{If } x_i \notin \text{ class } j \end{cases} \tag{19}$$

In this case, the sample based ML estimator is $(h = 0)$,

$$m_j = \frac{1}{n_j}\sum\nolimits_{i=1}^{n_j} x_i \tag{20}$$

$$\hat{\Sigma}_j = \frac{1}{n_j}\sum\nolimits_{i=1}^{n_j} (x_i - m_j)(x_i - m_j)^T, \tag{21}$$

where x_i is a sample from class j, and n_j is the training sample number of class j.

Using the classification rule equations (4) and (5) with the above covariance estimator is called quadratic discriminant analysis (QDA). When the class sample size n_j is approximately equal to or smaller than the dimension d, the covariance estimation with equation (21) will become highly variable, and it becomes a poorly-posed or an ill-posed classification problem. To improve such kind of problem, regularization is one of the solution.

One of the regularization methods to deal with the poorly-posed problem is linear discriminant analysis (LDA). In LDA, the Σ_j in equation (5) is replaced with a pooled covariance matrix

$$\Sigma = \frac{1}{N}\sum_{j=1}^{k} n_j \Sigma_j \tag{22}$$

This applies a considerable degree of regularization by substantially reducing the number of parameters to be estimated.

Regularized discriminant analysis (RDA) is another regularization method which was proposed by Friedman[2]. RDA is designed for the small number sample case, where the covariance matrix takes the following form:

$$\mathbf{\Sigma_j}(\lambda,\gamma) = (1-\gamma)\mathbf{\Sigma}_j(\lambda) + \gamma\frac{Trace[\mathbf{\Sigma}_j(\lambda)]}{d}I_d \tag{23}$$

where

$$\mathbf{\Sigma}_j(\lambda) = \frac{(1-\lambda)n_j\mathbf{\Sigma}_j + \lambda N\mathbf{\Sigma}}{(1-\lambda)n_j + \lambda N} \tag{24}$$

The two parameters λ and γ, which are restricted to the range between 0 and 1, are regularization parameters to be selected according to the maximum of the leave-one-out classification accuracy. λ controls the amount of the $\mathbf{\Sigma}_j$ that is shrunk towards $\mathbf{\Sigma}$, while γ controls the shrinkage of the eigenvalues towards equality as Trace$[\mathbf{\Sigma}_j(\lambda)]/d$ is equal to the average of the eigenvalues of $\mathbf{\Sigma}_j(\lambda)$.

The KLIM is derived under the frame of Kullback–Leibler information measure, while RDA is heuristicly proposed. KLIM and RDA are similar in that they both consider ML estimated covariance matrix and the addition of extra matrices. Namely, they both have an identity matrix multiplied by a scalar; however, the scalar term is different from each other. There is a term of weighted parameter with regularized ML estimation in KLIM, which relates to the difference between averaged classes quantities and single class quantities. RDA, on the other hand, considers LDA estimation.

In KLIM, the regularization parameter is the smoothing parameter in kernel density estimation, which can be selected based on KL divergence with total training samples. While in RDA, we have to use some statistical method, such as bootstrap, leave-one-out cross validation, to optimize the regularization parameter. At this point, RDA requires much more computation than KLIM.

Another advantage of KLIM is that it can be used to classify total un-labeled samples since it was related to mixture model analysis. The so-called smoothed EM algorithm[12] is the first order approximation of KLIM with ordinary EM algorithm.

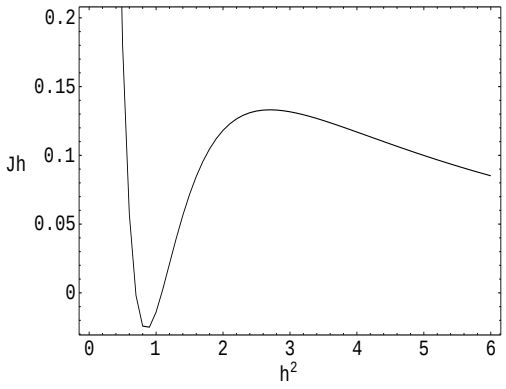

Figure 1: Typical curve for determining the smoothing parameter h using equations (13) and (14). corresponding local minima of $J(h)$ is proper h value.

3 Experiments

In order to investigate the performance of KLIM, we use both synthetic data and real world wine data set[1] to conduct experiments.

In the experiments, the synthetic data set was generated under different conditions. Three experiments with various distributions adapted from Friedman' paper[2] and four dimension (d = 6, 10, 20, 40) were performed. The 15 training samples in each class were randomly drawn from three different Gaussian distribution, and the mean and covariance matrix were estimated based on these training samples. Additional 100 independent test samples from each class were generated to verify the classification accuracy.

In the experiments, the smoothing parameter h was estimated using equations (13) and

[1]This data set was obtained from ftp://ftp.ics.uci. edu/pub/machine-learning-databases/

Table 1: Mean classification accuracy for experiment 1

	$d = 6$	$d = 10$	$d = 20$	$d = 40$
LDA	84.5(3.58)	75.3(6.86)	- - -	- - -
QDA	84.5(3.58)	75.3(6.88)	- - -	- - -
RDA	90.2(1.43)	88.57(5.37)	87.16(2.58)	91.2(2.09)
KLIM	90.2(1.43)	91.73(1.29)	88.4(1.4)	91.26(1.29)

Table 2: Mean classification accuracy for experiment 2

	$d = 6$	$d = 10$	$d = 20$	$d = 40$
LDA	98.1(0.9)	100(0.01)	- - -	- - -
QDA	98.1(0.9)	100(0.01)	- - -	- - -
RDA	98.9(0.8)	100(0.0)	100(0.0)	100(0.0)
KLIM	99.88(0.16)	100(0.0)	100(0.0)	100(0.0)

(14). Figure 1 is a typical $J(h)$ vs. h curve. We select h value corresponding to local minima of $J(h)$. In the case $n_j > d$, we can use equation (18) for quick estimation of h as an initial value. In RDA, the values of both λ and γ were sampled over a very coarse grid, $(0.0, 0.25, 0.50, 0.75, 1.0)$, resulting in 25 data points.

In experiment 1, the covariance matrices of all three classes were equal to the identity matrix, that is, the equal spherical covariance matrices. The means of the classes are hardly different from each other. In experiment 2, all three classes had identical, highly ellipsoidal covariance matrices, but classes are well separated. In experiment 3, the mean vector of all three classes was the same, but the class covariance matrices were unequally highly ellipsoidal. Here the results of experiments were shown in tables 1-3, respectively. In the tables, the value in parentheses represents the standard deviation and dashed lines indicate the covariance matrix is singular in which case reliable results cannot be obtained.

In the experiments 1 and 2, in most cases, KLIM led to higher classification accuracy than LDA, QDA, and was nearly the same as RDA. In the experiment 3, the KLIM classification accuracy is higher than others' except in one case ($d = 20$).

The real world wine data set is 13-dimensional with three classes. This well-posed data set is large with 59, 71 and 48 training samples per class. In order to study the performance of regularized methods, 15 training samples were randomly drawn from each class, whereas the remaining samples were used to verify classification accuracy. Based on this split data set, the result for RDA gives an averaged classification accuracy 94.6. The corresponding measure for LDA is 87.37, and for QDA is 94.9. With a roughly estimated smoothing parameter, the classification accuracy for KLIM is 95.2.

From these experiments, we also know that the smoothing parameter value for KLIM depends on training samples distribution, and it is not an accurate requirement. In most cases the smoothing parameter selection method work well, and the experimental results indicate that the KLIM covariance matrix estimator can lead to a high classification accuracy.

Table 3: Mean classification accuracy for experiment 3

	$d = 6$	$d = 10$	$d = 20$	$d = 40$
LDA	38.8(4.79)	42.2(4.25)	43.16(4.5)	39.64(5.2)
QDA	84.2(3.77)	84.1(6.3)	- - -	- - -
RDA	84.0(3.27)	84.9(5.78)	89.73(2.62)	74.2(8.6)
KLIM	85.8(2.26)	92.7(2.65)	85.84(3.15)	81.75(3.47)

4 Summary

In this paper, based on Kullback–Leibler information measure, the KLIM covariance matrix estimation is investigated for classification problems. An efficient smoothing parameter approximation formula was derived, and the approximation was found from experiments to be valid for most cases. With the Kullback–Leibler information measure, all training samples can be used to estimate the smoothing parameter without the need of validation samples, which is less computation expensive than using the leave-one-out cross-validation method. With the KL information measure based estimation method, all experiments show that the obtained estimator works well, and can lead to a higher classification accuracy than QDA, LDA and RDA estimators.

References

[1] Stefan Aeberhard, Danny Coomans and Olivier de Vel, "Comparative analysis of statistical pattern recognition methods in high dimensional settings," *Patt. Recog.*, vol. 27, no. 8, pp. 1065–1077, 1994.

[2] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, pp. 165–175, 1989.

[3] Isabelle Rivals and Leon Personnaz, "On cross validation for model selection," *Neural Computation*, vol. 11, pp. 863–870, 1999.

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

[5] L. Devroye, *A Course in Density Estimation*, Birhhauser Publisher, Boston, 1987.

[6] N. M. Laird A. P. Dempster and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Society*, vol. B39, pp. 1–38, 1977.

[7] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM Review*, vol. 26, pp. 195–239, 1984.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Boston, second edition, 1990.

[9] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.

[10] George S. Fishman, *Monte Carlo: concepts, algorithms, and applications*, Springer-Verlag, New York, 1996.

[11] James E. Gentle, *Random number generation and Monte Carlo methods*, Springer, New York, 1998.

[12] Lei Xu, "Bayesian Ying-Yang system and theory as a unified statistical learning approach (VII): Data smoothing," in *Proceedings of Intentional Conference on Neural Information Processing* (ICONIP'98), Kitakyushu, Japan, 1998, 1, pp. 243–248.