Neurocomputing 129 (2014) 33–40

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Graphical lasso quadratic discriminant function and its application to character recognition

Bo Xu [a], Kaizhu Huang [e,*], Irwin King [c], Cheng-Lin Liu [b], Jun Sun [d], Naoi Satoshi [d]

[a] Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, PR China
[b] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, PR China
[c] Department of CSE, The Chinese University of Hong Kong, Hong Kong
[d] Fujitsu Research and Development Center, Beijing, China
[e] Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, 111 Ren Ai Road, Dushu Lake Higher Education Town, Suzhou, Jiangsu 215123, China

ARTICLE INFO

Article history:
Received 31 January 2012
Received in revised form
16 June 2012
Accepted 11 August 2012
Available online 7 November 2013

Keywords:
Graphical lasso
Quadratic discriminant function
Character recognition

ABSTRACT

Multivariate Gaussian distribution is a popular assumption in many pattern recognition tasks. The quadratic discriminant function (QDF) is an effective classification approach based on this assumption. An improved algorithm, called modified QDF (or MQDF in short) has achieved great success and is widely recognized as the state-of-the-art method in character recognition. However, because both of the two approaches estimate the mean and covariance by the maximum-likelihood estimation (MLE), they often lead to the loss of the classification accuracy when the number of the training samples is small. 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 nine real datasets (including both benchmark digit and Chinese character data).

1. Introduction

Statistical techniques are widely used for classification in various pattern recognition problems [14]. Statistical classifiers include linear discriminant function (LDF), quadratic discriminant function (QDF), Parzen window classifier, nearest-neighbor (1-NN), *k*-NN rules and margin classifiers [13,12]. QDF is derived under the assumption of multivariate Gaussian distribution for each class. Despite its simplicity, QDF and its variants have achieved great success in many fields. In a performance evaluation study of classifiers in handwritten character recognition, QDF and its variants were shown to be superior in the resistance to noncharacters even though they were not trained with noncharacter data. The parameters involved in QDF, e.g., the mean and the covariance, are often obtained via the principle of the maximum-likelihood estimation (MLE) [10]. 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. 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 to not be computed reliably.

To solve this problem, there have been a number of approaches in the literature. Modified quadratic discriminant function (MQDF) [15] 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. [19] 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 error (MCE). Similar to MQDF, DLQDF has the same problem in parameter selection. Alternatively, the regularized discriminant analysis (RDA) [6] improves the performance of QDF by covariance matrix interpolation. Hoffbeck and Landgrebe further extended RDA by optimizing the interpolation coefficients [11]. 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

* Corresponding author.

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.08.073

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 detail. 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

In this section, we review the QDF and the MQDF and also present some basic notations used throughout the paper.

2.1. Quadratic discriminant function

In this section we briefly review the algorithm of QDF. Let $x=(x_1,\ldots,x_d)^T$ represent a feature of a pattern, the posteriori probability can be computed by the Bayes rule:

$$P(\omega_i|x)=\frac{P(\omega_i)p(x|\omega_i)}{p(x)},\quad i=1,\ldots,M \tag{1}$$

where $P(\omega_i)$ is the prior probability of class ω_i, $p(x|\omega_i)$ is the class probability density function (pdf) and $p(x)$ is the mixture density function. Since $p(x)$ is independent of class label, the nominator of Eq. (1) can be used as the discriminant function for classification:

$$g(x|\omega_i)=p(\omega_i)p(x|\omega_i). \tag{2}$$

Assume the pdf of each class is multivariate Gaussian:

$$p(x)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-u)^t\Sigma^{-1}(x-u)\right\}, \tag{3}$$

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. (3) as follows:

$$g(x|\omega_i)=(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)+\log|\Sigma_i|. \tag{4}$$

The QDF is actually a distance metric in the sense that the class of minimum distance is assigned to the input pattern.

By K–L transform, the covariance matrix can be diagonalized as

$$\Sigma=\Phi\Lambda\Phi^T \tag{5}$$

where $\Lambda=diag[\lambda_1,\ldots,\lambda_d]$ with λ_i, $i=1,\ldots,d$, being the eigenvalues (in decreasing order) of Λ, and $\Phi=[\phi_1,\ldots,\phi_d]$ with ϕ_i, $i=1,\ldots,d$, being the ordered eigenvectors.

Thus the QDF can be rewritten in the form of eigenvectors and eigenvalues:

$$\begin{aligned}g(x|\omega_i)&=[\Phi_i^T(x-\mu_i)]^T\Lambda_i^{-1}\Phi_i^T(x-\mu_i)+\log|\Lambda_i|\\&=\sum_{j=1}^{d}\frac{((x-\mu_i)^t\varphi_{ij})^2}{\lambda_{ij}}+\sum_{j=1}^{d}\log\ \lambda_{ij}.\end{aligned} \tag{6}$$

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

$$\begin{aligned}\hat{g}(x|\omega_i)&=[\hat{\Phi}_i^T(x-\hat{\mu}_i)]^T\hat{\Lambda}_i^{-1}\hat{\Phi}_i^T(x-\hat{\mu}_i)+\log|\hat{\Lambda}_i|\\&=\sum_{j=1}^{d}\frac{((x-\hat{\mu}_i)^t\hat{\varphi}_{ij})^2}{\hat{\lambda}_{ij}}+\sum_{j=1}^{d}\log\ \hat{\lambda}_{ij},\end{aligned} \tag{7}$$

here λ_{ij} is the i-th eigenvalue of $\hat{\Sigma}_i$ and $\hat{\varphi}_i$ is the eigenvector. It is well-known that small eigenvalues in Eq. (7) are usually inaccurate; this hence causes the reduction of recognition accuracy. Moreover, the computational cost of Eq. (7) 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 $x=(x_1,x_2,x_3,\ldots,x_d)^T$. 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

$$-2\log\ p(x|\omega_i)=(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)+\log|\Sigma_i|+CI \tag{8}$$

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 $\Lambda_i=diag[\lambda_{i1},\ldots,\lambda_{ik},\ldots,\lambda_{id}]$ 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 $diag\ [\lambda_{i1},\ldots,\lambda_{ik},\delta_i,\ldots,\delta_i]$ (k is the number of principal eigenvectors to be retained), the discriminant function of Eq. (7) becomes what we call MQDF:

$$\begin{aligned}g(x|\omega_i)=&\sum_{j=1}^{k}\frac{((x-\mu_i)^t\varphi_{ij})^2}{\lambda_{ij}}+\sum_{j=1}^{k}\log\ \lambda_{ij}\\&+\frac{1}{\delta_i}\left(\|x-\mu_i\|^2-\sum_{j=1}^{k}|(x-\mu_i)^T\varphi_{ij}|^2\right)+(d-k)\log\ \delta_i\end{aligned} \tag{9}$$

where $i,j=1,\ldots,k$ are the principal eigenvectors of the covariance matrix of class ω_i.

By defining

$$r_i(x)=\|x-\mu_i\|^2-\sum_{j=1}^{k}|(x-\mu_i)^T\phi_{ij}|^2 \tag{10}$$

where $r_i(x)$ is the residual of subspace projection, Eq. (9) can be rewritten as

$$g(x|\omega_i)=\sum_{j=1}^{k}\frac{((x-\mu_i)^t\varphi_{ij})^2}{\lambda_{ij}}+\sum_{j=1}^{k}\log\ \lambda_{ij}+\frac{1}{\delta_i}r_i(x)+(d-k)\log\ \delta_i \tag{11}$$

The parameters of MQDF are estimated as follows. The mean vector and covariance matrix of a class are estimated from the sample data of this class. The class-dependent δ_i is calculated by

the average of minor eigenvalues

$$\delta_i = \frac{\mathrm{tr}(\Sigma_i) - \sum_{j=1}^{k} \lambda_{ij}}{d-k} = \frac{1}{d-k} \sum_{j=k+1}^{d} \lambda_{ij} \tag{12}$$

where $\mathrm{tr}(\Sigma_i)$ denotes the trace of covariance matrix.

In classification, the input pattern is classified to the class of minimum quadratic distance and multiple candidate classes are ordered in the ascending order of distances.

There are at least three appealing features about MQDF. Firstly, it overcomes the bias of minor eigen-values (which are underestimated on small sample size) such that the classification performance can be improved. Second, for computing MQDF, only the principal eigenvectors and the eigenvalues are to be stored so that the memory space is reduced. Third, the computation effort is largely saved because the projections to minor axes are not computed [19].

3. Graphical lasso quadratic discriminant function

In this section, we focus on introducing the graphical lasso quadratic discriminant function. We will present the problem formulation, the related work, and the involved optimization method.

3.1. Problem formulation

The key problem in the QDF is the estimation of covariance matrix and mean. QDF applies maximum-likelihood to estimate the covariance which usually has a lower bias when there are enough training samples. However, when the number of training samples is small, the estimation results will have a large bias and thus decrease the classification accuracy. To solve this problem, we apply log-likelihood instead of the maximum-likelihood to estimate the covariance matrix.

Suppose we are given n samples independently drawn from an m-dimensional Gaussian distribution: $y^{(1)}, \ldots, y^{(n)} \sim N(\mu, \Sigma_p)$, where the covariance matrix Σ is to be estimated. Let S denote the second moment matrix about the mean:

$$S := \frac{1}{n} \sum_{k=1}^{n} (y^{(k)} - \mu)(y^{(k)} - \mu)^T. \tag{13}$$

Let $\Theta = \Sigma^{-1}$, the problem of graphical lasso is to maximize the penalized log-likelihood

$$\hat{\Sigma}^{-1} = \arg\max_{\Theta > 0} \log \det \Theta - \mathrm{tr}(S\Theta) - \rho \|\Theta\|_1, \tag{14}$$

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} [2]. The scalar parameter ρ controls the size of the penalty. In the case where $Smac; sc; 0$, the classical maximum likelihood estimate is recovered for $\rho = 0$. However, when the number of samples n is small compared to the number of variables p, the second moment matrix may not be invertible. In such cases, for $\rho > 0$, the estimator performs some regularization so that the estimate $\hat{\Sigma}$ is always invertible, no matter how small the ratio of samples to variables is.

3.2. Related work

In recent years, a number of researchers have proposed the estimation of Gaussian models through the use of L_1 (lasso) regularization, which increases the sparsity of the inverse covariance. Meinshausen and Bühlmann [20] 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 [23], Banerjee et al. [2], and Dahl et al. [4] adapted interior point optimization methods for the solution to this problem, Bigot and Biscay [3] used a matrix regression model for high-dimensional covariance matrix estimation by a group lasso penalty. All these papers revealed that the simpler approach of Meinshausen and Bühlmann [20] can be viewed as an approximation to the exact problem. Banerjee et al. [2] exploited the blockwise coordinate descent approach to solve the lasso problem. Friedman et al. [8] 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 and Bühlmann [20].

3.3. Graphical lasso solution

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

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix}, \tag{15}$$

the solution for w_{12} satisfies

$$w_{12} = \arg\min_y \{y^T W_{11}^{-1} y : \|y - s_{12}\|_\infty \le \rho\} \tag{16}$$

This is a box-constrained quadratic program (QP), which can be solved using an iterative interior-point procedure. At each iteration, the target column is the last by permuting the rows and columns. By solving Eq. (16) for each column, we obtain a column of the solution. This procedure is repeated until convergence. If

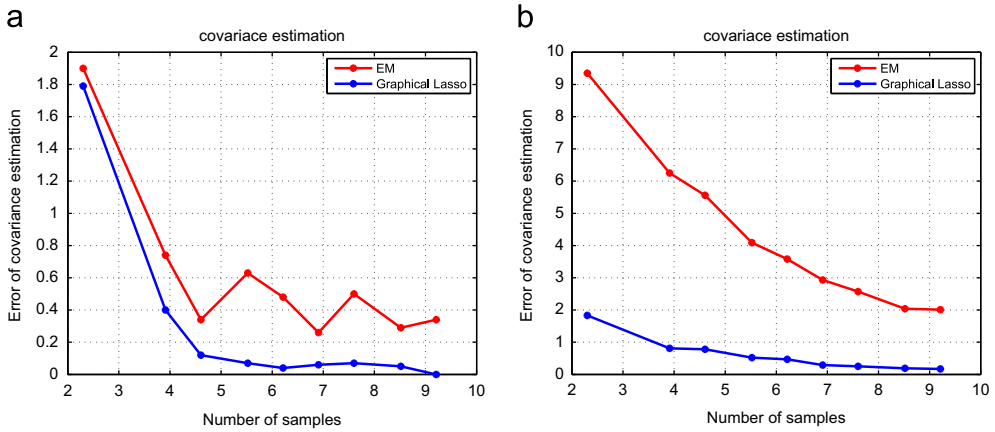

Fig. 1. Estimation error on synthetic data. (a) 2-dimensional estimation and (b) 10-dimensional estimation.

this procedure is initialized with a positive definite matrix, the iterates from this procedure remain positive definite and invertible, even if $p > N$.

Using convex duality, the solution of problem (16) is equivalent to solving the dual problem

$$\min_{\beta}\left\{\frac{1}{2}\|W_{11}^{1/2}\beta - b\|^2 + \rho\|\beta\|_1\right\}, \tag{17}$$

where $b = W_{11}^{-1/2}s_{12}$; if β solves Eq. (17), then $w_{12} = W_{11}\beta$ solves Eq. (16). Expression (17) 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 p-th variable on the others. When $W_{11} \neq S_{11}$ in general, we can use the fast coordinate descent algorithm [7], which makes solution of the lasso problem very attractive.

To solve problem (17), we use W_{11} and s_{12}, where W_{11} is the current estimate of the upper block of W. This algorithm updates w and cycles through all of the variables until convergence.

The detailed algorithm is listed in Algorithm 1.

Algorithm 1. Graphical lasso algorithm.

1: Start with $W = S + \rho I$. The diagonal of W remains unchanged in what follows.
2: **for** $j = 1, 2, \ldots, p,\ 1, 2, \ldots, p, \ldots$
3: input: W_{11} and s_{12}
4: solve the lasso problem (17)
5: give a $p-1$ vector solution $\hat{\beta}$
6: fill in the corresponding row and column of W using $w_{12} = W_{11}\hat{\beta}$
7: continue until convergence
8: **end for**

Table 1
Description of the used UCI datasets.

Datasets	# of classes	# of dimension	# of training	# of test
Ecoli	8	7	303	33
Wine	3	13	161	17
Car	4	6	1682	186
Optdigits	10	64	3823	1797
Sat	6	36	4435	2000
HW306	153	512	91 365	9141

3.4. Graphical lasso quadratic discriminant function algorithm

As a short summary of the above-mentioned QDF and the graphical lasso algorithm, the GLQDF algorithm can be divided into two steps. The first step is the estimate of class covariance and its inverse under the penalized log-likelihood criteria, which is realized by the graphical lasso algorithm. The input parameters of the graphical lasso algorithm include the empirical covariance of class and the penalized factor ρ. The output of the algorithm is the estimated covariance and its corresponding inverse matrix. In the second step, the parameters Σ^{-1} and $|\Sigma|$ are then input into Eq. (4) to achieve the final discriminant function.

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 problem involved in QDF. One appealing feature is that the whole process is parameter-insensitive. This presents one big advantage over the other methods.

Table 2
USPS and MNIST datasets for experiments.

Datasets	# of classes	Image size	# of training	# of test
USPS	10	16×16	7291	2007
MNIST	10	20×20	60 000	10 000

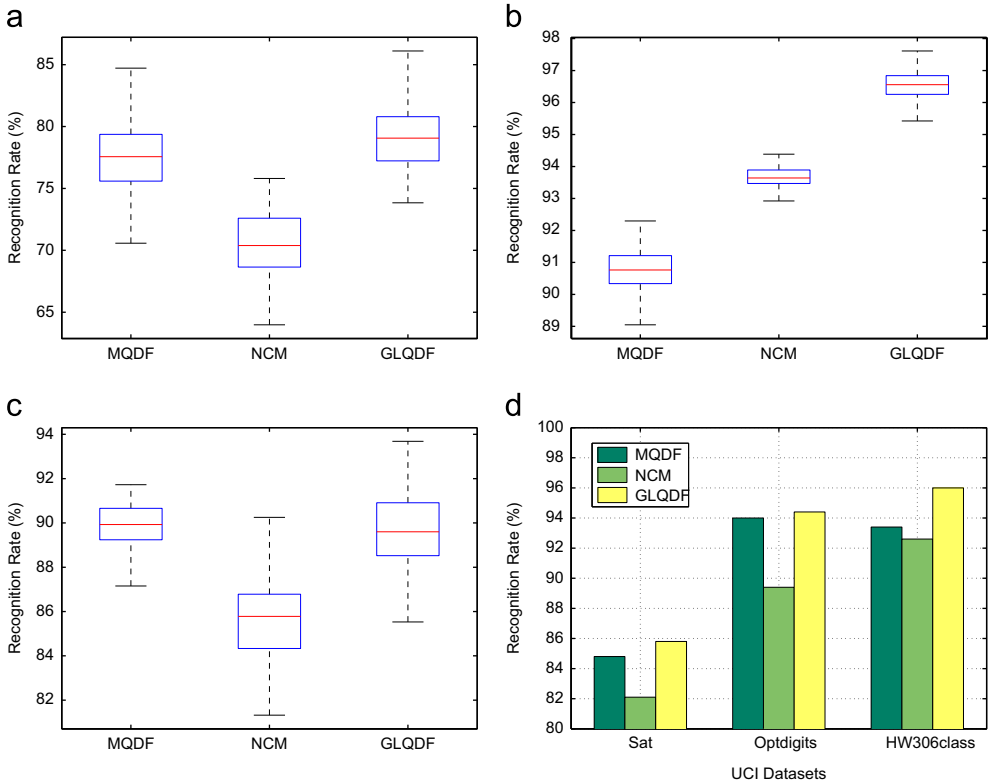

Fig. 2. Recognition rate comparison among different methods. (a) Wine, (b) car, (c) Ecoli, and (d) Sat, Optdigits and HW306.

4. Experimental results

We conduct extensive experiments to verify the effectiveness of the proposed algorithm for covariance estimation and classification. All the algorithms are implemented and run using Matlab on a PC with 3.0 GHz CPU and 2 GB 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 detail, 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 the estimated covariance. The estimation error is computed by the below equation:

$$D = sqrt\left(\sum_{i=1}^{m}\sum_{j=1}^{m}|C_{ij}-C_{ij}^{'}|\right). \quad (18)$$

We generate both two-dimensional data and ten-dimensional data, the number of samples are from 50 to 10 000. The results are listed in Fig. 1.

From the results, we can see that the graphical lasso estimates the covariance more precisely than the 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 six datasets from UCI repository [1], summarized in Table 1. These datasets have been used in many other studies [5,16,21]. We implemented the MQDF [15,22] and the popular nearest class mean (NCM) [9], and used them as the comparison methods with the proposed GLQDF.

For simplicity, we apply linear discriminant analysis (LDA) to reduce the dimensionality to the class number by 1 in the experiments. After the dimensionality reduction, the MQDF, NCM and GLQDF classifiers are then adopted to evaluate the performance. The reported test accuracies are acquired using 10-fold cross validation (CV) for the first three UCI datasets and the average results and their standard deviations are reported in Fig. 2(a)–(c). For Sat-log, Optdigits and HW306, the accuracies are calculated on their specified test sets and the results are reported in Fig. 2(d). It is clear that the GLQDF achieves better recognition rate in every dataset than MQDF and NCM. This clearly demonstrates the advantages of the proposed GLQDF.

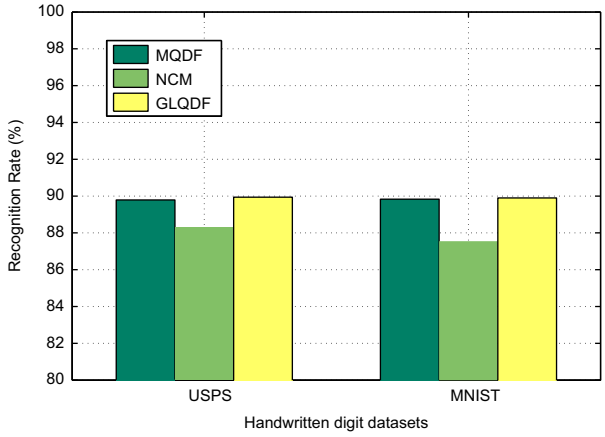

Fig. 4. Recognition rate comparison on USPS and MNIST using pixel-level features.

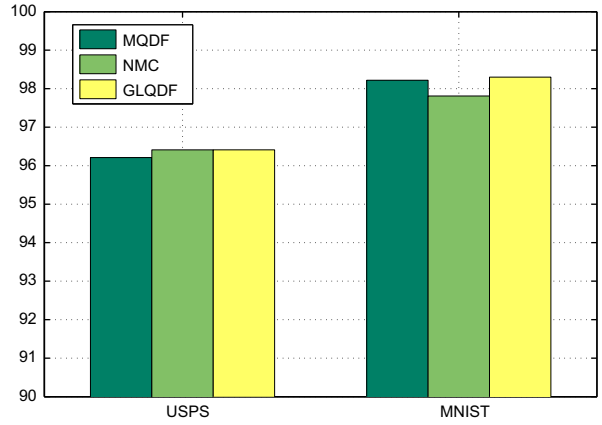

Fig. 5. Recognition rate comparison on USPS and MNIST using gradient features.

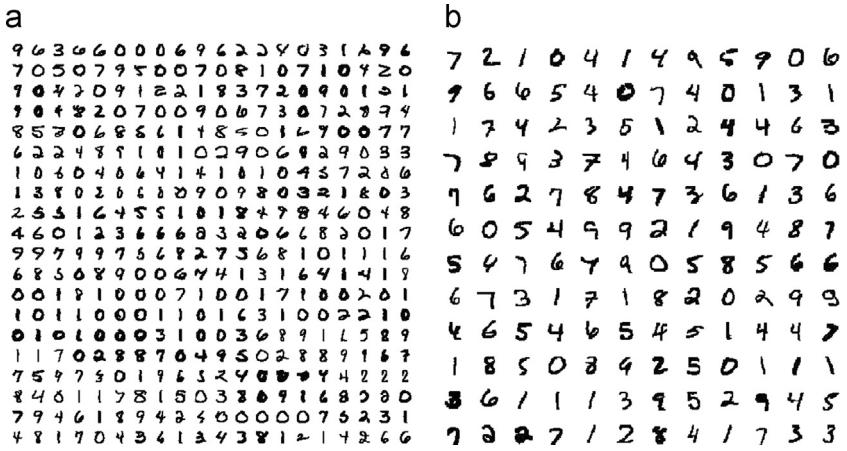

Fig. 3. USPS and MNIST samples. (a) USPS samples and (b) MNIST samples.

4.3. Results on handwritten digit datasets

In this section, we report the experimental results of the proposed algorithm on two handwritten digit datasets, the United States Postal Services (USPS) dataset and MNIST. The basic information is listed in Table 2. Fig. 3 illustrates some image samples from these two datasets.

We compare the recognition rate of different classifiers on both the pixel-level feature and the gradient feature. The pixel-level feature number of the two datasets is 256 and 400, respectively. The gradient feature is extracted by the algorithm in [17]. We specify 8 directions of gradient and 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 the MNIST and feed to the MQDF, NCM 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. We select the principle axes as 8. The final results on pixel feature are listed in Fig. 4 and the result on gradient feature is listed in Fig. 5.

From the results, using both the pixel features or gradient features, the recognition rate of GLQDF is better than the MQDF and NCM. This proves again the effectiveness of the lasso criterion based covariance estimation.

Fig. 6. Samples of CASIA.

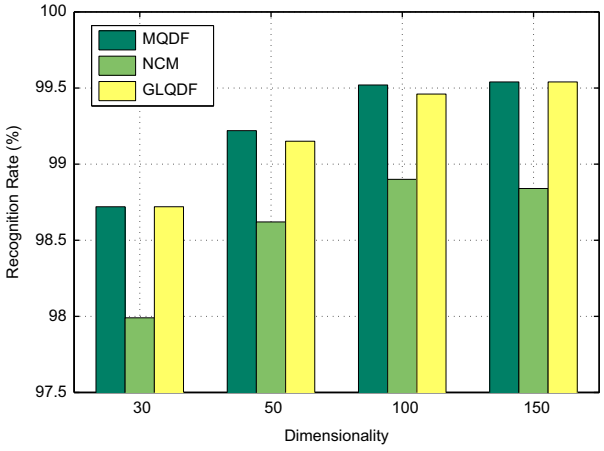

Fig. 7. Recognition rate on CASIA.

4.4. Results on handwritten Chinese character data

We exploited the CASIA dataset for comparison. The CASIA dataset, collected by the Institute of Automation, Chinese Academy of Sciences, contains 3755 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. Fig. 6 describes some image samples from the dataset. We 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 [18]. 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 the projected vectors were then fed to the MQDF, NCM and GLQDF classifier. The hyper-parameter of MQDF was learned by cross validation and its principle axes were set to 20 in different lower subspaces.

To compare the performance among MQDF, NCM and GLQDF, we projected the original features into different lower subspace and recorded the recognition rate of the corresponding classifier. The results are listed in Fig. 7. From the results, we can see that GLQDF achieves competitive performance than the MQDF, even when the number of lower subspace is equal to 150. However, since our GLQDF merely needs to tune one parameter (ρ) which proves not sensitive, it appears more stable than MQDF. Furthermore, compared to NCM, GLQDF demonstrates much better performance. This shows again the advantages of GLQDF.

4.5. Parameter insensitiveness analysis

In this section, we investigate how the parameter ρ of GLQDF influences the recognition performance in USPS and MNIST datasets by only using the pixel-level features. By varying ρ from 1 to 1000 gradually, we obtain the corresponding recognition rate and show the results in Fig. 8. As we can see, the performance curves are basically flat. This verifies that the final recognition rate is not much sensitive to the scalar factor ρ.

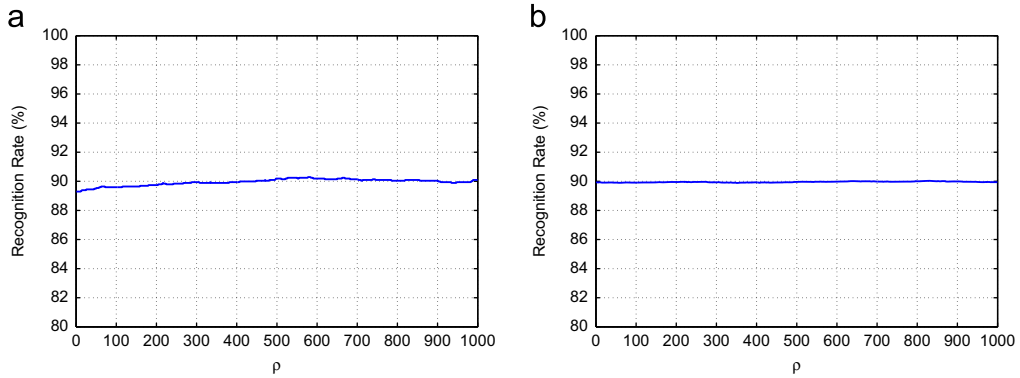

Fig. 8. Parameter insensitiveness analysis on USPS and MNIST. (a) USPS recognition and (b) MNIST recognition.

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 nine real datasets.

Acknowledgments

This work was supported by National Basic Research Program of China (973 Program) Grants 2012CB316301 and 2012CB316302, National Natural Science Foundation of China (NSFC) Grant nos. 61075052, 60825301 and the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 413210).

References

[1] A. Asuncion, D. Newman, UCI Machine Learning Repository, 2007.
[2] O. Banerjee, L. ElGhaoui, A. d'Aspremont, Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, J. Mach. Learn. Res. 9 (2008) 485–516.
[3] J. Bigot, R. Biscay, Group lasso estimation of high-dimensional covariance matrices, J. Mach. Learn. Res. 12 (2011) 3187–3225.
[4] J. Dahl, V. Roychowdhury, L. Vandenberghe, Maximum Likelihood Estimation of Gaussian Graphical Models: Numerical Implementation and Topology Selection, UCLA Preprint, 2005.
[5] R. Duin, M. Loog, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, IEEE Trans. Pattern Anal. Mach. Intell. 26 (6) (2004) 732–739.
[6] J. Friedman, Regularized discriminant analysis, J. Am. Stat. Assoc. 84 (405) (1989) 165–175.
[7] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, Ann. Appl. Stat. 1 (2) (2007) 302–332.
[8] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (3) (2008) 432–445.
[9] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.
[10] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, , 2001.
[11] J. Hoffbeck, D. Landgrebe, Covariance matrix estimation and classification with limited training data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (7) (1996) 763–767.
[12] K. Huang, H. Yang, I. King, M. Lyu, Maxi–min margin machine: learning large margin classifiers locally and globally, IEEE Trans. Neural Networks 19 (2) (2008) 260–272.
[13] K. Huang, H. Yang, I. King, M. Lyu, L. Chan, The minimum error minimax probability machine, J. Mach. Learn. Res. 5 (2004) 1253–1286.
[14] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.
[15] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to Chinese character recognition, IEEE Trans. Pattern Anal. Mach. Intell. 9 (1) (1987) 149–153.
[16] T. Li, S. Zhu, M. Ogihara, Using discriminant analysis for multi-class classification, in: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM), 2003, pp. 589–592.
[17] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, Handwritten digit recognition: benchmarking of state-of-the-art techniques, Pattern Recognition 36 (10) (2003) 2271–2285.
[18] C.-L. Liu, H. Sako, H. Fujisawa, Handwritten Chinese character recognition: alternatives to nonlinear normalization, in: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR), 2003, pp. 524–528.
[19] C.-L. Liu, H. Sako, H. Fujisawa, Discriminative learning quadratic discriminant function for handwriting recognition, IEEE Trans. Neural Networks 15 (2) (2004) 430–444.
[20] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Ann. Stat. 34 (3) (2006) 1436–1462.
[21] B. Xu, K. Huang, C.-L. Liu, Dimensionality reduction by minimal distance maximization, in: Proceedings of the 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 569–572.
[22] Z. Xu, K. Huang, J. Zhu, I. King, M. Lyu, A novel kernel-based maximum a posteriori classification method, Neural Networks 22 (7) (2009) 977–987.
[23] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika 94 (1) (2007) 19–35.

Bo Xu is currently an Assistant Professor at Institute of Automation, the Chinese Academy of Sciences from July 2011. He received M.E. degree in Automation in 2006 from Xi'an JiaoTong University and the Ph.D. degree from Pattern Recognition and Artificial Intelligence from the Institute of Automation, Chinese Academy of Sciences, in 2011. His research interests include pattern recognition, image processing, machine learning and especially the applications to character recognition.

Kaizhu Huang holds the Associate Professor position at Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University. Meanwhile, he has been the Visiting Fellow at University of Bristol since 2009. He was an Associate Professor at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, The Chinese Academy of Sciences from 2009 to 2012. Dr. Huang was a student of the Special Class for Gifted Youth at Xi'an Jiaotong University and received the B.Sc. degree in Engineering in 1997. He received the M.Sc. degree in Engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China in July 2000 and the Ph.D. degree from The Chinese University of Hong Kong in 2004. He worked as a researcher in Fujitsu Research Center from 2004 to 2007. During 2008 and 2009, he was a research fellow in The Chinese University of Hong Kong and a research assistant at University of Bristol, UK.

Huang has published three books in Springer and over 80 research papers (24 indexed by SCI and 50+ by EI) in book chapters, journals (JMLR, IEEE-T-PAMI, IEEE T-NN, IEEE T-BME, IEEE T-SMC, Neural Computation, NN, etc.) and conferences (NIPS, IJCAI, UAI,CIKM, ICDM, ICML,ECML, CVPR, etc.). In addition, he also holds 6 patents in China. He is a member of IEEE, ACM, INNS, and CCF. He served on the programme committees in many international conferences such as ICONIP, IJCNN, IWACI, EANN, KDIR. Especially, he serves as chairs in several major conferences or workshops, e.g., ICONIP 2014 (Program Chair), DMC 2012, 2013, 2014 (Organizing co-Chair), ICDAR 2011 (Publication Chair), ACPR 2011 (Publicity Chair), ICONIP2006, 2009-2011 (Session Chair).

Irwin King (SM'08) received his B.Sc. degree in Engineering and Applied Science from California Institute of Technology, Pasadena, and his M.Sc. and Ph.D. degrees in Computer Science from the University of Southern California, Los Angeles. In 1993, he joined the Chinese University of Hong Kong, where he is now a Professor in the Department of Computer Science and Engineering. He is currently on leave to be with AT&T Labs Research and also a visiting professor with the School of Information, University California at Berkeley. Dr. Kings research interests include machine learning, social computing, web intelligence, data mining, and multimedia processing. In these research areas, he has over 210 technical publications in journals and top conferences. In addition, he has contributed over 20 book chapters and edited volumes. Moreover, Dr. King has over 30 research and applied grants. One notable system he has developed is the VeriGuide system, which detects similar sentences and performs readability analysis of text-based documents in both English and in Chinese to promote academic integrity and honesty. Dr. King is the Book Series Editor for "Social Media and Social Computing" with Taylor and Francis (CRC Press). He is also an Associate Editor of the IEEE Transactions on Neural Networks (TNNs) and IEEE Computational Intelligence Magazine (CIM). He has served as a Special Issue Guest Editor for Neurocomputing, International Journal of Intelligent Computing and Cybernetics (IJICC), Journal of Intelligent Information Systems (JIIS), and International Journal of Computational Intelligent Research (IJCIR). He is a senior member of IEEE and a member of ACM, International Neural Network Society (INNS), and Asian Pacific Neural Network Assembly (APNNA). Currently, he is serving the Neural Network Technical Committee (NNTC) and the Data Mining Technical Committee under the IEEE Computational Intelligence Society (formerly the IEEE Neural Network Society). He is a member of the Board of Governors of INNS and the Vice-President and a Governing Board Member of APNNA. He also serves INNS as the Vice-President for Membership in the Board of Governors. In addition, he has served as a member in the RGC Engineering Panel for the Hong Kong SAR Government, in the Review Panel of the Natural Sciences and Engineering Research Council of Canada (NSERC), and also in the Review Panel of the Natural Science, and Engineering of Academy of Finland.

Cheng-Lin Liu received the B.S. degree in Electronic Engineering from Wuhan University, Wuhan, China, the M.E. degree in Electronic Engineering from Beijing Polytechnic University, Beijing, China, the Ph.D. degree in Pattern Recognition and Artificial Intelligence from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1989, 1992 and 1995, respectively. He was a postdoctoral fellow at Korea Advanced Institute of Science and Technology (KAIST) and later at Tokyo University of Agriculture and Technology from March 1996 to March 1999. From 1999 to 2004, he was a research staff member and later a senior researcher at the Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan. From 2005, he has been a Professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, and is now the Deputy Director of the laboratory. His research interests include pattern recognition, image processing, neural networks, machine learning, and especially the applications to character recognition and document analysis. He has published over 70 technical papers at international journals and conferences. He won the IAPR/ICDAR Young Investigator Award of 2005.

Jun Sun received the B.S. degree in Industrial Automation from Hunan University, Hunan, China, the M.E. degree and Ph.D degree in pattern analysis and intelligent system from Tsinghua University, Beijing, China, in 1995, 1998 and 2001, respectively. He joint Fujitsu R&D Center Co., Ltd. (FRDC) in Beijing in 2002. His research interests including document image analysis, pattern recognition and machine learning. He has published over 30 papers at journals and conferences. He served as programme committees in many international conference such as ICDAR, DAS, ICONIP, ACPR, ICFHR, CBDAR, etc. He is now the director of Information Technology Laboratory of FRDC and the member of the Industrial Liaison Committee of IAPR.

Satoshi Naoi received a B.E. and M.E. degrees from Keio University in 1983 and 1985 respectively. He joined Fujitsu Laboratories Ltd. in 1985 and is currently a president in Fujitsu R&D Center in China. His research interests include character recognition, pattern recognition and image processing, Ph.D. He was a visiting associate professor and a visiting professor of Tokyo Institute of Technology from 2001 to 2005 and from 2005 to 2012, respectively. He is a member of the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan and IEEE.