Graphical Lasso Quadratic Discriminant Function for Character Recognition

Bo Xu[1], Kaizhu Huang[1], Irwin King[2,4], Cheng-Lin Liu[1],
Jun Sun[3], and Naoi Satoshi[3]

[1] National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences, Beijing 100190, P.R. China
[2] Dept. of CSE, The Chinese University of Hong Kong
[3] Fujitsu Research and Development Center, Beijing, China
[4] AT&T Labs Research, San Francisco, USA
{box,kzhuang,liucl}@nlpr.ia.ac.cn, king@cse.cuhk.edu.hk,
{sunjun,naoi}@cn.fujitsu.com

Abstract. The quadratic discriminant function (QDF) derived from the multivariate Gaussian distribution is effective for classification in many pattern recognition tasks. In particular, a variant of QDF, called MQDF, has achieved great success and is widely recognized as the state-of-the-art method in character recognition. However, when the number of training samples is small, covariance estimation involved in QDF will usually be ill-posed, and it leads to the loss of the classification accuracy. To attack this problem, in this paper, we engage the graphical lasso method to estimate the covariance and propose a new classification method called the Graphical Lasso Quadratic Discriminant Function (GLQDF). By exploiting a coordinate descent procedure for the lasso, GLQDF can estimate the covariance matrix (and its inverse) more precisely. Experimental results demonstrate that the proposed method can perform better than the competitive methods on two artificial and six real data sets (including both benchmark digit and Chinese character data).

Keywords: Graphical Lasso, Quadratic Discriminant Function, Character Recognition.

1 Introduction

In many pattern recognition tasks, it is very common to assume that the data follow a Gaussian distribution. The quadratic discriminant function (QDF) derived from the multivariate Gaussian distribution can then be used for classification. Despite of its simplicity, QDF or its variants have achieved great success in many fields. The parameters involved in QDF, e.g., the mean and the covariance, are often obtained via the principle of the maximization-likelihood Estimation (MLE) [6]. MLE has a number of attractive features. First, it usually has good convergence properties as the number of training samples increases. Furthermore, it can often be simpler than alternative methods, such as Bayesian techniques.

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part III, LNCS 7064, pp. 747–755, 2011.

However, when the number of training samples is small (especially when compared to dimensionality), the estimated covariance based on MLE could be often ill-posed, making the covariance matrix singular; this further leads its inverse matrix cannot be computed reliably.

To solve this problem, there have been a number of approaches in the literature. Modified Quadratic Discriminant Function (MQDF) [8] is proposed to replace the minor eigenvalues of covariance matrix of each class with a constant parameter. This small change proves very effective and has made MQDF a state-of-the-art classifier in character recognition. However, the substitution of minor eigenvalues with a constant inevitably loses some class information. Meanwhile, the cutoff threshold of minor eigenvalues and the constant selection are critical for the final performance. Liu et al. [11] proposed a discriminative learning algorithm called Discriminative Learning QDF (DLQDF). It optimizes the parameters of MQDF with the aim to improve the classification accuracy based on the criterion of Minimum Classification Estimation (MCE). Similar to MQDF, DLQDF has the same problem in parameter selection. Alternatively, the Regularized Discriminant Analysis (RDA) [5] improves the performance of QDF by covariance matrix interpolation. Hoffbeck and Landgrebe further extended RDA by optimizing the interpolation coefficients [7]. Empirical results showed that these two algorithms can usually improve the classification performance of QDF. However, the improvements are also dependent on two critical parameters β and γ. In short, all of the above-mentioned methods need empirical settings of parameters to achieve the best results, which are however both time-consuming and task-dependent in real applications.

Different from the above approaches, in this paper, we present a novel method, called the Graphical Lasso Quadratic Discriminant Function (GLQDF). By engaging the graphical lasso, the covariance estimation of the ordinal QDF can be successfully conducted even when the number of training samples is very small. Moreover, we can estimate the inverse of the covariance directly and hence avoid singular problems involved in QDF. One appealing feature is that the whole process is parameter-insensitive. This presents one big advantage over the other methods.

The rest of the paper is organized as follows. In the next section, we make an overview of QDF and MQDF. In Section 3, we introduce our novel GLQDF in details. In Section 4, we conduct a series of experiments to verify our method. Finally, we set out concluding remarks in Section 5.

2 Review of QDF and MQDF

2.1 Quadratic Discriminant Function

Let d be the dimension of the feature. The probability density function of d-dimensional normal distribution is:

$$p(x) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\{-\frac{1}{2}\,(x-u)^t\,\Sigma^{-1}(x-u)\}\ , \tag{1}$$

where x is a d-component vector, μ is the mean vector, and Σ is the $d \times d$ covariance matrix. The quadratic discriminant function is derived from Eq.(1) as follows:

$$g(x) = (x-\mu)^t \Sigma^{-1} (x-\mu) + \log |\Sigma| = \sum_{i=1}^{d} \frac{((x-\mu)^t \varphi_i)^2}{\lambda_i} + \sum_{i=1}^{d} \log \lambda_i \ , \qquad (2)$$

where λ_i is the i-th eigenvalue of Σ sorted by descending order and φ_i is the eigenvector that corresponds to λ_i. This function will lead to the optimal classifier, provided that (1) the actual distribution is normal, (2) the prior probabilities of all categories are equal and (3) the parameters μ and Σ can be reliably provided. However, since the parameters are usually unknown, the sample mean vector $\hat{\mu}$ and sample covariance matrix $\hat{\Sigma}$ are used.

$$\begin{aligned} \hat{g}(x) &= (x-\hat{\mu})^t \hat{\Sigma}^{-1} (x-\hat{\mu}) + \log \left|\hat{\Sigma}\right| \\ &= \sum_{i=1}^{d} \frac{((x-\hat{\mu})^t \hat{\varphi}_i)^2}{\hat{\lambda}_i} + \sum_{i=1}^{d} \log \hat{\lambda}_i \ . \end{aligned} \qquad (3)$$

Here, λ_i is the i-th eigenvalue of $\hat{\sigma}$ and $\hat{\varphi}_i$ is the eigenvector. It is well-known that small eigenvalues in Eq.(3) are usually inaccurate; this hence causes the reduction of recognition accuracy. Moreover, the computational cost of Eq.(3) is $O(d^3)$ for d-dimensional vectors, which may be computationally difficult when the dimension is high.

2.2 Modified Quadratic Discriminant Function

MQDF is a modified version of the ordinary QDF. QDF suffers from the quadratic number of parameters, which cannot be estimated reliably when the number of samples per class is smaller than the feature dimensionality. MQDF reduces the complexity of QDF by replacing the small eigenvalues of covariance matrix of each class with a constant. Consequently, the small eigenvectors will disappear in the discriminant function. This reduces both the space and the computational complexity. More importantly, this small change proves to improve the classification performance significantly. Denote the input sample by a d-dimensional feature vector $x = (x_1, x_2, x_3, , x_d)^T$. For classification, each class c_i is assumed to have a Gaussian density $p(x|c_i) = N(u_i, \sigma_i)$, where μ_i and σ_i are the class mean and covariance matrix, respectively. Assuming equal a priori class probabilities, the discriminant function is given by the log-likelihood:

$$-2 \log p\,(x|c_i) = (x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) + \log |\Sigma_i| + CI \qquad (4)$$

where CI is a class-independent term, and is usually omitted. We take the minus log-likelihood to make the discriminant function a distance measure. The covariance matrix Σ_i can be diagonalized as: Λ_i, where $\Lambda_i = diag[\lambda_{i1}, ..., \lambda_{ik}, ..., \lambda_{id}]$ has the eigenvalues of λ_{ik}(in descending order) as diagonal elements, φ_{ik} is an ortho-normal matrix comprising as columns the eigenvectors of λ_{ik}. Replacing the minor eigenvalues with a constant, i.e., replacing Λ_i with diag$[\lambda_{i1}, , \lambda_{ik}, \delta_i, , \delta_i]$

(k is the number of principal eigenvectors to be retained), the discriminant function of Eq. (3) becomes what we call MQDF:

$$\begin{aligned} g(x,c_i) = &\sum_{j=1}^{k} \frac{((x-\mu_i)^t\varphi_{ij})^2}{\lambda_{ij}} + \sum_{j=1}^{k} \log\lambda_{ij} \\ &+\frac{1}{\delta_i}\left(\|x-\mu_i\|^2 - \sum_{j=1}^{k}\left|(x-\mu_i)^T\varphi_{ij}\right|^2\right) + (d-k)\log\delta_i \ , \end{aligned} \qquad (5)$$

where $i, j = 1, \ldots, k$, are the principal eigenvectors of the covariance matrix of class ω_i. In classification, the input pattern is classified to the class of minimum quadratic distance (MQDF), and multiple candidate classes are ordered in ascending order of distances.

3 Graphical Lasso Quadratic Discriminant Function

In recent years, a number of researchers have proposed the estimation of Gaussian models through the use of L_1 (lasso) regularization, which increase the sparsity of the inverse covariance. Meinshausen and Buhlmann [12] took a simple approach to this problem. They estimated a sparse model by fitting a lasso model to each variable while using the others as predictors. Other researchers have proposed algorithms for the exact maximization of the L_1-penalized log-likelihood. For example, Yuan and Lin [13], Banerjee et al.[1] and Dahl et al. [2] adapted interior point optimization methods for the solution to this problem. Both papers revealed that the simpler approach of Meinshausen and Buhlmann [12] can be viewed as an approximation to the exact problem. Banerjee et al. [1] exploited the blockwise coordinate descent approach to solve the lasso problem. Friedman et al. [4] invented the graphical lasso and applied fast coordinate descent algorithms to solve the lasso problem. Graphical lasso is faster than previous methods and also provides a conceptual link between the exact problem and the approximation suggested by Meinshausen et al. [12]. In the following, we introduce the details on how to apply the graphical lasso on QDF.

The graphical lasso estimates the covariance matrix of Gaussian distribution by recursively solving and updating the lasso problem. Suppose, we have N multivariate normal observations of dimension d, with mean μ and covariance Σ. Let $\Theta = \Sigma^{-1}$ and let S be the empirical covariance matrix, the problem of graphical lasso is to maximize the penalized log-likelihood

$$\log\det\Theta - tr(S\Theta) - \rho\,\|\Theta\|_1 \qquad (6)$$

Here, tr denotes the trace and $||\Theta||_1$ is the L_1 norm$-$the sum of the absolute values of the elements of Σ^{-1}. ρ is a trade-off parameter, which however proves insensitive to the optimization. We set it to 10^{-4} in all the experiments of this paper. Expression (6) is the Gaussian log-likelihood of the data, partially maximized with respect to the mean parameter μ.

Let W be the estimation of Σ. We can solve the problem by optimizing over each row and corresponding column of W in a block coordinate descent approach. Partitioning W and S

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix} , \tag{7}$$

the solution for w_{12} satisfies

$$w_{12} = \arg\min_{y} \left\{ y^T W_{11}^{-1} y : \|y - s_{12}\|_\infty \leq \rho \right\} \tag{8}$$

This is a box-constrained quadratic program (QP), which can be solved using an interior-point procedure. By permuting the rows and columns, the target column is always the last. We can then solve a problem similar to Eq. (8) for each column and update their estimate of W after each stage. This is repeated until convergence. If this procedure is initialized with a positive definite matrix, the iterates from this procedure remains positive definite and invertible, even if $p > N$.

Using convex duality, the solution of problem (8) is equivalent to solving the dual problem

$$\min_{\beta}\{\frac{1}{2}\left\|W_{11}^{1/2}\beta - b\right\|^2 + \rho\|\beta\|_1\} , \tag{9}$$

where $b = W_{11}^{-1/2} s_{12}$; if β solves Eq. (9), then $w_{12} = W_{11}\beta$ solves Eq. (8). Expression (9) resembles a lasso (L_1regularized) least squares problem. If $W_{11} = S_{11}$, the solutions $\hat{\beta}$ are easily seen to equal the lasso estimates for the pth variable on the others. When $W_{11} \neq S_{11}$ in general, we can use fast coordinate descent algorithm [3], which makes solution of the lasso problem very attractive.

To solve problem (9), graphical lasso uses W_{11} and s_{12}, where W_{11} is the current estimate of the upper block of W. The algorithm updates w and cycles through all of the variables until convergence.

The detailed algorithm is listed as below:

Algorithm 1. Graphical lasso algorithm

```
Start with $W = S + \rho I$. The diagonal of $W$ remains unchanged in what follows.
for $j = 1, 2, ...p, 1, 2, ...p, ...$
    input: $W_{11}$ and $s_{12}$
    solve the lasso problem (9)
    gives a $p - 1$ vector solution $\hat{\beta}$.
    fill in the corresponding row and column of $W$ using $w_{12} = W_{11}\hat{\beta}$
    continue until convergence
end for
```

4 Experimental Results

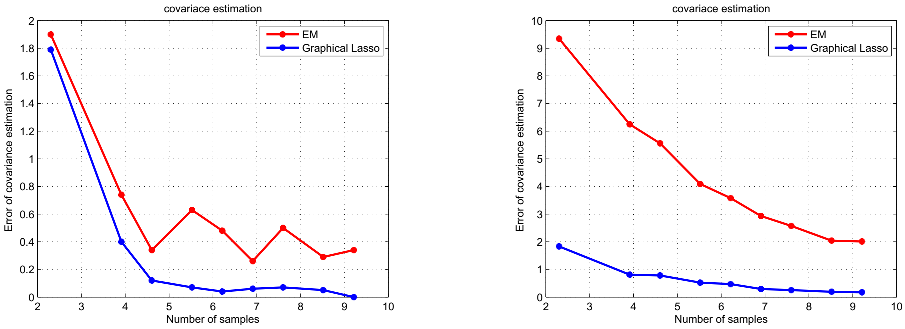
We conduct extensive experiments to verify the effectiveness of the proposed algorithm for covariance estimation and classification. We compare our algorithm to the state-of-the-art algorithm MQDF. All the algorithms are implemented and run using matlab on a PC with 3.0Ghz CPU and 2G RAM.

4.1 Results on Synthetic Data

In this section, we perform experiments on synthetic data to measure how accurate the proposed graphical Lasso covariance estimate will be. We compared the estimated covariance obtained by graphical lasso and the EM algorithm, which is used in QDF. In more details, we first generate samples following a specific Gaussian distribution. We then use EM and Graphical Lasso to estimate the covariance. Finally we examine the estimation error between the ground truth covariance and estimated covariance. The estimation error is computed by Eq. (10)

$$D = sqrt(\sum_{i=1}^{m}\sum_{j=1}^{m}|C_{ij} - C'_{ij}|). \tag{10}$$

We generate both two-dimensional data and ten-dimensional data, the number of samples are from 50 to 10000. The results are listed in Fig. 1.



(a) 2-dimensional estimation

(b) 10-dimensional estimation

Fig. 1. Estimation Error on Synthetic Data

From the results, we can see that the graphical lasso estimates the covariance more precisely than EM estimator both on 2-dimensional data and 10-dimensional data. The superiority is more distinctive when the number of samples is smaller than the data dimensionality. This can be seen in the left part of Fig. 1(b).

4.2 Results on UCI

To examine the classification performance of GLQDF, we conduct a series of experiments on three data sets from UCI repository: 1)Optdigits: with 10 class and 64 dimension, 3,823 training and 1,797 test samples. 2) Sat, with 6 class and 36 dimension, 4,435 training and 2,000 test samples. 3) HW306class: with 153 class and 512 dimension, 91,365 training and 9,141 test samples. For simplicity, we apply Linear Discriminant Analysis (LDA) to reduce the dimensionality to the class number minus 1 in the experiments. The recognition rate of MQDF and GLQDF is listed in Table 1. It is clear that the GLQDF achieves better recognition rate in every dataset than MQDF. This clearly demonstrates the advantages of the proposed GLQDF.

Table 1. Recognition rate on UCI data sets

dateset	MQDF	GLQDF
Optdigits	94.0	**94.4**
Sat	84.8	**85.8**
HW306class	93.4	**96.0**

4.3 Results on Handwritten Digital Datasets

In this section, we report the experimental results of the proposed algorithm on two handwritten digital datasets, USPS and MNIST. USPS contains $9,298$ handwriting character measurements divided into 10 classes. The entire USPS data set is divided into two parts, a training set with $7,291$ measurements and a test set with $2,007$ measurements. The original image size is 16×16. The MNIST dataset is another handwritten digits data collection, in which a training set of 60,000 examples and a test set of $10,000$ examples in 10 classes. The original image size is 20×20. We compare the recognition rate of different classifier on both the pixel-level feature and gradient feature. The pixel-level feature number of those two datasets is 256 and 400. The gradient feature is extracted by the algorithm in [9]. We specify 8 direcgions of gradient, choose grid structure of 4×4 for USPS and 5×5 for MNIST. Thus, the gradient feature dimensionality of USPS and MNIST is 128 and 200, respectively. We reduce the dimensionality to c - 1 by LDA in both the USPS and MNIST and feed to the MQDF and GLQDF for training and test. We obtain the hyper-parameter of MQDF, which is a multiplier used for the selection of constant δ_i, by cross validation and we select the principle axes as 8. The final results on pixel feature is listed in Table 2 and the result on gradient feature is listed in Table 3.

From the results, either on the pixel feature or gradient feature, the recognition rate of GLQDF is better than the MQDF. This proves again the effectiveness of the lasso criterion based covariance estimation.

Table 2. Recognition rate on handwritten digits data set of pixel feature

dataset	MQDF	GLQDF
USPS	89.09	**89.74**
MNIST	89.91	**90.07**

Table 3. Recognition rate on handwritten digits data set of gradient feature

dataset	MQDF	GLQDF
USPS	95.96	**96.16**
MNIST	**98.21**	**98.21**

4.4 Results on Handwritten Chinese Character Data

We exploit the CASIA data set for comparison. The CASIA data set, collected by the Institute of Automation, Chinese Academy of Sciences, contains $3,755$ Chinese characters of the level-1 set of the standard GB2312-80, 300 samples per class. We choose 250 samples per class for training and the remaining 50 samples per class for test. To save time, we only selected the first 200 classes from CASIA data for our experiment. Each binary image of CASIA data was firstly normalized to gray-scale image of 64×64 pixels by the bi-moment normalization method [10]. Then the 8-direction gradient direction features were extracted. The resulting 512-dimensional feature vector was projected into a low dimensional subspace learned by the global LDA. All of projected vectors were then fed to the MQDF classifier and GLQDF classifier. The hyper-parameter of MQDF was learned by cross validation and its principle axes was set as 20 in different lower subspace.

To compare the performance between MQDF and GLQDF, we projected the original features into different lower subspace and recorded the recognition rate of the corresponding classifier. The results were listed in the Table 4. From the results, we can see that GLQDF almost achieves the same recognition rate as the MQDF, even when the number of lower subspace is equal to 150.

Table 4. Recognition rate on CASIA data set

Dimensionality	MQDF	GLQDF
LDA = 30	**98.72**	**98.72**
LDA = 50	**99.22**	99.15
LDA = 100	**99.52**	99.46
LDA = 150	99.51	**99.54**

5 Conclusion

In this paper, we engage the graphical lasso method to estimate the covariance and propose a new quadratic method called the Graphical Lasso Quadratic Discriminant Function (GLQDF). By exploiting a coordinate descent procedure for the lasso, GLQDF can estimate the covariance matrix more precisely. We can even compute the inverse of the covariance. This solves the singular problem in covariance estimation, especially when the number of samples is smaller than the dimensionality. Extensive experiments demonstrate that the proposed method can perform better than the competitive methods on two artificial and six real data sets.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (NSFC) under grants No. 60825301, No. 61075052 and the Council of the Hong Kong Special Administrative Region, China under the Research Grants No. CUHK 413210.

References

1. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. The Journal of Machine Learning Research 9, 485–516 (2008)
2. Dahl, J., Roychowdhury, V., Vandenberghe, L.: Maximum likelihood estimation of gaussian graphical models: numerical implementation and topology selection. UCLA preprint (2005)
3. Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. The Annals of Applied Statistics 1(2), 302–332 (2007)
4. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3), 432–445 (2008)
5. Friedman, J.: Regularized discriminant analysis. Journal of the American Statistical Association 84(405), 165–175 (1989)
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning (2001)
7. Hoffbeck, J., Landgrebe, D.: Covariance matrix estimation and classification with limited training data. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(7), 763–767 (1996)
8. Kimura, F., Takashina, K., Tsuruoka, S., Miyake, Y.: Modified quadratic discriminant functions and the application to chinese character recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (1), 149–153 (1987)
9. Liu, C., Nakashima, K., Sako, H., Fujisawa, H.: Handwritten digit recognition: benchmarking of state-of-the-art techniques. Pattern Recognition 36(10), 2271–2285 (2003)
10. Liu, C., Sako, H., Fujisawa, H.: Handwritten chinese character recognition: alternatives to nonlinear normalization. In: Proc. 7th ICDAR, pp. 524–528 (2003)
11. Liu, C., Sako, H., Fujisawa, H.: Discriminative learning quadratic discriminant function for handwriting recognition. IEEE Transactions on Neural Networks 15(2), 430–444 (2004)
12. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34(3), 1436–1462 (2006)
13. Yuan, M., Lin, Y.: Model selection and estimation in the gaussian graphical model. Biometrika 94(1), 19–35 (2007)