

Optimal Regularization Parameter Estimation for Spectral Regression Discriminant Analysis

Wei Chen, Caifeng Shan, *Member, IEEE*, and Gerard de Haan, *Senior Member, IEEE*

Abstract—Spectral regression discriminant analysis (SRDA) is an efficient subspace learning method proposed recently. One important unsolved issue of SRDA is how to automatically determine an appropriate regularization parameter. In this letter, we present a method to estimate the optimal regularization parameter for SRDA. We test our method in different applications including head pose estimation, face recognition, and text categorization. Our extensive experiments evidently illustrate the effectiveness and efficiency of our approach.

Index Terms—Regularization parameter estimation, spectral regression discriminant analysis, subspace learning.

I. INTRODUCTION

SPECTRAL regression discriminant analysis (SRDA) [1]–[3] is an efficient subspace learning method proposed recently. By casting projective function learning into a regression framework, it avoids eigen-decomposition of dense matrices. Compared to the existing subspace learning algorithms [4] with cubic-time complexity, SRDA can be implemented in linear time. SRDA has shown promising performance in different applications including face recognition [2], text clustering and categorization [1], spoken letter recognition [3], and handwritten digit classification [3].

One unsolved issue of SRDA is how to automatically determine an appropriate regularization parameter α [1]. The parameter α , which was empirically set in the existing work, controls the smoothness of the estimator. Experiments in [1]–[3] imply that the performance of SRDA is closely related to the choice of α . Therefore, estimating an optimal α is an essential problem for SRDA. In this letter, by formulating the problem as a constrained optimization problem, we present a method to estimate the optimal regularization parameter for SRDA. Compared to the existing regularization parameter estimation methods including general cross-validation (GCV) [5] and the L-curve [6], our approach is much more efficient, and provides more accurate estimation. We test our method in different applications including head

pose estimation, face recognition, and text categorization. Our extensive experiments evidently illustrate the effectiveness and efficiency of our approach. The rest of the letter is organized as follows. Section II gives a brief introduction on SRDA. Our approach is described in Section III. We present the experiments in Section IV, and finally Section V concludes the letter.

II. SPECTRAL REGRESSION DISCRIMINANT ANALYSIS

Given a data set $\{\mathbf{x}_i\}_{i=1}^m$ in \mathbb{R}^n , dimensionality reduction methods aim to find a low-dimensional representation of $\{\mathbf{x}_i\}$. In the graph-based methods [4], a symmetric matrix $W (= [w_{ij}]_{m \times m})$ is built, where w_{ij} measures the similarity between \mathbf{x}_i and \mathbf{x}_j . Let $\mathbf{y} = [y_1, \dots, y_m]^T$ be the 1-D projection of $X = [\mathbf{x}_1, \dots, \mathbf{x}_m]$, the optimal \mathbf{y} is given by minimizing [7]

$$\sum_{i,j} (y_i - y_j)^2 w_{ij} = 2\mathbf{y}^T (D - W)\mathbf{y} = 2\mathbf{y}^T L\mathbf{y} \quad (1)$$

where D is a diagonal matrix whose entries are column (or row) sums of W . A constraint $\mathbf{y}^T D\mathbf{y} = 1$ can be imposed [7], and the minimization problem reduces to finding the optimal \mathbf{y}^*

$$\mathbf{y}^* = \arg \min_{\mathbf{y}^T D\mathbf{y}=1} \mathbf{y}^T L\mathbf{y} = \arg \min_{\mathbf{y}} \frac{\mathbf{y}^T L\mathbf{y}}{\mathbf{y}^T D\mathbf{y}} = \arg \max_{\mathbf{y}} \frac{\mathbf{y}^T W\mathbf{y}}{\mathbf{y}^T D\mathbf{y}} \quad (2)$$

which is solved as the maximum eigen-problem

$$W\mathbf{y} = \lambda D\mathbf{y}. \quad (3)$$

To obtain a projective mapping for all samples, including new testing samples, a linear function $y_i = f(\mathbf{x}_i) = \mathbf{a}^T \mathbf{x}_i$ is chosen, i.e., $\mathbf{y} = X^T \mathbf{a}$, (3) can be reduced to the maximum eigen-problem

$$XWX^T \mathbf{a} = \lambda XDX^T \mathbf{a}. \quad (4)$$

With different choices of W , the above framework leads to different subspace learning methods. A common problem of these methods is the high computational cost from the eigen-decomposition of dense matrices. To address this problem, Cai *et al.* [1]–[3] introduced SRDA which, instead of solving the eigen-problem in (4), derives the linear projective functions via two steps.

Manuscript received October 01, 2008; revised February 27, 2009. First version published July 7, 2009; current version published December 1, 2009. This paper was recommended by Associate Editor Y. Rui.

The authors are with Philips Research, 5656 AE Eindhoven, The Netherlands (e-mail: w.chen@philips.com; caifeng.shan@philips.com; g.de.haan@philips.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCSVT.2009.2026953

- 1) Solve the eigen-problem in (3) to get \mathbf{y} .
- 2) Find \mathbf{a} which satisfies $X^T \mathbf{a} = \mathbf{y}$. In reality, such \mathbf{a} might not exist. A possible solution is to find \mathbf{a} which best fits the equation in the least squares sense

$$\mathbf{a}^* = \arg \min_{\mathbf{a}} \sum_{i=1}^m (\mathbf{a}^T \mathbf{x}_i - y_i)^2. \quad (5)$$

In the first step, SRDA constructs the weight matrix W by incorporating the label information. Suppose c classes in the data set and m_t samples in the t th class, i.e., $m_1 + \dots + m_c = m$, W is defined as

$$w_{ij} = \begin{cases} 1/m_t, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ both belong to the } t\text{th class} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

In the second step, the minimization problem in (5) is usually ill-posed in reality. Instead of using maximum likelihood estimation, which leads to the ordinary least squares (OLS) estimator

$$\hat{\mathbf{a}} = (X X^T)^{-1} X \mathbf{y} \quad (7)$$

SRDA adopts the regularization technique to obtain the regularized estimator

$$\hat{\mathbf{a}}^* = (X X^T + \alpha I)^{-1} X \mathbf{y} \quad (8)$$

where α (≥ 0) is a regularization parameter to control the smoothness of the estimator $\hat{\mathbf{a}}^*$.

As illustrated in [1]–[3], the performance of SRDA varies greatly as α changes; a nonzero α was empirically set in these experiments. An inappropriate setting of α may result in poor performance in practice. Before we present an approach for estimating α , we briefly discuss two existing methods for regularization parameter estimation. Generalized cross-validation (GCV) [5] is based on statistical consideration that a good regularization parameter should predict the missing data. If an arbitrary data point y_i of \mathbf{y} is left out, the corresponding regularized solution $\hat{\mathbf{a}}^*$ should be able to predict it correctly. GCV has the time complexity of $O(m^3)$. The L-curve method [6] is based on a log–log plot of the norm of a regularized solution $\|\hat{\mathbf{a}}^*\|_2$ versus the norm of the residual $\|X^T \hat{\mathbf{a}}^* - \mathbf{y}\|_2$. The α corresponding to the corner of the L-shaped plot suggests a solution wherein both the solution norm and the residual norm attain low values. The corner is derived by examining the curvature of points, and the L-curve method has the time complexity of $O(n^3)$.

III. REGULARIZATION PARAMETER ESTIMATION FOR SRDA

The difference between the regularized estimator $\hat{\mathbf{a}}^*$ in (8) and the OLS estimator $\hat{\mathbf{a}}$ in (7) can be analyzed using singular value decomposition (SVD). Suppose X is a wide matrix ($m > n$), we have $X^T = USV^T$, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are unitary matrices, and $S \in \mathbb{R}^{m \times n}$ is the singular

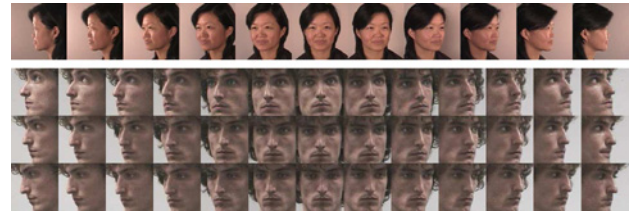


Fig. 1. Example images in the FacePix database (top) and the Pointing '04 database (bottom).

value matrix with the rank of r ($r \leq n$). The solution $\hat{\mathbf{a}}^*$ in (8) can be reduced as

$$\hat{\mathbf{a}}^* = (VS^T U^T USV^T + \alpha VV^T)^{-1} VS^T U^T \mathbf{y} \quad (9)$$

$$= V (S^T S + \alpha I)^{-1} S^T U^T \mathbf{y} \quad (10)$$

$$= \sum_{i=1}^n \mathbf{v}_i (s_i^2 + \alpha)^{-1} s_i \mathbf{u}_i^T \mathbf{y} \quad (11)$$

$$= \sum_{i=1}^n \mathbf{v}_i \left(\frac{\mathbf{u}_i^T \mathbf{y}}{s_i} \cdot \frac{s_i^2}{s_i^2 + \alpha} \right) \quad (12)$$

where \mathbf{u}_i and \mathbf{v}_i denote the orthonormal column vectors in U and V respectively, s_i represents the i th largest singular value of X^T (when $i > r$, $s_i = 0$), and \mathbf{y} is the constant response calculated in (3). $\frac{s_i^2}{s_i^2 + \alpha} \in [0, 1]$ is called *filter factor* in [6]. Similarly, the solution $\hat{\mathbf{a}}$ in (7) can be reduced as

$$\hat{\mathbf{a}} = \sum_{i=1}^r \mathbf{v}_i \left(\frac{\mathbf{u}_i^T \mathbf{y}}{s_i} \right). \quad (13)$$

By comparing (12) and (13), we can find that both $\hat{\mathbf{a}}^*$ and $\hat{\mathbf{a}}$ are linear combinations of basis vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, and the regularization technique in SRDA changes only the coefficients of the linear combination by scaling with a filter factor. The coefficients can be seen as functions of the singular values of X and the regularization parameter α .

In order to estimate the optimal α , we first investigate the constraint on α itself. SRDA is solved as the multivariate linear regression problem

$$X^T \mathbf{a} + \varepsilon = \mathbf{y} \quad (14)$$

where ε is an $(n \times 1)$ vector of random error with $E[\varepsilon] = 0$ and $\text{Var}[\varepsilon] = \sigma^2 I_n$. A good regularization parameter α should reduce the mean square error (MSE) of the regularized estimator $\hat{\mathbf{a}}^*$. In order to evaluate the MSE of $\hat{\mathbf{a}}^*$ with respect to α , it is necessary to derive $E[D^2(\alpha)]$, where $D(\alpha)$ denotes the distance from $\hat{\mathbf{a}}^*$ to \mathbf{a} . For the OLS estimator $\hat{\mathbf{a}}$, we have

$$\hat{\mathbf{a}} = \mathbf{a} + (X X^T)^{-1} X \varepsilon \quad (15)$$

$$E[\hat{\mathbf{a}}] = \mathbf{a}. \quad (16)$$

From (7) and (8), we can obtain the relationship between $\hat{\mathbf{a}}$ and $\hat{\mathbf{a}}^*$ as follows:

$$\begin{aligned} \hat{\mathbf{a}}^* &= (X X^T)(X X^T + \alpha I)^{-1} \hat{\mathbf{a}} \\ &= (I - \alpha(X X^T + \alpha I)^{-1}) \hat{\mathbf{a}} \\ &=: R \hat{\mathbf{a}} \end{aligned} \quad (17)$$

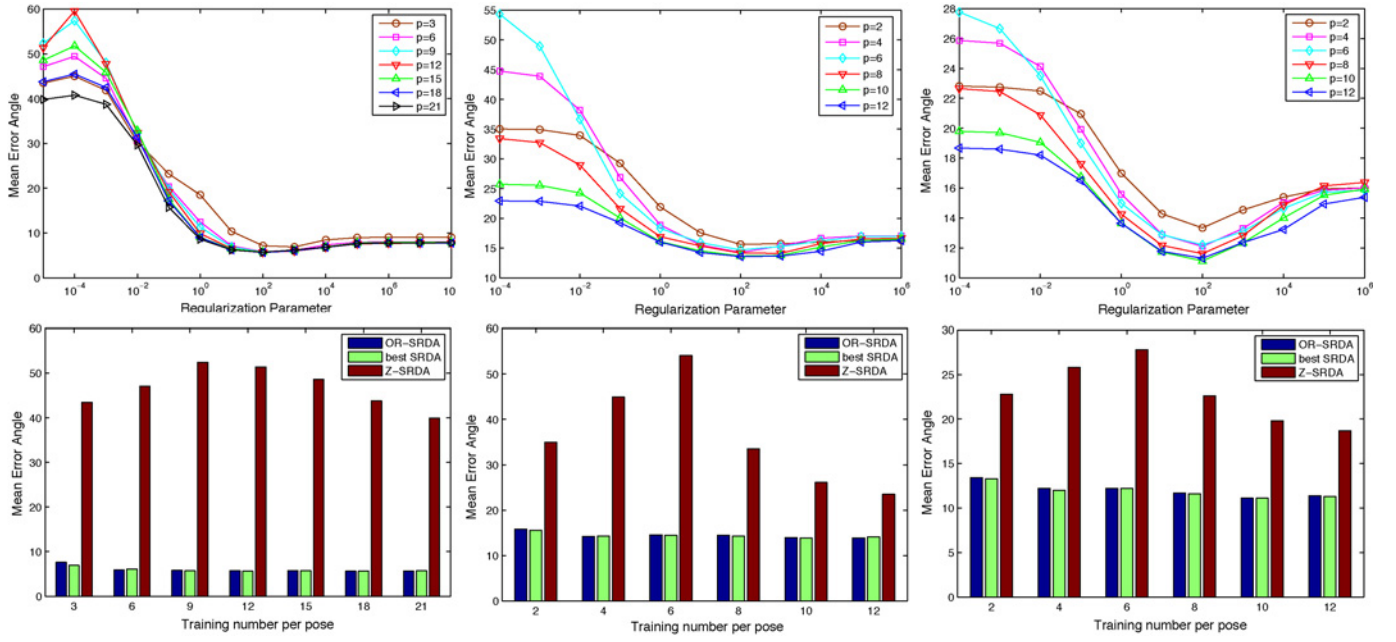


Fig. 2. Average estimation performance of SRDA with respect to different α (top row) and comparisons of different methods (bottom row). Left: FacePix; middle: Pointing '04-Yaw; right: Pointing '04-Pitch.

where R is used for simplicity. Hence we have

$$E[D^2(\alpha)] = E[(\hat{\mathbf{a}}^* - \mathbf{a})^T (\hat{\mathbf{a}}^* - \mathbf{a})] \quad (18)$$

$$= E[(R\hat{\mathbf{a}} - R\mathbf{a} + R\mathbf{a} - \mathbf{a})^T (R\hat{\mathbf{a}} - R\mathbf{a} + R\mathbf{a} - \mathbf{a})] \quad (19)$$

$$= E[(\hat{\mathbf{a}} - \mathbf{a})^T R^T R (\hat{\mathbf{a}} - \mathbf{a})] + (R\mathbf{a} - \mathbf{a})^T (R\mathbf{a} - \mathbf{a}). \quad (20)$$

Substituting (15) in the first term of (20), we obtain

$$E[(\hat{\mathbf{a}} - \mathbf{a})^T R^T R (\hat{\mathbf{a}} - \mathbf{a})] = E[\boldsymbol{\varepsilon}^T X^T (XX^T)^{-1} R^T R (XX^T)^{-1} X \boldsymbol{\varepsilon}] \quad (21)$$

$$= \text{Tr}(X^T (XX^T)^{-1} R^T R (XX^T)^{-1} X \text{Var}[\boldsymbol{\varepsilon}]) + E[\boldsymbol{\varepsilon}] \text{Tr}(X^T (XX^T)^{-1} R^T R (XX^T)^{-1} X) E[\boldsymbol{\varepsilon}] \quad (22)$$

$$= \sigma^2 \text{Tr}((XX^T)^{-1} R^T R). \quad (23)$$

With (17), we then have

$$E[D^2(\alpha)] = \sigma^2 \text{Tr}((XX^T)^{-1} R^T R) + \mathbf{a}^T (R - I)^T (R - I) \mathbf{a} \quad (24)$$

$$= \sigma^2 \text{Tr}((XX^T + \alpha I)^{-1} (I - \alpha (XX^T + \alpha I)^{-1})) + \mathbf{a}^T (\alpha^2 (XX^T + \alpha I)^{-2}) \mathbf{a} \quad (25)$$

$$= \sigma^2 (\text{Tr}(XX^T + \alpha I)^{-1} - \alpha \text{Tr}(XX^T + \alpha I)^{-2}) + \alpha^2 \mathbf{a}^T (XX^T + \alpha I)^{-2} \mathbf{a}. \quad (26)$$

Let $\mathbf{c} = [c_1, c_2, \dots, c_n]^T$, which satisfies $\mathbf{c} = V^T \mathbf{a}$, we obtain

$$E[D^2(\alpha)] = \sigma^2 \left(\sum_{i=1}^n \frac{1}{(s_i^2 + \alpha)} - \sum_{i=1}^n \frac{\alpha}{(s_i^2 + \alpha)^2} \right) + \alpha^2 \mathbf{c}^T (S^T S + \alpha I)^{-2} \mathbf{c} \quad (27)$$

$$= \sigma^2 \sum_{i=1}^n \frac{s_i^2}{(s_i^2 + \alpha)^2} + \alpha^2 \sum_{i=1}^n \frac{c_i^2}{(s_i^2 + \alpha)^2}. \quad (28)$$

It is obvious that, for any $\alpha > 0$, the first and second terms in (28) are monotonically decreasing and increasing functions of α respectively. Taking the derivative of (28) with respect to α , we have

$$\frac{\partial E[D^2(\alpha)]}{\partial \alpha} = 2 \sum_{i=1}^n \frac{s_i^2 (\alpha c_i^2 - \sigma^2)}{(s_i^2 + \alpha)^3}. \quad (29)$$

Now we can see that

$$\frac{\partial E[D^2(\alpha)]}{\partial \alpha} < 0, \quad \text{for } 0 < \alpha < \min \left\{ \frac{\sigma^2}{c_i^2}, \forall i \right\} \quad (30)$$

and

$$\frac{\partial E[D^2(\alpha)]}{\partial \alpha} > 0, \quad \text{for } \max \left\{ \frac{\sigma^2}{c_i^2}, \forall i \right\} < \alpha < \infty. \quad (31)$$

Thus, the minimum of MSE falls in the interval of α

$$\left[\min \left\{ \frac{\sigma^2}{c_i^2} \right\}, \max \left\{ \frac{\sigma^2}{c_i^2} \right\} \right] \quad \forall i. \quad (32)$$

Therefore, the optimal α should be neither too large nor too small.

The criteria we consider for estimating α is the robustness of the regularized estimator $\hat{\mathbf{a}}^*$ to noises in the data X . More precisely, $\hat{\mathbf{a}}^*$ with respect to the optimal α should be robust to the perturbation in the parameter space of the singular values $\{s_i\}$ of X , since $\hat{\mathbf{a}}^*$ can be seen as a function of α and $\{s_i\}$. In this way, estimating the optimal α is reduced to solving the minimization problem

$$\alpha^* = \arg \min_{\alpha} E[\|\hat{\mathbf{a}}^*(\alpha, s + \epsilon) - \hat{\mathbf{a}}^*(\alpha, s)\|_2^2], \quad \text{s.t. } \alpha^* \in \left[\min \left\{ \frac{\sigma^2}{c_i^2} \right\}, \max \left\{ \frac{\sigma^2}{c_i^2} \right\} \right] \quad \forall i \quad (33)$$

where $\epsilon \sim \mathcal{N}(0, \delta^2)$ is the perturbation in the parameter space. Since

$$\|\hat{\mathbf{a}}^*(\alpha, s + \epsilon) - \hat{\mathbf{a}}^*(\alpha, s)\|_2^2 = (\hat{\mathbf{a}}^*(\alpha, s + \epsilon) - \hat{\mathbf{a}}^*(\alpha, s))^T (\hat{\mathbf{a}}^*(\alpha, s + \epsilon) - \hat{\mathbf{a}}^*(\alpha, s)) \quad (34)$$

$$= \sum_{i=1}^n \sum_{j=1}^n \mathbf{y}^T \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_j \mathbf{u}_j^T \mathbf{y} \left(\frac{s_i}{s_i^2 + \alpha} - \frac{s_i + \epsilon}{(s_i + \epsilon)^2 + \alpha} \right)^2 \quad (35)$$

where $\mathbf{u}_i \mathbf{u}_j^T = 0$, $\mathbf{v}_i^T \mathbf{v}_j = 0$, when $i \neq j$, and $\mathbf{u}_i^T \mathbf{u}_j = 1$, $\mathbf{v}_i^T \mathbf{v}_j = 1$, when $i = j$, and only the terms with $i = j$ remain in (35). So we have

$$\alpha^* = \arg \min_{\alpha} E \left[\sum_{i=1}^n \mathbf{y}^T \mathbf{y} \left(\frac{s_i}{s_i^2 + \alpha} - \frac{s_i + \epsilon}{(s_i + \epsilon)^2 + \alpha} \right)^2 \right]. \quad (36)$$

Note that $\mathbf{y}^T \mathbf{y}$ is a positive constant, where \mathbf{y} is calculated from (3). Equation (36) is equivalent to

$$\alpha^* = \arg \min_{\alpha} E \left[\sum_{i=1}^n \left(\frac{(\alpha - s_i^2)\epsilon - s_i \epsilon^2}{(s_i^2 + \alpha)((s_i + \epsilon)^2 + \alpha)} \right)^2 \right]. \quad (37)$$

Considering ϵ is very small, we neglect the term $s_i \epsilon^2$. Thus, we have

$$\alpha^* = \arg \min_{\alpha} \sum_{i=1}^n \left(\frac{\alpha - s_i^2}{(s_i^2 + \alpha)((s_i + \epsilon)^2 + \alpha)} \right)^2 E(\epsilon^2) \quad (38)$$

$$= \arg \min_{\alpha} \sum_{i=1}^n \frac{(\alpha - s_i^2)^2}{((s_i^2 + \alpha)((s_i + \epsilon)^2 + \alpha))^2} \cdot \delta^2 \quad (39)$$

$$= \arg \min_{\alpha} \sum_{i=1}^n \frac{(\alpha - s_i^2)^2}{\rho(\alpha)^2} \quad (40)$$

where $\rho(\alpha) := (s_i^2 + \alpha)((s_i + \epsilon)^2 + \alpha)$. Now the minimization problem in (33) can be rewritten as follows:

$$\alpha^* = \arg \min_{\alpha} \sum_{i=1}^n \frac{(\alpha - s_i^2)^2}{\rho(\alpha)^2}, \quad \text{s.t. } \alpha^* \in \left[\min \left\{ \frac{\sigma_i^2}{c_i^2} \right\}, \max \left\{ \frac{\sigma_i^2}{c_i^2} \right\} \right] \quad \forall i. \quad (41)$$

It is difficult to solve the minimization problem in (41) analytically, as α^* and c_i are coupled. Considering the intrinsic bound of α and $\rho(\alpha)$, the problem is relaxed to a simple form

$$\alpha^* = \arg \min_{\alpha} \sum_{i=1}^n (\alpha - s_i^2)^2. \quad (42)$$

By setting the derivative with respect to α equal to 0, we obtain the solution

$$\alpha^* = \frac{1}{n} \sum_i s_i^2. \quad (43)$$

In practice, the above solution can be computed without extra computational cost, because $\{s_1^2, \dots, s_n^2\}$ are the eigenvalues of the symmetric matrix XX^T and their sum is the trace of XX^T , which has been calculated in SRDA.

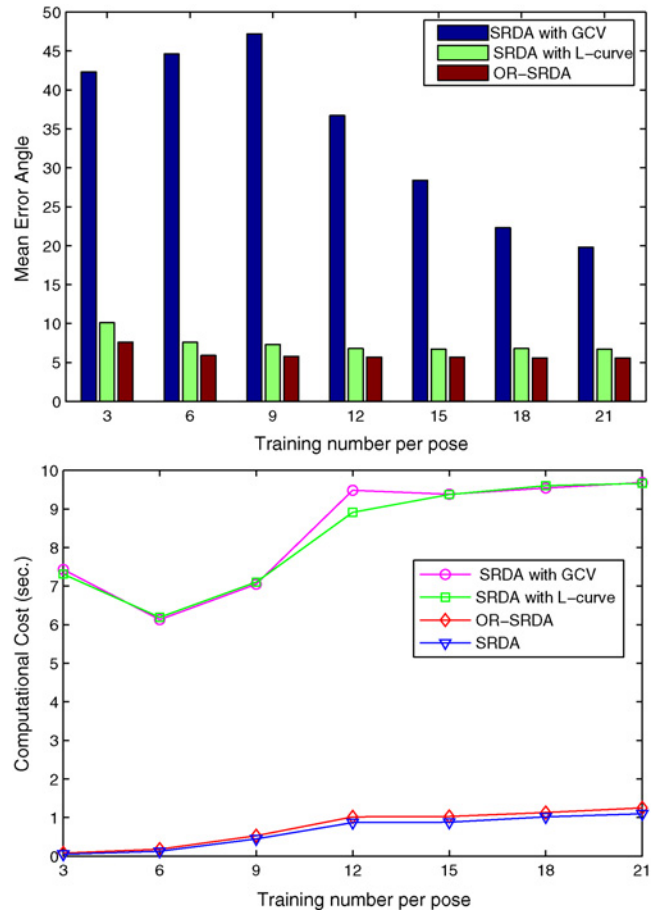


Fig. 3. Comparison of different regularization parameter estimation methods on the FacePix database: (top) the performance of SRDA with α estimated by three methods; (bottom) the computational cost of three methods compared to SRDA.

IV. EXPERIMENTS

We carried out experiments on head pose estimation [8]–[10], face recognition, and text categorization. All these tasks are treated as the multiclass recognition problem. For simplicity, the nearest-neighbor classifier with the L2-norm metric is adopted for classification in these experiments.

A. Head Pose Estimation

The FacePix database [11] and Pointing '04 database [12] were used in our experiments. The FacePix database contains images from 30 subjects. For each subject, we selected 91 images, representing pose angles from -90° to $+90^\circ$ at increments of 2° . The Pointing '04 database contains 15 subjects, each of which has images at different poses, including 13 yaw poses and seven pitch poses, plus two extreme cases with yaw angle 0° and pitch angle $90^\circ/-90^\circ$. We used all these images in our experiments. Some example images of these two databases are shown in Fig. 1. In our experiments, each image was normalized to 32×32 pixels in gray-scale space, thus represented as a 1024-dimensional vector. p subjects ($p = 3, 6, 9, 12, 15, 18, 21$ in the FacePix database and $p = 2, 4, 6, 8, 10, 12$ in the database) were randomly selected for training, and the rest were used for testing. For each p , we averaged the estimation error over 30 random splits.

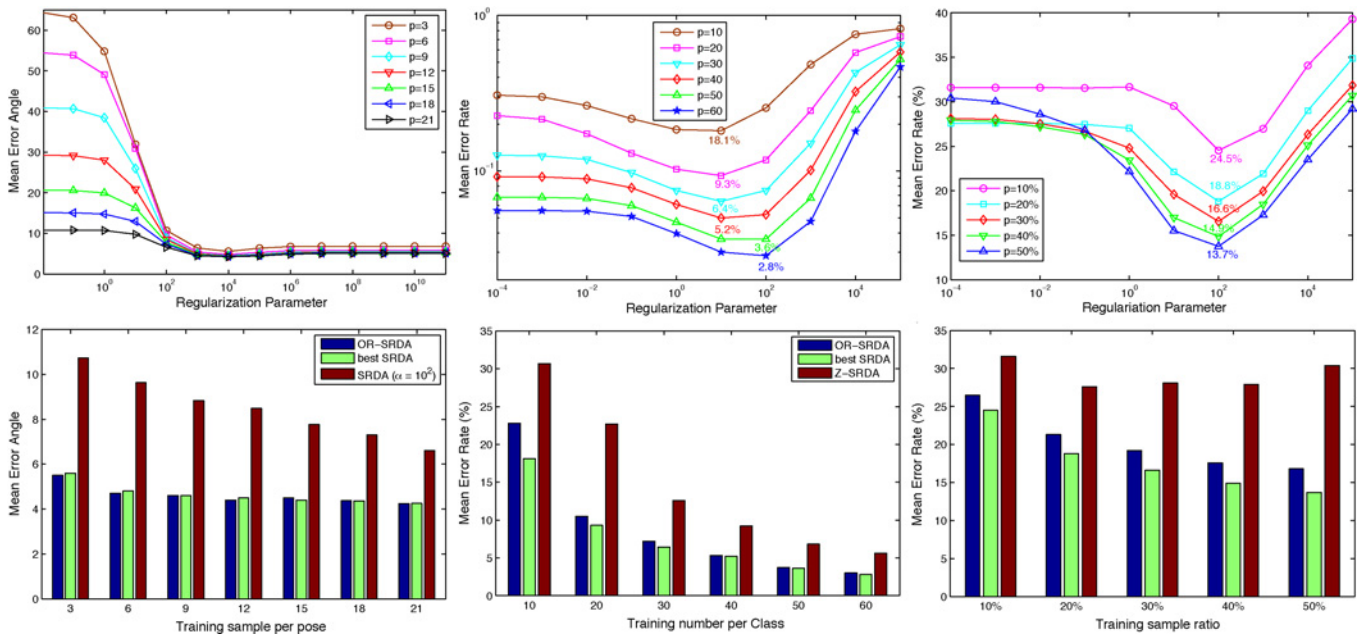


Fig. 4. Average performance of SRDA with respect to different α (top row) and comparison of different methods (bottom row). Left: head pose estimation using DCT coefficients on the FacePix database; middle: face recognition on the PIE database; right: text categorization on the 20Newsgroups database.

The pose estimation performance as the function of α is plotted in the top row of Fig. 2, where the reduced dimension is $c - 1$, and c is the number of different head poses, i.e., $c = 91$ for the FacePix database and $c = 93$ for the database. An exponentially incremental sampling of α is chosen to present the complete variation. It is evident that the performance of SRDA changes greatly as α varies. SRDA achieves significantly better performance when the projected space is smoothed (with $\alpha > 10^{-1}$) than that with α close to 0. There always exists an optimal α in all experiments, although the performance of SRDA does not change too much for large α on the FacePix and Pointing '04-Yaw data. The average performance of SRDA with the regularization parameter estimated using our method (denoted as OR-SRDA) is shown in the bottom row of Fig. 2, where the best performance of SRDA obtained by examining different α and the performance of SRDA with $\alpha = 0$ (denoted as Z-SRDA) are included for comparison. We can conclude from these comparisons that: 1) the regularization parameter is essential for SRDA, since SRDA with $\alpha = 0$ provides much worse performance than that with a proper α ; and 2) with the α estimated by our method, SRDA achieves very similar results to the best performance of SRDA by exhaustively examining different α . This illustrates the effectiveness of our method.

We compare our method with the GCV and L-curve methods on the FacePix database. Fig. 3 shows the performance of SRDA with the estimated parameters by different methods, and the corresponding computational cost.¹ These experiments were performed on a Linux PC (CPU 3.0 GHz, Cache 1024 kb, RAM 4 GB). It is observed that GCV fails to estimate the

appropriate regularization parameters for SRDA, while the L-curve performs much better, and our method outperforms both GCV and L-curve methods. Regarding the computational cost, our method is significantly more efficient than GCV and L-curve methods.

It is interesting that in the above experiments the best performance of SRDA was always achieved when $\alpha = 10^2$. To investigate this problem, we conducted further experiments on the FacePix database by adopting a different facial representation. Specifically, discrete cosine transform (DCT) coefficients are utilized [13]: each face image was down-scaled to 64×64 pixels in gray-scale space, and uniformly divided into 64 blocks of 8×8 pixels. The second till the sixth DCT coefficients of each block were used to construct a feature vector with the length of 320. With the same experiment settings as before, we obtain the experimental results shown in the left column of Fig. 4. As can be observed, the performance of SRDA with $\alpha = 10^4$ is much better than that with $\alpha = 10^2$, which implies that we cannot always empirically set α as 10^2 . Moreover, in these experiments, our OR-SRDA again achieved very similar performance to that of SRDA by exhaustively searching α . It further verifies the effectiveness of our method.

In the above experiments on head pose estimation, SRDA produces very poor performance with the small α , but, with the increased α , it achieves much better results; its performance becomes stable with larger α . This is possible because of the inherent continuity of head pose space, where neighboring pose classes are very much similar. With a larger α , the constructed head pose space is more smoothed, which helps to deal with appearance variance cross different subjects in the same head pose class. However, this is not always the case, as shown in experiments discussed in the next two sections.

¹We used the implementation of the GCV and L-curve methods in Regularization Tools (version 4.1), <http://www2.imm.dtu.dk/~pch/Regutools/>.

B. Face Recognition

Following [2], we conducted face recognition experiments using the CMU-PIE data online,² which consists of 68 subjects, with 170 images for each subject. p ($= 10, 20, 30, 40, 50, 60$) images of each subject were randomly selected for training, and the rest were used for testing. For each p , we averaged the recognition error rates over 30 random splits. The experimental results are shown in the middle column of Fig. 4, where the reduced dimension is $c - 1$ and $c = 68$ is the number of the subjects. Similarly, we observe that the regularization parameter α has great impact on the performance of SRDA. Compared to experiments on head pose estimation, the “oversmooth” effect due to a large α is more significant for face recognition. For example, the mean error rate with $\alpha = 10^3$ is much bigger than that with $\alpha = 1$. The experiments also illustrate that SRDA does not always achieve the best results with $\alpha = 10^2$. From the comparison between different methods, again we can see our OR-SRDA consistently achieves the performance close to the best performance of SRDA obtained by exhaustively searching α . Furthermore, their difference becomes much smaller with larger training data.

C. Text Categorization

We also performed experiments on text categorization using the 20Newsgroups data online,² which has 18 941 documents, evenly distributed across 20 classes. The data set was randomly split into training and testing sets. We ran tests with the training set containing 10%, 20%, 30%, 40%, and 50% documents, respectively. For each test, we averaged the error rate over 20 random splits. The experiment results are shown in the right column of Fig. 4, where the reduced dimension is $c - 1$ and $c = 20$ is the number of classes. The results reinforce the finding in the FacePix, Pointing '04, and CMU-PIE databases that the performance of SRDA varies greatly with the changes of α , and there always exists an optimal α . Again, we can see our method shows its effectiveness by estimating an appropriate α for SRDA. It is observed that the difference between the performance of OR-SRDA and the best performance of SRDA is slightly larger than that on the FacePix, Pointing '04, and PIE databases. This is possible because the 20Newsgroups data has a high feature dimension ($n = 26\,214$), a large sample number ($m = 18\,941$) but a small number of classes ($c = 20$).

V. CONCLUSION

How to determine an appropriate regularization parameter α is a crucial problem for SRDA. In this letter, we present an efficient approach to estimate the optimal α for SRDA. Our experiments on different databases (head poses, faces, texts) illustrate that our method can effectively estimate the regularization parameter for SRDA. Compared to the existing regularization parameter estimation methods, our approach substantially reduces the computational cost, and provides more accurate estimation.

REFERENCES

- [1] D. Cai, X. He, and J. Han, “Regularized locality preserving indexing via spectral regression,” in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2007, pp. 741–750.
- [2] D. Cai, X. He, and J. Han, “Spectral regression for efficient regularized subspace learning,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
- [3] D. Cai, X. He, and J. Han, “SRDA: An efficient algorithm for large scale discriminant analysis,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 1–12, Jan. 2008.
- [4] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: A general framework for dimensionality reduction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
- [5] G. Golub, M. Heath, and G. Wahba, “Generalized cross-validation as a method for choosing a good ridge parameter,” *Technometrics*, vol. 21, pp. 215–223, 1979.
- [6] P. C. Hansen, “Problems with ill-determined rank,” in *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. Philadelphia, PA: SIAM, 1998, ch. 4, pp. 69–98.
- [7] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, “Face recognition using laplacianfaces,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
- [8] Y. Fu and T. S. Huang, “Graph embedded analysis for head pose estimation,” in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 3–8.
- [9] J. Tu, Y. Fu, and T. S. Huang, “Locating nose-tips and estimating head poses in images by tensorposes,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 90–102, Jan. 2009.
- [10] S. Yan, H. Wang, Y. Fu, J. Yan, X. Tang, and T. S. Huang, “Synchronized submanifold embedding for person-independent pose estimation and beyond,” *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 202–210, Jan. 2009.
- [11] V. N. Balasubramanian, S. Krishna, and S. Panchanathan, “Person-independent head pose estimation using biased manifold embedding,” *Eur. Assoc. Signal Speech Image Process. J. Adv. Signal Process.*, vol. 8, no. 1, pp. 1–15, 2008.
- [12] N. Gourier and J. Letessier, “Estimating face orientation from robust detection of salient facial features,” in *Proc. Int. Workshop Vis. Observ. Deictic Gestures (ICPR)/Workshop Vis. Observation Deictic Gestures*, Cambridge, U.K., 2004, pp. 270–280.
- [13] H. K. Ekenel and R. Stiefelhagen, “Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization,” in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2006, pp. 34–41.

²<http://www.cs.uiuc.edu/homes/dengcai2/Data/data.html>