Neural Process Lett (2008) 27:17–24
DOI 10.1007/s11063-007-9056-7

A Gradient Linear Discriminant Analysis for Small Sample Sized Problem

Alok Sharma · Kuldip K. Paliwal

Published online: 2 November 2007

Abstract The purpose of conventional linear discriminant analysis (LDA) is to find an orientation which projects high dimensional feature vectors of different classes to a more manageable low dimensional space in the most discriminative way for classification. The LDA technique utilizes an eigenvalue decomposition (EVD) method to find such an orientation. This computation is usually adversely affected by the small sample size problem. In this paper we have presented a new direct LDA method (called gradient LDA) for computing the orientation especially for small sample size problem. The gradient descent based method is used for this purpose. It also avoids discarding the null space of within-class scatter matrix and between-class scatter matrix which may have discriminative information useful for classification.

Keywords Gradient linear discriminant analysis · Small sample size problem · Fisher's criterion function · Dimensionality reduction

1 Introduction

Linear discriminant analysis (LDA) is a well known technique for dimensionality reduction. It finds an orientation $\mathbf{W}$ that reduces a high dimensional feature vectors belonging to different classes to a lower dimensional feature space such that the projected feature vectors of a class on this lower dimensional space are well separated from the feature vectors of other classes. If the dimensionality reduction is from d-dimensional ($\mathbf{R}^d$) space to h-dimensional ($\mathbf{R}^h$) space (where $h < d$) then the size of the orientation matrix $\mathbf{W}$ would be $d \times h$. Therefore $\mathbf{W}$ has h column vectors known as the basis vectors. The orientation $\mathbf{W}$ is evaluated so that the Fisher's criterion function $J(\mathbf{W})$ is maximum. The criterion function depends on three factors: orientation $\mathbf{W}$, within-class scatter matrix (S_W) and between-class scatter matrix (S_B). For a c-class problem the value of h will be $c - 1$ or less, a constraint due to S_B. In the

A. Sharma (✉) · K. K. Paliwal
Signal Processing Lab, Griffith University, Brisbane, Australia
e-mail: sharma_al@usp.ac.fj

basic or conventional LDA technique, the orientation $\mathbf{W}$ is computed by using eigenvalue decomposition (EVD) method where scatter matrix S_W is arranged in such a way that it restricts the computation of $\mathbf{W}$ if it is being singular or reduced rank matrix. This limitation (quite often arises in human face recognition problem) is due to the high dimensionality of original feature vectors in comparison with the low number of feature vectors available. This drawback of LDA is known as small sample size problem [1]. To overcome this problem, several authors [2–6] have used intermediate techniques like principal component analysis (PCA) prior to the application of LDA. The PCA technique is used in such a way that the projected feature vectors on h-dimensional space give a full rank S_W matrix. Thereby the computation of the inverse of S_W is feasible and thus orientation $\mathbf{W}$ can then be found by the basic LDA method. The application of intermediate techniques would, however, sacrifice some classification performance. There are some techniques recently developed to solve small sample size problems. Chen et al. [7] have proposed a new LDA-based method. Their new LDA is based on the modified Fisher's criterion and involves discarding the null space of S_W, which contains the most discriminative information useful for classification [7,8]. Yu and Yang [8] presented a direct LDA method which discards the null space of S_B, however, prevents discarding the null space of S_W. Lu et al. [9] presented an approach based on the combination of direct LDA and fractional-step LDA [10] methods that overcomes shortcomings and limitations of individual methods used in the combination.

In this paper we do not extend any techniques presented in [7–10]. However, we have presented a new way of computing the orientation $\mathbf{W}$ which is derived directly from the conventional LDA technique. We used gradient descent method to solve for the orientation $\mathbf{W}$. The learning rate parameter is taken to be unity and $J(\mathbf{W})$ is used adaptively in the iterative process. This makes the convergence fast and reliable which is empirically presented. For brevity we call the proposed technique as gradient LDA technique. The gradient LDA technique can compute orientation $\mathbf{W}$ for both singular and non-singular S_W. This technique does not discard any null spaces of S_W and S_B thereby preserving discriminative information that may be useful for classification.

2 LDA Revisited

To explicitly define S_B and S_W for the Fisher's criterion function, in a c-class (assuming $c > 2$) problem let χ denotes d-dimensional set of n feature vectors, $\Omega = \{\omega_i : i = 1, 2, \ldots, c\}$ be the finite set of c states of nature or class labels where ω_i denotes the ith class label. The set χ can be subdivided into c subsets $\chi_1, \chi_2 \ldots, \chi_c$ where each subset χ_i belongs to ω_i and consists of n_i number of samples such that:

$$n = \sum_{i=1}^{c} n_i$$

The samples or patterns of set χ can be written as:

$$\chi = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \quad \text{where } \mathbf{x}_j \in \mathbf{R}^d$$

$$\chi_i \subset \chi \quad \text{and} \quad \chi_1 \cup \chi_2 \cup \ldots \cup \chi_c = \chi.$$

Let $\boldsymbol{\mu}_j$ be the centroid of χ_j and $\boldsymbol{\mu}$ be the centroid of χ, then the between class scatter matrix is given as

$$S_B = \sum_{j=1}^{c} n_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathrm{T}} \tag{1}$$

It can be observed from Eq. 1 that S_B is the sum of c matrices of rank one or less, and because only $c - 1$ of these are independent, S_B is of rank $c - 1$ or less [11].

The within-class scatter matrix which is the sum of c scatter matrices is defined as

$$S_W = \sum_{i=1}^{c} S_i \tag{2}$$

where

$$S_i = \sum_{\mathbf{x}\in\chi_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \tag{3}$$

It can be observed from Eqs. 2 and 3 that S_W is the sum of c scatter matrices and each of the scatter matrices is the sum of n_i matrices, and because only $c(n_{avg} - 1)$ or less are independent (where $n_{avg} = \frac{1}{c}\sum_{j=1}^{c} n_j = n/c$), the rank of S_W (for $n \geq c$) is

$$rank(S_W) \leq c(n_{avg} - 1) = n - c$$

If $n - c \geq d$ then S_W is full rank matrix i.e. non-singular and its inversion is possible. Now given scatter matrices S_B and S_W we can define Fisher's criterion as a function of $\mathbf{W}$ as [11]

$$J(\mathbf{W}) = \frac{|\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}|}{|\mathbf{W}^{\mathrm{T}} S_W \mathbf{W}|} \tag{4}$$

where $|\bullet|$ is the determinant. The orientation $\mathbf{W}$ is taken so that the Fisher's criterion function $J(\mathbf{W})$ is maximum. In a c-class problem the LDA projects from d-dimensional space to $c-1$ or less dimensional space i.e. $\mathbf{W} : \mathbf{x} \to \mathbf{y}$ or $\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$ where $\mathbf{x} \in \mathbf{R}^d$, $\mathbf{y} \in \mathbf{R}^h$ such that $1 \leq h \leq c - 1$. The orientation $\mathbf{W}$ is a rectangular matrix of size $d \times h$ which is the solution of the conventional eigenvalue problem

$$S_W^{-1} S_B \mathbf{w}_i = \lambda_i \mathbf{w}_i \tag{5}$$

where $\mathbf{w}_i$ are the column vectors of $\mathbf{W}$ that correspond to the largest eigenvalues (λ_i) in Eq. 5. It is evident from Eq. 5 that the explicit solution of the orientation can be found when S_W is non-singular. If $S_W$is singular (i.e. $n - c < d$) then it is not possible to obtain the orientation $\mathbf{W}$ by using Eq. 5. To overcome this singularity problem, we have presented the gradient LDA method which is described in the next section.

3 Gradient LDA for Reduced Rank Within-class Scatter Matrix

It is possible to find the desired leading h eigenvectors of the orientation $\mathbf{W}$ for reduced rank S_W matrix provided $rank(S_W) \geq h$ and $rank(S_B) \geq h$. A direct computation of $\mathbf{W}$ can be achieved by applying gradient descent method on the Fisher's criterion function. Here we are interested in the orientation $\mathbf{W}$ that gives maximum $J(\mathbf{W})$ value. However, denoting $\hat{J}(\mathbf{W}) = 1/J(\mathbf{W})$, the maximization problem becomes the minimization problem, where we investigate the orientation $\mathbf{W}$ that minimizes $\hat{J}(\mathbf{W})$ value. To derive the gradient LDA method we first find the derivative of $\hat{J}(\mathbf{W})$ then update $\mathbf{W}$ using gradient descent method

Table 1 Gradient LDA algorithm for computing the orientation **W**

```
1. Choose $h$, the number of leading eigenvectors required to estimate.
2. Initialize the orientation $\mathbf{W}$ of size $d \times h$ e.g. randomly or using identity matrix[a]
3. while (true)
4. Compute $\hat{J}(\mathbf{W}) = |\mathbf{W}^{\mathrm{T}} S_W \mathbf{W}|/|\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}|$
5. $\mathbf{W} \leftarrow \mathbf{W} - \alpha 2\hat{J}(\mathbf{W})[S_W \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_W \mathbf{W})^{-1} - S_B \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_B \mathbf{W})^{-1}]$
6. Normalize column vectors of $\mathbf{W}$
     for $j = 1$ to $h$
       $\mathbf{W}(:, j) \leftarrow \mathbf{W}(:, j)/||\mathbf{W}(:, j)||$[b]
     end
7. end
```

[a] In a $d \times h$ identity matrix $I_{d\times h}$, the first h rows and columns is an identity matrix $\mathrm{I}_{h\times h}$ and the last $d - h$ rows are zero elements i.e. $\mathrm{I}_{d\times h} = [\mathrm{I}_{h\times h} 0_{h\times d-h}]^{\mathrm{T}}$

[b] In $\mathbf{W}(:, j)$, ':, j' indicates elements of all the rows of jth column (i.e. jth column vector) and $||\bullet||$ denotes the norm value of this column vector

while normalizing the column vectors of **W** for each of the iterations. The derivative of $\hat{J}(\mathbf{W})$ can be given from Appendix 1 as

$$\frac{\partial \hat{J}(\mathbf{W})}{\partial \mathbf{W}} = 2\hat{J}(\mathbf{W})[S_W \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_W \mathbf{W})^{-1} - S_B \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_B \mathbf{W})^{-1}] \tag{6}$$

It can be observed from Eq. 6 that the inverse of S_W (a $d \times d$ sized matrix) is not computed in the equation as has been done in Eq. 5. However, inverse of $(\mathbf{W}^{\mathrm{T}} S_W \mathbf{W})$ and $(\mathbf{W}^{\mathrm{T}} S_B \mathbf{W})$ are computed to find the derivative of $\hat{J}(\mathbf{W})$ which are full rank $h \times h$ sized matrices. Eq. 6 can be utilized in the gradient descent algorithm to solve for the values of **W**

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha \frac{\partial \hat{J}(\mathbf{W})}{\partial \mathbf{W}} \tag{7}$$

$$\mathbf{W} \leftarrow \textit{Normalize each of the column vectors of } \mathbf{W} \textit{ separately} \tag{8}$$

where α is a learning rate parameter. It can also be observed from substituting Eq. 6 in Eq. 7 that $\hat{J}(\mathbf{W})$ is updated for each of the iterations. The gradient LDA algorithm is illustrated in Table 1. It will be empirically seen in the next section that the algorithm converges fast when unity value of α is taken and when $\hat{J}(\mathbf{W})(\mathbf{or}\ J(\mathbf{W}))$ is utilized adaptively in the algorithm. This makes the algorithm fast converging for the iteration process. The iterative process of the algorithm can be terminated when $J(\mathbf{W})$ becomes stable.

The convergence relation proof of the gradient LDA technique can be easily shown. Since it is a typical gradient descent based algorithm, the convergence proof will be similar to that of the LMS (least-mean-squared) algorithm.

4 An Illustration

In this section we first compare the performance of the proposed gradient LDA technique with that of the basic LDA technique using the Fisher's criterion value as a prototype. Since the basic LDA can be applied only for full ranked S_W matrix we have taken the dataset accordingly. For this purpose Sat-Image dataset from UCI repository [12] is used. The Sat-Image dataset consists of six distinct classes with 36 dimensions or attributes. It has 4,435 feature vectors for training purpose and 2,000 feature vectors for testing purpose. However,

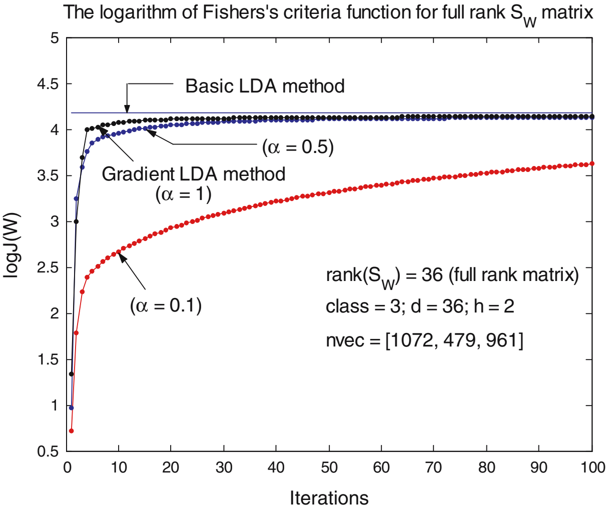

Fig. 1 A comparison between basic LDA method and Gradient LDA method using Fisher's criterion value as a prototype

for this comparison we have taken the features from the first three classes of the training set. The coordinates of dataset taken are given as follows:

	Class 1	Class 2	Class3
Number of training feature vectors per class (nvec):	1072	479	961

The dimension is reduced from 36-dimensional space to 2-dimensional plane. The total number of feature vectors minus the number of classes $(n - c)$ is 2,509 which is greater than the original dimension $(d = 36)$. Therefore S_W is a full rank matrix of size 36 × 36. Figure 1 illustrates the comparison between both the techniques for this dataset. The x-axis represents the number of iterations used for gradient LDA method and y-axis represents Fisher's criterion in logarithmic scale $(\log J(\mathbf{W}))$ for both the techniques. Five different values (0.1, 0.5, 1, 2 and 5) of α are taken for the gradient LDA algorithm. The α values[1] 2 and 5 do not converge and provide negative $J(\mathbf{W})$ values which cannot be plotted on the figure (since $\log J(\mathbf{W})$ will yield a complex value). It can be observed from the figure that gradient LDA algorithm converges fast for $\alpha = 1$. Substituting this unity value for α (in Eq. 7) means that the convergence becomes independent of any learning rate parameter or initial settings. One of the reasons for this fast convergence is the use of parameter $\hat{J}(\mathbf{W})$(or $J(\mathbf{W})$) adaptively in the algorithm (Table 1) i.e.$J(\mathbf{W})$ is updated for each of the iterations or for every single change in the value of $\mathbf{W}$. This adaptation makes the process fast and reliable. The value of $J(\mathbf{W})$ for gradient LDA is very close to the value of $J(\mathbf{W})$ of the basic LDA method. This test indicates that the orientation $\mathbf{W}$ obtained by both the techniques will discriminate different classes of feature vectors in a similar fashion.

Next, we have taken feature vectors such that S_W is no longer full rank matrix to demonstrate its use in solving the small sample size problem. The same Sat-Image dataset is used where only four vectors from each of the three classes are taken i.e.

[1] The α values above 1 usually do not provide very stable $J(\mathbf{W})$ values i.e. convergence is not guaranteed.

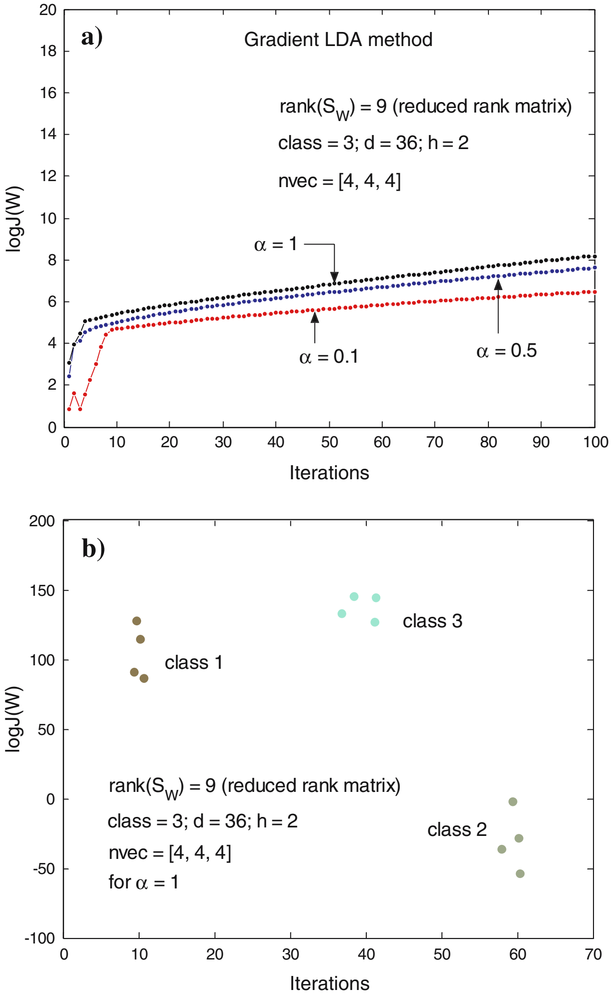

Fig. 2 Gradient LDA application for small sample size problem. (**a**) The logarithm of Fisher's criterion function for reduced rank S_W matrix. (**b**) Projection of feature vectors onto 2-dimensional plane using Gradient LDA method for $\alpha = 1$

	Class 1	Class 2	Class3
Number of training feature vectors per class (nvec):	4	4	4

The rest of the parameters are not altered (i.e. $d = 36$ and $h = 2$). The size of S_W matrix is still 36 × 36. However, its rank is now $n - c = 9$. The basic LDA method cannot be applied here since S_W is singular thereby its inverse is not possible. The gradient LDA method is applied in this case for the same five values of α. The Fisher's criterion in logarithmic scale is depicted in Fig. 2a and projected samples ($\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$) on 2-dimensional plane is depicted in Fig. 2b (for $\alpha = 1$).

Here also the α values greater than unity (2 and 5) diverge and give complex $\log J(\mathbf{W})$ values which cannot be plotted in Fig 2a. The convergence using other values of α is depicted in the figure. It can be observed from the figure that the algorithm achieves stable Fisher's criterion value somewhere before the tenth iteration. The orientation **W** at this iteration is adequate for providing the discrimination between different classes of feature vectors in the reduced dimensional space. In this case as well, the unity value of α is giving better results than the other presented values. This means that $\alpha = 1$ is a suitable choice for the convergence of the algorithm.

In Fig 2a the iteration count is shown up to 100, however, we have experimented the algorithm up to serval thousand iteration counts and it was observed that the log $J(\mathbf{W})$(or $J(\mathbf{W})$) value increases towards positive infinity. This means increasing the iteration count the $\mathbf{W}^{\mathrm{T}} S_W \mathbf{W}$ value tends towards zero whereas $\mathbf{W}^{\mathrm{T}} S_b \mathbf{W}$ remains non-zero tending $J(\mathbf{W})$ towards positive infinity which will give the value of **W** close to the global optimal solution of **W**.

Figure 2b illustrates projection of 36-dimensional feature vectors onto 2-dimensional plane using orientation **W** which is obtained by the gradient LDA method (for $\alpha = 1$). It is evident from the figure that different classes of feature vectors are well separated.

It can be concluded from the experiments that gradient LDA method is an efficient substitute of basic LDA method especially for reduced rank within-class scatter matrix (small sample size problem).

5 Conclusion

We have presented a new way of computing the orientation **W** in LDA which addresses small sample size problem. The proposed method (called gradient LDA) is based on gradient descent method but the convergence fast and reliable. The gradient LDA method does not discard any null spaces of S_W and S_B matrices and thus preserves discriminative information which is useful for classification.

Appendix 1

Lemma 1 *Let the scalar function $\hat{J}(\mathbf{W}) = |\mathbf{W}^{\mathrm{T}} S_W \mathbf{W}|/|\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}|$ be a differentiable function of a $d \times h$ rectangular matrix* **W** *such that $h < d$. The size of both the symmetric matrices S_B and S_W is $d \times d$ and the rank for both is greater or equal to h. Then the derivative of $\hat{J}(\mathbf{W})$ is defined as*

$$\partial \hat{J}(\mathbf{W})/\partial \mathbf{W} = 2\hat{J}(\mathbf{W})[S_W \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_W \mathbf{W})^{-1} - S_B \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_B \mathbf{W})^{-1}].$$

Proof 1 Using the quotient rule of differentiation we can differentiate $\hat{J}(\mathbf{W})$ with respect to **W** as

$$\frac{\partial \hat{J}(\mathbf{W})}{\partial \mathbf{W}} = [|\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}| \frac{\partial}{\partial \mathbf{W}}(|\mathbf{W}^{\mathrm{T}} S_W \mathbf{W}|) - \frac{\partial}{\partial \mathbf{W}}(|\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}|)|\mathbf{W}^{\mathrm{T}} S_W \mathbf{W}|]/|\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}|^2 \tag{A1}$$

from Appendix 2 we can write Eq. A1 as

$$= 2|\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}||\mathbf{W}^{\mathrm{T}} S_W \mathbf{W}|[S_W \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_W \mathbf{W})^{-1} - S_B \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_B \mathbf{W})^{-1}]/|\mathbf{W}^{\mathrm{T}} S_B \mathbf{W}|^2$$

Therefore

$$\partial \hat{J}(\mathbf{W})/\partial \mathbf{W} = 2\hat{J}(\mathbf{W})[S_W \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_W \mathbf{W})^{-1} - S_B \mathbf{W}(\mathbf{W}^{\mathrm{T}} S_B \mathbf{W})^{-1}] \qquad \square$$

Appendix 2

Lemma 2 *Let the scalar function* $g(\mathbf{W}) = |\mathbf{W}^T\mathbf{SW}|$ *be a differentiable function of a* $d \times h$ *rectangular matrix* $\mathbf{W}$ *such that* $h < d$*.The size of symmetric matrix* $\mathbf{S}$ *is* $d \times d$ *and* $rank(\mathbf{S}) \geq h$*. Then derivative of* $g(\mathbf{W})$ *with respect to* $\mathbf{W}$ *is defined as* $\partial g(\mathbf{W})/\partial\mathbf{W} = 2|\mathbf{W}^{\mathrm{T}}\mathbf{SW}|\mathbf{SW}(\mathbf{W}^{\mathrm{T}}\mathbf{SW})^{-1}$.

Proof 2 The derivative of any determinant $\mathbf{X}$ is given by [13]

$$\partial|\mathbf{X}|/\partial\mathbf{X} = |\mathbf{X}|(\mathbf{X}^{\mathrm{T}})^{-1} \tag{A2}$$

Equation A2 can also be written in the *trace* format as

$$\partial|\mathbf{X}| = |\mathbf{X}|trace[(\mathbf{X}^{\mathrm{T}})^{-1}\partial\mathbf{X}^{\mathrm{T}}] \tag{A3}$$

from Eq. A3 the derivative of $g(\mathbf{W})$ is

$$\begin{aligned}
\partial g(\mathbf{W}) &= |\mathbf{W}^{\mathrm{T}}\mathbf{SW}|trace[(\mathbf{W}^{\mathrm{T}}\mathbf{SW})^{\mathrm{T}^{-1}}\partial(\mathbf{W}^{\mathrm{T}}\mathbf{SW})^{\mathrm{T}}] \\
&= |\mathbf{W}^{\mathrm{T}}\mathbf{SW}|\{trace[\mathbf{S}^{\mathrm{T}}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{SW})^{\mathrm{T}^{-1}}\partial\mathbf{W}^{\mathrm{T}}] \\
&\quad +trace[\mathbf{SW}(\mathbf{W}^{\mathrm{T}}\mathbf{SW})^{-1}\partial\mathbf{W}^{\mathrm{T}}]\} \\
&\quad \{\because trace(\mathrm{A}^{\mathrm{T}}) = trace(\mathrm{A}) \text{ and } trace(\mathrm{AB}) = trace(\mathrm{BA})\} \\
&= 2|\mathbf{W}^{\mathrm{T}}\mathbf{SW}|trace\left[\mathbf{SW}(\mathbf{W}^{\mathrm{T}}\mathbf{SW})^{-1}\partial\mathbf{W}^{\mathrm{T}}\right] \\
&\quad \{\because \mathbf{S} \text{ is a symmetric matrix therefore} (\mathbf{W}^{\mathrm{T}}\mathbf{SW}) \text{ is symmetric too}\} \\
&\quad \therefore \partial g(\mathbf{W})/\partial\mathbf{W} = 2|\mathbf{W}^{\mathrm{T}}\mathbf{SW}|\mathbf{SW}(\mathbf{W}^{\mathrm{T}}\mathbf{SW})^{-1} \qquad \square
\end{aligned}$$

References

1. Fukunaga K (1990) Introduction to statistical pattern recognition. Academic Press Inc., Hartcourt Brace Jovanovich, Publishers
2. Swets DL, Weng J (1996) Using discriminative eigenfeatures for image retrieval. IEEE Trans. Pattern Anal Mach Intell 18(8):831–836
3. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
4. Zhao W, Chellappa R, Nandhakumar N (1998) Empirical performance analysis of linear discriminant classifiers. Proc. IEEE Conf Comput Vision Pattern Recogn, 164–169
5. Zhao W, Chellappa R, Phillips PJ (1999) Subspace linear discriminant analysis for face recognition, Tech. Rep. CAR-TR-914, Center for Automation Research, University of Maryland, College Park
6. Sharma A, Paliwal KK, Onwubolu GC (2006) Class-dependent PCA, MDC and LDA: a combined classifier for pattern classification. Pattern Recogn 39(7):1215–1229
7. Chen L-F, Liao H-YM, Ko M-T, Lin J-C, Yu G-J (2000) A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recogn 33:1713–1726
8. Yu H, Yang J (2001) A direct LDA algorithm for high-dimensional data-with application to face recognition. Pattern Recogn 34:2067–2070
9. Lu J, Plataniotis KN, Venetsanopoulos AN (2003) Face recognition using LDA-based algorithms. IEEE Trans Neural Netw 14(1):195–200
10. Lotlikar R, Kothari R (2000) Fractional-step dimensionality reduction. IEEE Trans Pattern Anal Mach. Intell 22(6):623–627
11. Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York
12. Blake CL, Merz CJ (1998) UCI repository of machine learning databases, http://www.ics.uci.edu/~mlearn. University of Calif., Dept. of Information and Comp. Sci., Irvine
13. Magnus JR, Neudecker H (1994) Matrix differential calculus with applications in statistics and econometrics. Wiley