Sparse Discriminant Analysis Based on the Bayesian Posterior Probability Obtained by L1 Regression

Akinori Hidaka and Takio Kurita

[1] Tokyo Denki University
[2] Hiroshima University

Abstract. Recently the kernel discriminant analysis (KDA) has been successfully applied in many applications. However, kernel functions are usually defined a priori and it is not known what the optimum kernel function for nonlinear discriminant analysis is. Otsu derived the optimum nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities similar with the Bayesian decision theory. Kurita derived discriminant kernels function (DKF) as the optimum kernel functions in terms of the discriminant criterion by investigating the optimum discriminant mapping constructed by the ONDA. The derived kernel function is defined by using the Bayesian posterior probabilities. We can define a family of DKFs by changing the estimation method of the Bayesian posterior probabilities. In this paper, we propose a novel discriminant kernel function based on L1-regularized regression, called L1 DKF. L1 DKF is given by using the Bayesian *posterior* probabilities estimated by L1 regression. Since L1 regression yields a sparse representation for given samples, we can naturally introduce the sparseness into the discriminant kernel function. To introduce the sparseness into LDA, we use L1 DKF as the kernel function of LDA. In experiments, we show sparseness and classification performance of L1 DKF.

1 Introduction

Recently the kernel discriminant analysis (KDA), a non-linear extension of linear discriminant analyasis (LDA), has been successfully applied in many applications [1, 8]. KDA constructs a nonlinear discriminant mapping by using kernel functions. Usually the kernel function is defined a priori, and it is not known what the best kernel function for nonlinear discriminant analysis (NDA) is. Also the class information is usually not introduced in kernel functions.

On the other hand, Otsu derived the optimum nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities [9–11] similar with the Bayesian decision theory [2]. He showed that the optimum nonlinear discriminant mapping was obtained by using variational calculus and was closely related to Bayesian decision theory (The *posterior* probabilities). The optimum nonlinear discriminant mapping can be defined as a linear combination of the Bayesian *posterior* probabilities and the coefficients of the linear combination are obtained

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 648–656, 2012.

by solving the eigenvalue problem of the matrices defined by using the Bayesian *posterior* probabilities.

Kurita showed that the best kernel function is derived from the optimum discriminant mapping constructed by ONDA by investigating the dual problem of the eigenvalue problem of ONDA [7]. The derived kernel function, called the discriminant kernel function (DKF), is also given by using the *posteriori* probabilities. This means the class information is naturally introduced in the kernel function. As like ONDA, the DKF is also optimum in terms of the discriminant criterion. Kurita also showed that a family of DKFs can be defined by changing the estimation method of the Bayesian *posterior* probabilities [7].

Recently, many researchers have actively studied about *sparseness* of features or classifiers [3][13]. It is known that the sparse representation often brings several good properties for classification problems; lower computational load, higher classification accuracy or a feature representation which is easy to interpret.

One of the approach to give sparse representation to existing methods is to introduce the L1-regularized penalty into optimization problems. Based on this approach, Sparse principal component analysis (PCA) by Zou et al. [13] and sparse LDA by Clemmensen et al. [3] were proposed.

In this paper, we propose a novel discriminant kernel function based on L1-regularized regression, called L1 DKF. L1 DKF is obtained by using the Bayesian *posterior* probabilities estimated by L1 regression. Since L1 regression yields a sparse representation for given samples, we can naturally introduce the sparseness into DKF.

We use L1 DKF as the kernel function of LDA to introduce the sparseness into LDA indirectly. Our approach is different from Clemmensen's approach which brings the sparseness into LDA directly [3]. In experiments, we show sparseness and classification performance.

In Sec. 2, we briefly summarize LDA and its nonlinear extensions, KDA and ONDA. In Sec. 3, we describe about discriminant kernels. In Sec. 4, we propose L1 regression based discriminant kernel function. The experiments are shown in Sec. 5. The conclusions are described in Sec. 6.

2 Optimal Nonlinear Discriminant Analysis

2.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [4] is defined as a method to find the linear combination of features which best separates two classes of objects. LDA is regarded as one of the well known methods to extract the best discriminating features for multi-class classification.

Let an $m-$D feature vector be $\boldsymbol{x} = (x_1, \ldots, x_m)^T$. Consider K classes denoted by $\{C_1, \ldots, C_K\}$. Assume that we have N feature vectors $\{\boldsymbol{x}_i | i = 1, \ldots, N\}$ as training samples and they are labeled as one of the K classes. Then LDA constructs a dimension reducing linear mapping from the input feature vector $\boldsymbol{x}$ to a new feature vector $\boldsymbol{y} = A^T\boldsymbol{x}$ where $A = [a_{ij}]$ is the coefficient matrix.

The objective of LDA is to maximize the discriminant criterion,

$$J = \mathrm{tr}(\hat{\Sigma}_T^{-1}\hat{\Sigma}_B) \tag{1}$$

where $\hat{\Sigma}_T$ and $\hat{\Sigma}_B$ are respectively the total covariance matrix and the between-class covariance matrix of the new feature vectors $\boldsymbol{y}$.

The optimal coefficient matrix A is then obtained by solving the following generalized eigenvalue problem

$$\Sigma_B A = \Sigma_T A \Lambda \quad (A^T \Sigma_T A = I) \tag{2}$$

where $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_L)$ is a diagonal matrix of eigen values and I shows the unit matrix. The matrices Σ_T and Σ_B are respectively the total covariance matrix and the between-class covariance matrix of the input feature vectors $\boldsymbol{x}$.

2.2 Kernel Discriminant Analysis

The kernel discriminant analysis (KDA) is one of the nonlinear extensions of LDA. Consider a nonlinear mapping Φ from a input feature vector $\boldsymbol{x}$ to the new feature vector $\Phi(\boldsymbol{x})$. For the case of 1−D feature extraction, the discriminant mapping can be given as $y = \boldsymbol{a}^T\Phi(\boldsymbol{x})$. Since the coefficient vector $\boldsymbol{a}$ can be expressed as a linear combinations of the training samples as $\boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \Phi(\boldsymbol{x}_i)$, the discriminant mapping can be rewritten as

$$y = \sum_{i=1}^{N} \alpha_i \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\alpha}^T \boldsymbol{k}(\boldsymbol{x}), \tag{3}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}) = \Phi(\boldsymbol{x}_i)^T\Phi(\boldsymbol{x})$ and $\boldsymbol{k}(\boldsymbol{x}) = (K(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{x}_N, \boldsymbol{x}))$ are the kernel function defined by the nonlinear mapping $\Phi(\boldsymbol{x})$ and the vector of the kernel functions, respectively.

Then the discriminant criterion is given as

$$J = \frac{\sigma_B^2}{\sigma_T^2} = \frac{\boldsymbol{\alpha}^T \Sigma_B^{(K)} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \Sigma_T^{(K)} \boldsymbol{\alpha}}, \tag{4}$$

where σ_T^2 and σ_B^2 are respectively the total variance and the between-class variance of the discriminant feature y, and $\Sigma_T^{(K)}$ and $\Sigma_B^{(K)}$ are respectively the total covariance matrix and the between-class covariance matrix of the kernel feature vector $\boldsymbol{k}(\boldsymbol{x})$ (details are denoted in [7]).

The optimum coefficient vector $\boldsymbol{\alpha}$ can be obtained by solving the generalized eigenvalue problem $\Sigma_B^{(K)}\boldsymbol{\alpha} = \Sigma_W^{(K)}\boldsymbol{\alpha}\lambda$.

For the multi-dimension case, the kernel discriminant mapping is given by $\boldsymbol{y} = A^T\boldsymbol{k}(\boldsymbol{x})$, where the coefficinet matrix A is defined by $A^T = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N)$. The optimum coefficient matrix A is obtained by solving the eigenvalue problem

$$\Sigma_B^{(K)} A = \Sigma_W^{(K)} A \Lambda. \tag{5}$$

Usually the kernel function is defined a priori in KDA. However it is not noticed what the best kernel function for nonlinear discriminant analysis is. Also the class information is usually not introduced in these kernel functions.

2.3 Optimal Nonlinear Discriminant Analysis

Otsu derived the optimal nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities [9–11]. This assumption is similar with the Bayesian decision theory. Similar with LDA, ONDA constructs the dimension reducing optimum nonlinear mapping which maximizes the discriminant criterion J. Namely ONDA finds the optimum nonlinear mapping in terms of the discriminant criterion J.

By using Variational Calculus, Otsu showed that the optimal nonlinear discriminant mapping is obtained as

$$\boldsymbol{y} = \sum_{k=1}^{K} P(C_k|\boldsymbol{x})\boldsymbol{u}_k \tag{6}$$

where $P(C_k|\boldsymbol{x})$ is the Bayesian *posterior* probability of the class C_k given the input $\boldsymbol{x}$. The vectors $\boldsymbol{u}_k (k = 1, \ldots, K)$ are class representative vectors which are determined by the following generalized eigenvalue problem

$$\Gamma U = PU\Lambda \tag{7}$$

where $\Gamma = [\gamma_{ij}]$ is a $K \times K$ matrix whose elements are defined by

$$\gamma_{ij} = \int (P(C_i|\boldsymbol{x}) - P(C_i))(P(C_j|\boldsymbol{x}) - P(C_j))p(\boldsymbol{x})d\boldsymbol{x} \tag{8}$$

and the other matrices are defined as $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]^T$, $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_L)$, $P = \mathrm{diag}\,(P(C_1), \ldots, P(C_K))$. It is important to notice that the optimal nonlinear mapping is closely related to Bayesian decision theory, namely the *posterior* probabilities $P(C_k|\boldsymbol{x})$.

By using the eigen vectors obtained by solving the generalized eigenvalue problem (7), we can construct the optimum nonlinear discriminant mapping from a given input feature $\boldsymbol{x}$ to the new discriminant feature $\boldsymbol{y}$ as shown in the equation (6) if we can know or estimate all the *posterior* probabilities. This means that we have to estimate the *posterior* probabilities for real applications. It also implies a family of nonlinear discriminant mapping can be defined by changing the estimation method of the *posterior* probabilities.

3 Discriminant Kernel Functions

3.1 Dual Problem of ONDA

In the KDA, usually the kernel function is defined a priori. The polynomial functions or the Radial Basis functions are often used as the kernel functions but such kernel functions are general and are not related to the discrimination. Thus the class information is usually not introduced in these kernel functions. Also it is not known what the optimum kernel function for nonlinear discriminant analysis is.

Kurita showed the optimum kernel function, called discriminant kernel function (DKF), can be derived by investigating the dual problem of the eigenvalue problem of ONDA [7]. The DKF is also optimum in terms of the discriminant criterion.

The eigenvalue problem of ONDA given by the equation (7) is the generalized eigenvalue problem. By multiplying $P^{-1/2}$ from the left, this eigen equations can be rewritten as the usual eigenvalue problem as

$$P^{-1/2}\Gamma P^{-1/2}P^{1/2}U = P^{1/2}U\Lambda. \tag{9}$$

By denoting $\tilde{U} = P^{1/2}U$, we have the following usual eigenvalue problem as

$$(P^{-1/2}\Gamma P^{-1/2})\tilde{U} = \tilde{U}\Lambda. \tag{10}$$

Then the optimum nonlinear discriminant mapping of ODNA is rewritten as

$$\boldsymbol{y} = U^T\tilde{\boldsymbol{B}}(\boldsymbol{x}) = \tilde{U}^T P^{-1/2}\tilde{\boldsymbol{B}}(\boldsymbol{x}) = \tilde{U}^T\boldsymbol{\phi}(\boldsymbol{x}) \tag{11}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = P^{-1/2}\tilde{\boldsymbol{B}}(\boldsymbol{x})$ and $\tilde{\boldsymbol{B}}(\boldsymbol{x}) = (P(C_1|\boldsymbol{x}) - P(C_1), \ldots, P(C_K|\boldsymbol{x}) - P(C_K))^T$.

For the case of N training samples, the eigenvalue problem to determine the class representative vectors (10) is given by

$$(\Phi^T\Phi)\tilde{U} = \tilde{U}\Lambda, \tag{12}$$

where $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_N))^T$.

The dual eigenvalue problem of (12) is then given by

$$(\Phi\Phi^T)V = V\Lambda. \tag{13}$$

From the relation on the singular value decomposition of the matrix Φ, these two eigenvalue problems (12) and (13) have the same eigenvalues and there is the following relation between the eigenvectors $\tilde{U}$ and V as $\tilde{U} = \Phi^T V\Lambda^{-1/2}$.

By inserting this relation into the nonlinear discriminant mapping (11), we have

$$\boldsymbol{y} = \Lambda^{-1/2}V^T\Phi\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{N}\Lambda^{-1/2}\boldsymbol{v}_i\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) = \sum_{i=1}^{N}\boldsymbol{\alpha}_i K(\boldsymbol{x}_i, \boldsymbol{x}) - \boldsymbol{\alpha}_0 \tag{14}$$

where

$$\begin{aligned} K(\boldsymbol{x}_i, \boldsymbol{x}) &= \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) + 1 \\ &= \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x}_i) - P(C_k)(P(C_k|\boldsymbol{x}) - P(C_k))}{P(C_k)} + 1 \\ &= \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x}_i)P(C_k|\boldsymbol{x})}{P(C_k)}. \end{aligned} \tag{15}$$

This shows that the kernel function of the optimum nonlinear discriminant mapping is given by

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} \frac{P(C_k|\boldsymbol{x})P(C_k|\boldsymbol{y})}{P(C_k)}. \tag{16}$$

This is called the discriminant kernel function (DKF).

The derived DKF is defined by using the Bayesian *posterior* probabilities $P(C_k|\boldsymbol{x})$. This means that the class information is explicitly introduced in this kernel function. Also there is no kernel parameters. This means that we do not need to estimate the kernel parameters.

4 Sparse LDA Based on L1 DKF

There are many ways to estimate the Bayesian *posterior* probabilities. Depending on the estimation method, we can define the corresponding DKF [6][7].

In this paper, we propose L1-regularized discriminant kernel function which is defined by using the Bayesian posterior probability obtained from L1-regularized regression. We use L1 DKF as the kernel function for LDA.

4.1 L1-Regularized Regression

Given training samples $\{\mathbf{x}_n, t_n\}_{n=1}^N$ where $\mathbf{x}_n$ is n-th observation and t_n is the corresponding target value of $\mathbf{x}_n$, the objective of regression is to estimate the value t for a new data $\mathbf{x}$.

The simplest estimation model is given as the linear model:

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} \tag{17}$$

where $\mathbf{x} = (1, x_1, \cdots, x_D)$, $\mathbf{w} = (w_0, w_1, \cdots, w_D)$ and $y(\mathbf{x}, \mathbf{w})$ is the predicted value of t. An appropriate cofficient vector $\mathbf{w}$ is obtained by minimizing a certain error function $E_D(\mathbf{w})$. A sum-of-squares error function is commonly used:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2. \tag{18}$$

To control over-fitting, we can add a regularization term $E_W(\mathbf{w})$ with a regularization parameter λ into the error function. Given general regularizer $E_W(\mathbf{w}) = \sum_{j=1}^{M}|w_j|^q$, we obtain a regularized error function,

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2 + \frac{1}{\lambda}\sum_{j=1}^{M}|w_j|^q. \tag{19}$$

The case of $q = 1$ is known as the L1-regularized regression [12]. L1 regression has the property that if $\lambda(> 0)$ is sufficiently large, some of the coefficients w_j

Table 1. Classification accuracy for the test set (agerage of 10 trials)

dataset	bre	dna	ger	hea	iri	seg	sem	spl	veh	win
# of classes	2	3	2	2	3	7	10	2	4	3
# of samples	683	2586	1000	270	4	2310	1593	3175	846	178
# of features	10	180	24	13	150	19	256	60	18	13
LDA	96.3%	93.0%	72.1%	83.9%	86.1%	88.9%	81.9%	83.9%	76.3%	98.5%
sparse LDA	96.4%	93.5%	72.8%	85.2%	89.6%	89.9%	85.7%	84.4%	77.3%	99.3%

are driven to zero. It leads to a sparse model in which the corresponding basis functions play no role.

for K class classification problems, L1 regression can be used as the Baysian *posterior* probability estimator by the *one-vs-all* manner. Let $\mathbf{x}_n$ denote n-th d-dimensional feature vector $(n = 1, \cdots, N)$. For all $k = 1, \cdots, K$, the regression about *k-th class vs other classes* is performed between independent varible $\mathbf{x}_n$ and following dependent variable t_n^k,

$$t_n^k = \begin{cases} 1 & \text{if sample } \mathbf{x}_n \text{ belongs to class } k, \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

After K times regression, we obtain summarized projection vector of $\mathbf{x}_n$,

$$\mathbf{p}(\mathbf{x}_n) = (p_1, \cdots, p_K) = (\mathbf{w}_1^T \mathbf{x}_n, \cdots, \mathbf{w}_K^T \mathbf{x}_n) \tag{21}$$

where $\mathbf{w}_k$ is k-th projection cofficients. Ideally, if $\mathbf{x}_n$ belongs to class k, the k-th component of $\mathbf{p}$ should be 1, and the others should be 0. By normalizing $\mathbf{p}(\mathbf{x}_n)$ to satisfy the condition $\forall k(p_k \geq 0)$ and $p_1 + \cdots + p_K = 1$, we can consider p_k as the estimation of the Bayesian *posterior* probabilities $P(C_k|\boldsymbol{x})$.

4.2 L1 DKF and Sparse LDA

In this paper, we use L1 regression for the estimator of the Baysian *posterior* probability in the K class problems. For the input vector $\mathbf{x}$, the regression outputs probabilistic vector $(p_1, \cdots, p_K)$ in Eq. (21) as the estimation of the Bayesian *posterior* probabilities $(P(C_1|\mathbf{x}), \cdots, P(C_K|\mathbf{x}))$.

Then the corresponding discriminant kernel function, L1 DKF, is given as

$$K(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} \frac{p_k(\mathbf{x}) p_k(\mathbf{y})}{p(C_k)}. \tag{22}$$

We use L1 DKF as the kernel function of LDA to introduce the sparseness into LDA indirectly.

5 Experiments

We confirmed the performance of L1 DKF for the kernel of LDA, by using several data sets in UCI machine learning repository [5]: Breast-cancer (bre),

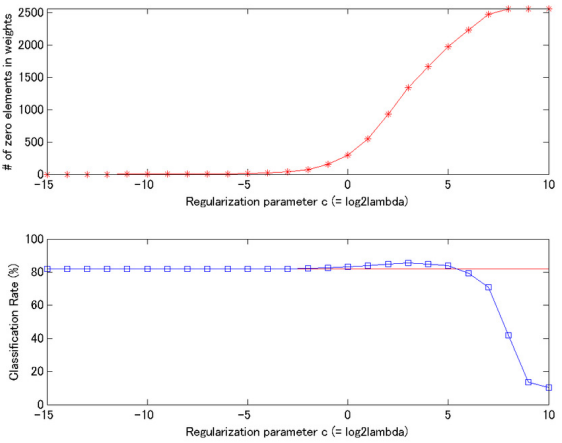

Fig. 1. Results for semeion data. The top figure shows the sparseness of the regression coefficients. The bottom figure shows the classification rate for the test set. The horizontal line shows LDA's performance. The curve shows sparse LDA's performance. Both graphs show the average of 10 trials' results.

dna, german (ger), heart (hea), iris (iri), segment (seg), semeion (sem), splice (spl), vehicle (veh) and wine (win) data. Each data set was divided into a training set (2/3 of all samples) and a test set (remaining samples), at random. For classification experiments, we made 10 different divisions of the training and test sets. For all experiments, we used class prior $P(C_k) = N_k/N$ where N_k is the number of samples in C_k. We use a nearest mean classifier for usual LDA and sparse LDA.

We train L1 DKF by using different regularization parameter $\lambda = 2^{-15}, 2^{-14}, \cdots, 2^{10}$. Fig. 1 shows the training result for the semeion data. Note that the figure shows the average of results of 10 trials.

The top figure shows the sparseness of the L1 DKF. The semeion data has 10 classes and 256 features, therefore the summarized projection matrix $[\mathbf{w}_1, \cdots, \mathbf{w}_K]$ has totally 2560 elements. The vertical axis shows the number of zero elements in 2560 elements. The number of zero elements is increasing in proportion to the regularization parameter λ.

The bottom figure shows the classification accuracy for the test set. As the baseline performance, LDA has 81.9% accuracy. The accuracy of sparse LDA is better than LDA in some part. The highest averaged accuracy of sparse LDA is 85.7 % ($\lambda = 2^3$). In this case, about 1,400 features in 2,560 original features did not be used in the classification task. It is considered that the features which are not suitable for the classification task were removed by L1 regression.

Tab.1 shows the classification performances of LDA and sparse LDA for each data set. In all cases, the highest performance of sparse LDA was better than the performance of LDA.

6 Conclusions

In this paper, we propose a novel discriminant kernel function based on L1 regression (called L1 DKF), and we use it for the kernel of LDA to introduce the sparseness into LDA. In experiments, we show L1 DKF is appropriate as the kernel for LDA. Our sparse LDA has better classification performance than usual LDA.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 23500211.

References

1. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Computation 12(10), 2385–2404 (2000)
2. Chow, C.K.: An optimum character recognition system using decision functions. IRE Trans. EC-6, 247–254 (1957)
3. Clemmensen, L., Hastie, T., Witten, D., Ersboll, B.: Sparse discriminant analysis (2011)
4. Fisher, R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7, 179–188 (1936)
5. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, `http://archive.ics.uci.edu/ml`
6. Hidaka, A., Kurita, T.: Discriminant Kernels based Support Vector Machine. In: The First Asian Conference on Pattern Recognition (ACPR 2011), Beijing, China, November 28-30, pp. 159–163 (2011)
7. Kurita, T.: "Discriminant Kernels derived from the Optimum Nonlinear Discriminant Analysis. In: Proc. of 2011 International Joint Conference on Neural Networks, San Jose, California, USA, July 31-August 5 (2011)
8. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Smola, A., Muller, K.: Fisher discriminant analysis with kernels. In: Proc. IEEE Neural Networks for Signal Processing Workshop, pp. 41–48 (1999)
9. Otsu, N.: Nonlinear discriminant analysis as a natural extension of the linear case. Behavior Metrika 2, 45–59 (1975)
10. Otsu, N.: Mathemetical Studies on Feature Extraction In Pattern Recognition. Researches on the Electrotechnical Laboratory 818 (1981) (in Japanease)
11. Otsu, N.: Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In: Proceedings of the 6th International Conference on Pattern Recognition, pp. 557–560 (1982)
12. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B. 58(1), 267–288 (1996)
13. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics 15(2), 262–286 (2006)