Int. J. Appl. Math. Comput. Sci., 2011, Vol. 21, No. 3, 549–558
DOI: 10.2478/v10006-011-0043-9

amcs

ANALYSIS OF CORRELATION BASED DIMENSION REDUCTION METHODS

YONG JOON SHIN, CHEONG HEE PARK

Department of Computer Science and Engineering
Chungnam National University, 220 Gung-dong, Yuseong-gu, Daejeon, 305-764, Korea
e-mail: {yjsin,cheonghee}@cnu.ac.kr

Dimension reduction is an important topic in data mining and machine learning. Especially dimension reduction combined with feature fusion is an effective preprocessing step when the data are described by multiple feature sets. Canonical Correlation Analysis (CCA) and Discriminative Canonical Correlation Analysis (DCCA) are feature fusion methods based on correlation. However, they are different in that DCCA is a supervised method utilizing class label information, while CCA is an unsupervised method. It has been shown that the classification performance of DCCA is superior to that of CCA due to the discriminative power using class label information. On the other hand, Linear Discriminant Analysis (LDA) is a supervised dimension reduction method and it is known as a special case of CCA. In this paper, we analyze the relationship between DCCA and LDA, showing that the projective directions by DCCA are equal to the ones obtained from LDA with respect to an orthogonal transformation. Using the relation with LDA, we propose a new method that can enhance the performance of DCCA. The experimental results show that the proposed method exhibits better classification performance than the original DCCA.

Keywords: canonical correlation analysis, dimension reduction, discriminative canonical correlation analysis, feature fusion; linear discriminant analysis.

1. Introduction

Dimension reduction is a widely used preprocessing step in pattern recognition and data mining. It can help to avoid the curse of dimensionality and give a compact representation of original data with a limited loss of information. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are traditional dimension reduction methods which use data scatter information (Jolliffe, 1986; Duda *et al*., 2001). PCA finds a projective direction giving the greatest total scatter of data, and LDA searches for a projection vector that maximizes class separability in the reduced dimensional space by maximizing between-class scatter and minimizing within-class scatter.

However, for undersampled problems where the number of data samples is smaller than the data dimension, scatter matrices used in LDA become singular and their inverses are not defined. In order to overcome the problems caused by the singularity of scatter matrices, several methods have been proposed, where the singularity problem is avoided by performing the maximization of between-class scatter and the minimization of within-class scatter one after the other (Chen *et al*., 2000; Yu and Yang, 2001; Yang and Yang, 2003; Howland and Park, 2004). Comparative studies of generalized LDA methods can be found in the works of Park and Park (2008) or Ye (2005). While PCA and LDA are linear dimension reduction methods, nonlinear dimension reduction methods using kernel based approaches or local learning algorithms have been developed for data with nonlinear structures (Billings and Lee, 2002; Baudat and Anouar, 2000; Sugiyama, 2006; Nie *et al*., 2007). In kernel methods, an original data space is transformed to a feature space by an implicit nonlinear mapping through kernel functions so that any linear dimension reduction methods formulated with inner product computations can be performed in the transformed feature space.

Recently, graph based dimension reduction methods have been actively investigated (He and Niyogi, 2003; Hou *et al*., 2009; Yan *et al*., 2007; Pardalos and Hansen, 2008). Locality Preserving Projection (LPP) (He and Niyogi, 2003) pursues the minimization of the local scatter by minimizing the distances between near points. LPP is different from other manifold-based dimension reduction methods such as Isomap (Tenenbaum *et al.*, 2000) or LLE (Roweis and Saul, 2000) in the sense

that it is a linear method and the transformation is explicitly composed. Various modifications of the objective function in LPP have been proposed, for example, by adding the maximization of the global scatter of distant points (Yan *et al*., 2007), or orthogonal and smooth regularization (Hou *et al*., 2009).

When the data are described by multiple feature sets, dimension reduction combined with feature fusion becomes necessary (Yang *et al*., 2003; Sun *et al*., 2005; Garthwaite, 1994). The union of multiple feature sets obtained from various sources can make the data dimension large. Canonical Correlation Analysis (CCA) is a feature fusion method based on correlation (Hotelling, 1936). It finds directions which maximize the correlation between feature vectors of two feature sets (Sun *et al*., 2005). While CCA is an unsupervised feature fusion method, Discriminative Canonical Correlation Analysis (DCCA) is a supervised feature fusion method (Sun *et al*., 2008). It utilizes class information by maximizing the correlation between feature vectors in the same class and minimizing the correlation between feature vectors belonging to different classes.

It is known that Linear Discriminant Analysis (LDA) can be considered a special case of CCA. When the second feature set corresponding to the original feature set is constructed by class label information, performing CCA gives the same results as LDA. Based on the relation between CCA and LDA, we analyze the relationship between DCCA and LDA, showing that the projective directions by DCCA are equal to the ones obtained from LDA with respect to an orthogonal transformation. On the other hand, LDA is optimal for data which have normal class distributions but may not work well for data with complex class structures. Similarly, DCCA can fail when the data have nonnormal class distributions. We propose a discriminative feature fusion method that can be effective for the data with general class distributions. Our proposed method can reflect complex class shapes by using local neighborhood scatters within classes instead of the global within-class scatter.

This paper is organized as follows. In Section 2, a brief review of LDA, CCA and DCCA is given. In Section 3, a theoretical analysis of DCCA is given by revealing the relation of DCCA and LDA. In Section 4, we propose an improved method of DCCA which can be applied effectively for data with nonnormal class distributions. In Section 5, experimental results show that the proposed method exhibits better classification performance than DCCA. Discussions are given in Section 6.

2. Reviews of LDA, CCA and DCCA

2.1. LDA. The goal of LDA is to find a linear transformation that maximizes class separability in the reduced dimensional space (Duda *et al*., 2001). Hence the criterion for dimension reduction in LDA is to maximize between-class scatter and minimize within-class scatter. The scatters are measured by using scatter matrices. Denoting by x_j^i the j-th example in class i, the data set X is given as

$$X = [x_1^1, \ldots, x_{n_1}^1, \ldots, x_1^r, \ldots, x_{n_r}^r] \in \mathbb{R}^{p\times n}. \quad (1)$$

The between-class scatter matrix S_b, within-class scatter matrix S_w and total scatter matrix S_t are defined as

$$S_b = \sum_{i=1}^{r} n_i(\overline{x^i} - \overline{x})(\overline{x^i} - \overline{x})^T,$$

$$S_w = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_j^i - \overline{x^i})(x_j^i - \overline{x^i})^T,$$

$$S_t = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_j^i - \overline{x})(x_j^i - \overline{x})^T,$$

where $n = \sum_{i=1}^{r} n_i$, and

$$\overline{x^i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j^i,$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_j^i$$

are the class means and the global mean, respectively.

The objective function of LDA can be formulated as a maximization problem,

$$\arg\max_{G\in\mathbb{R}^{p\times l}} \frac{|G^T S_b G|}{|G^T S_w G|}, \quad (2)$$

where the data dimension is reduced from p to l by a linear transformation such that $G^T : x \longmapsto G^T x$. It is well known that the eigenvectors corresponding to the $r-1$ largest eigenvalues of

$$S_b g = \lambda S_w g \quad (3)$$

form the columns of a linear transformation matrix G for LDA (Fukunaga, 1990; Duda *et al*., 2001).

2.2. CCA. CCA is an unsupervised feature fusion method for two feature sets describing the same data objects (Sun *et al*., 2005). CCA finds projective directions which maximize the correlation between the feature vectors of two feature sets.

Given a data set with n pairs of feature vectors

$$\{(x_i, y_i), i = 1, \ldots, n\},$$

the centered data are denoted as

$$X = [x_1 - \overline{x}, \ldots, x_n - \overline{x}] \in \mathbb{R}^{p\times n},$$
$$Y = [y_1 - \overline{y}, \ldots, y_n - \overline{y}] \in \mathbb{R}^{q\times n},$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the means of x_is and y_is, respectively. The objective function of CCA(X,Y) is expressed as

$$\arg\max_{g_x,g_y} \frac{g_x^T X Y^T g_y}{\sqrt{g_x^T X X^T g_x}\sqrt{g_y^T Y Y^T g_y}}, \tag{4}$$

which can be restated as

$$\arg\max_{g_x,g_y} g_x^T X Y^T g_y \tag{5a}$$

$$g_x^T X X^T g_x = 1, \quad g_y^T Y Y^T g_y = 1. \tag{5b}$$

After finding a pair of projective directions, (g_{x1}, g_{y1}), satisfying (5), the second pair of projective directions can be found by solving the optimization problem

$$\arg\max_{g_x,g_y} g_x^T X Y^T g_y \tag{6a}$$

$$g_x^T X X^T g_x = g_y^T Y Y^T g_y = 1, \tag{6b}$$

$$g_{x1}^T X X^T g_x = g_{y1}^T Y Y^T g_y = 0. \tag{6c}$$

Repeating the above process is converted to the solving of the paired eigenvalue problem

$$\begin{aligned} X Y^T \left(Y Y^T\right)^{-1} Y X^T g_x &= \lambda X X^T g_x, \\ Y X^T \left(X X^T\right)^{-1} X Y^T g_y &= \lambda Y Y^T g_y, \end{aligned} \tag{7}$$

and the eigenvectors (g_{xi}, g_{yi}), $i = 1, \ldots, l$, corresponding to the l largest eigenvalues are the pairs of projective directions for CCA (Sun *et al.*, 2005). Hence

$$\{(g_{xi})^T X, i = 1, \ldots, l\}$$

and

$$\{(g_{yi})^T Y, i = 1, \ldots, l\}$$

compose the feature sets extracted from X and Y by CCA. The number l is determined as the number of nonzero eigenvalues.

2.3. DCCA. While CCA does not use class label information, DCCA is a supervised feature fusion method which aims at improving classification performance by utilizing class information (Sun *et al.*, 2008). DCCA uses two types of correlations between two feature sets, maximizing the correlation between feature vectors in the same classes and minimizing the correlation between feature vectors in different classes.

We are given two centered feature sets $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{q\times n}$ such as

$$\begin{aligned} X &= \left[x_1^1 - \overline{x}, \ldots, x_{n_1}^1 - \overline{x}, \ldots, x_1^r - \overline{x}, \ldots, x_{n_r}^r - \overline{x}\right], \\ Y &= \left[y_1^1 - \overline{y}, \ldots, y_{n_1}^1 - \overline{y}, \ldots, y_1^r - \overline{y}, \ldots, y_{n_r}^r - \overline{y}\right], \end{aligned} \tag{8}$$

where x_j^i denotes the j-th example in class i. When C_w and C_b denote the correlations between the same classes and between different classes, respectively, such that

$$C_w = \sum_{i=1}^{r}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i} \left(x_k^i - \overline{x}\right)\left(y_l^i - \overline{y}\right)^T$$

and

$$C_b = \sum_{i=1}^{r}\sum_{j=1,j\neq i}^{r}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} \left(x_k^i - \overline{x}\right)\left(y_l^j - \overline{y}\right)^T,$$

an objective function of DCCA (X, Y) is defined as

$$\arg\max_{g_x,g_y} g_x^T C_w g_y - g_x^T C_b g_y \tag{9a}$$

$$g_x^T X X^T g_x = 1, \quad g_y^T Y Y^T g_y = 1. \tag{9b}$$

By simple computation, we can get the relation

$$C_b = -C_w,$$

and the problem (9) is simplified to

$$\arg\max_{g_x,g_y} g_x^T C_w g_y, \tag{10a}$$

subject to

$$g_x^T X X^T g_x = 1, \quad g_y^T Y Y^T g_y = 1. \tag{10b}$$

Similarly as in CCA, the problem (10) can be solved by the eigenvalue problems

$$\begin{aligned} C_w (Y Y^T)^{-1} C_w^T g_x &= \lambda^2 X X^T g_x, \\ C_w^T (X X^T)^{-1} C_w g_y &= \lambda^2 Y Y^T g_y. \end{aligned} \tag{11}$$

For any $d \leq \min(p, q, r)$, the eigenvectors (g_{xi}, g_{yi}), $i = 1, \ldots, d$, corresponding to the d largest eigenvalues are the ones that satisfy the objective function (10), and therefore

$$\{(g_{xi})^T X, i = 1, \ldots, d\}$$

and

$$\{(g_{yi})^T Y, i = 1, \ldots, d\}$$

form the feature sets extracted from X and Y by DCCA. For high dimensional data where the number of classes, r, is smaller than the number of features, the rank of C_w is generally $r - 1$ and we can get $r - 1$ nonzero eigenvalues, indicating $d = r - 1$.

2.4. Solving coupled eigenvalue problems. The coupled eigenvalue problems in (11) can be solved by Singular Value Decomposition (SVD) (Sun *et al.*, 2008). Let

$$\begin{aligned} H &= (X X^T)^{-\frac{1}{2}} C_w (Y Y^T)^{-\frac{1}{2}}, \\ u &= (X X^T)^{\frac{1}{2}} g_x, \quad v = (Y Y^T)^{\frac{1}{2}} g_y, \\ \operatorname{rank}(H) &= s. \end{aligned} \tag{12}$$

Then the equations in (11) can be written as

$$HH^T u = \lambda^2 u,$$
$$H^T H v = \lambda^2 v.$$

Let the SVD of H be

$$H = UDV^T = \sum_{i=1}^{s} \lambda_i u_i v_i^T,$$

where u_i and v_i are the column vectors of orthogonal matrices U and V, respectively. This means that u_is are the eigenvectors of HH^T and v_is are the eigenvectors of $H^T H$. Hence $g_{xi} = (XX^T)^{-\frac{1}{2}} u_i$ and $g_{yi} = (YY^T)^{-\frac{1}{2}} v_i$ yield the solution for the problem (11). The problem (7) in Section 2.2 can be solved in the same way.

3. Analysis of the relation between DCCA and LDA

For labeled data, when the second feature set corresponding to the original feature set is composed of label information, performing CCA for those two feature sets is equal to applying LDA to the original feature set (Sun and Chen, 2007). Based on the relation between CCA and LDA, we derive a relation between DCCA and LDA.

3.1. DCCA(X, Y) equals CCA(X, C) + CCA(Y, C).

Suppose we are given data sets denoted by (8), and the matrix C represents class label information of the data points such that

$$C = [\underbrace{c^1, \ldots, c^1}_{n_1}, \ldots, \underbrace{c^r, \ldots, c^r}_{n_r}],$$

where $c^i \in \mathbb{R}^r$ is a column vector whose i-th element is 1 and the others are zero. By denoting the projective directions obtained from DCCA(X, Y), CCA(X, C) and CCA(Y, C) as (g_x^d, g_y^d), (g_x^{cx}, g_c^{cx}) and (g_y^{cy}, g_c^{cy}), respectively, the objective functions for DCCA(X, Y), CCA(X, C) and CCA(Y, C) are

$$\text{DCCA}(X,Y) \quad : \quad \arg\max_{g_x^d, g_y^d} \left(g_x^d\right)^T C_w g_y^d, \quad \text{s.t.} \quad \begin{matrix} \left(g_x^d\right)^T XX^T g_x^d = 1, \\ \left(g_y^d\right)^T YY^T g_y^d = 1, \end{matrix}$$

$$\text{CCA}(X,C) \quad : \quad \arg\max_{g_x^{cx}, g_c^{cx}} \left(g_x^{cx}\right)^T XC^T g_c^{cx}, \quad \text{s.t.} \quad \begin{matrix} \left(g_x^{cx}\right)^T XX^T g_x^{cx} = 1, \\ \left(g_c^{cx}\right)^T CC^T g_c^{cx} = 1, \end{matrix}$$

$$\text{CCA}(Y,C) \quad : \quad \arg\max_{g_y^{cy}, g_c^{cy}} \left(g_y^{cy}\right)^T YC^T g_c^{cy}, \quad \text{s.t.} \quad \begin{matrix} \left(g_y^{cy}\right)^T YY^T g_y^{cy} = 1, \\ \left(g_c^{cy}\right)^T CC^T g_c^{cy} = 1. \end{matrix}$$

When $\overline{x^i} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j^i$ and $\overline{y^i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_j^i$ are the class means, we let

$$\hat{X} = \left[n_1 \left(\overline{x^1} - \overline{x} \right), \ldots, n_r \left(\overline{x^r} - \overline{x} \right) \right],$$
$$\hat{Y} = \left[n_1 \left(\overline{y^1} - \overline{y} \right), \ldots, n_r \left(\overline{y^r} - \overline{y} \right) \right].$$

Then C_w, XC^T, YC^T and CC^T can be written as

$$\begin{aligned} C_w &= \sum_{i=1}^{r} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} \left(x_k^i - \overline{x}\right) \left(y_l^i - \overline{y}\right)^T \\ &= \sum_{i=1}^{r} n_i \left(\overline{x^i} - \overline{x} \right) n_i \left(\overline{y^i} - \overline{y} \right)^T \\ &= \hat{X} \hat{Y}^T, \\ XC^T &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(x_j^i - \overline{x}\right) \left(c^i\right)^T \\ &= \sum_{i=1}^{r} n_i \left(\overline{x^i} - \overline{x} \right) \left(c^i\right)^T \\ &= \hat{X} I = \hat{X}, \\ YC^T &= \hat{Y}, \\ CC^T &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} c^i (c^i)^T = \sum_{i=1}^{r} n_i c^i (c^i)^T \\ &= \begin{bmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_r \end{bmatrix}. \end{aligned} \tag{13}$$

As in Section 2.4, we define H as

$$\begin{aligned} H_d &= \left(XX^T\right)^{-\frac{1}{2}} C_w \left(YY^T\right)^{-\frac{1}{2}}, \\ u^d &= \left(XX^T\right)^{\frac{1}{2}} g_x^d, \\ v^d &= \left(YY^T\right)^{\frac{1}{2}} g_y^d, \end{aligned} \tag{14a}$$

$$\begin{aligned} H_{cx} &= \left(XX^T\right)^{-\frac{1}{2}} XC^T \left(CC^T\right)^{-\frac{1}{2}}, \\ u^{cx} &= \left(XX^T\right)^{\frac{1}{2}} g_x^{cx}, \end{aligned} \tag{14b}$$

$$\begin{aligned} H_{cy} &= \left(YY^T\right)^{-\frac{1}{2}} YC^T \left(CC^T\right)^{-\frac{1}{2}}, \\ u^{cy} &= \left(YY^T\right)^{\frac{1}{2}} g_y^{cy}. \end{aligned} \tag{14c}$$

From (13) and (14), we get

$$\begin{aligned} &H_{cx} CC^T H_{cy}^T \\ &\quad = \left(XX^T\right)^{-\frac{1}{2}} XC^T \left(CC^T\right)^{-\frac{1}{2}} CC^T \\ &\quad \cdot \left(CC^T\right)^{-\frac{1}{2}} CY^T \left(YY^T\right)^{-\frac{1}{2}} \\ &\quad = \left(XX^T\right)^{-\frac{1}{2}} \hat{X} \hat{Y}^T \left(YY^T\right)^{-\frac{1}{2}} \\ &\quad = \left(XX^T\right)^{-\frac{1}{2}} C_w \left(YY^T\right)^{-\frac{1}{2}} \\ &\quad = H_d. \end{aligned}$$

Since the rank of $\hat{X}$ and $\hat{Y}$ is generally $r-1$, so is the rank of C_w. Therefore, H_d, H_{cx} and H_{cy} have rank $r-1$. Now let SVDs of H_{cx}, H_{cy} and H_d be

$$H_{cx} = U_{cx} D_{cx} V_{cx}^T,$$
$$H_{cy} = U_{cy} D_{cy} V_{cy}^T,$$
$$H_d = U_d D_d V_d^T = \sum_{i=1}^{r-1} u_i^d \lambda_i^d (v_i^d)^T,$$

respectively. Then H_d can be written as

$$\begin{aligned} H_d &= H_{cx} \left(CC^T\right) H_{cy}^T \\ &= U_{cx} \underbrace{D_{cx} V_{cx}^T CC^T V_{cy} D_{cy}}_{A} U_{cy}^T \\ &= \sum_{i,\,j=1}^{r-1} u_i^{cx} \alpha_{ij} u_j^{cyT}, \end{aligned}$$

where u_i^{cx} and u_i^{cy} denote the i-th column vectors of U_{cx} and U_{cy}, respectively. Here α_{ij} is the (i,j)-th element of the matrix $A = D_{cx} V_{cx}^T CC^T V_{cy} D_{cy}$. Since the column vectors of U_{cy} are orthonormal,

$$H_d u_k^{cy} = \sum_{i=1}^{r-1} u_i^{cx} \alpha_{ik}, \quad k = 1, \ldots, r-1.$$

Hence $\{u_i^{cx},\, i = 1, \ldots, r-1\}$ spans the range space of H_d and is a basis of the range space of H_d (Anton and Busby, 2003). Since both $\left\{u_1^d,\, u_2^d,\, \ldots,\, u_{r-1}^d\right\}$ and $\left\{u_1^{cx},\, u_2^{cx},\, \ldots,\, u_{r-1}^{cx}\right\}$ are bases of the range space of H_d, there exists an orthogonal matrix Q_x satisfying

$$\left[u_1^{cx},\, u_2^{cx}\, \ldots,\, u_{r-1}^{cx}\right] Q_x = \left[u_1^d,\, u_2^d,\, \ldots,\, u_{r-1}^d\right].$$

From (14) it follows that

$$g_{xi}^d = (XX^T)^{-\frac{1}{2}} u_i^d, \quad g_{xi}^{cx} = (XX^T)^{-\frac{1}{2}} u_i^{cx},$$

and therefore

$$\begin{aligned} &[g_{x1}^d, \cdots, g_{x(r-1)}^d] \\ &\quad = (XX^T)^{-\frac{1}{2}} [u_1^d,\, u_2^d,\, \ldots,\, u_{r-1}^d] \\ &\quad = (XX^T)^{-\frac{1}{2}} \left[u_1^{cx},\, u_2^{cx}\, \ldots,\, u_{r-1}^{cx}\right] Q_x \\ &\quad = [g_{x1}^{cx}, \cdots, g_{x(r-1)}^{cx}] Q_x. \end{aligned}$$

This shows that the projective directions for X in DCCA(X, Y) are equivalent to the projective directions for X in CCA(X, C) with respect to an orthogonal transformation.

Similarly, working with H_d^T instead of H_d, we can get an orthogonal matrix Q_y such that

$$\left[u_1^{cy},\, u_2^{cy}\, \ldots,\, u_{r-1}^{cy}\right] Q_y = \left[v_1^d,\, v_2^d,\, \ldots,\, v_{r-1}^d\right].$$

This implies that the projective directions for Y in DCCA(X, Y) are equivalent to the projective directions for Y in CCA(Y, C) with respect to an orthogonal transformation.

3.2. CCA and LDA. The relation between CCA and LDA can complete the analysis of the relation between DCCA and LDA. Recall from (7) that g_x^{cx} is a solution obtained from CCA(X,C) such that

$$XC^T \left(CC^T\right)^{-1} CX^T g_x^{cx} = \lambda XX^T g_x^{cx}. \tag{15}$$

From (13), we get

$$\begin{aligned} &XC^T \left(CC^T\right)^{-1} CX^T \\ &\quad = \hat{X} \begin{bmatrix} \frac{1}{n_1} & & \\ & \ddots & \\ & & \frac{1}{n_r} \end{bmatrix} \hat{X}^T \\ &\quad = \sum_{i=1}^{r} n_i \left(\overline{x^i} - \overline{x}\right) \left(\overline{x^i} - \overline{x}\right)^T \\ &\quad = S_b, \end{aligned}$$

and $S_t = XX^T$. Hence (15) becomes

$$S_b g_x^{cx} = \lambda S_t g_x^{cx},$$

and, since $S_t = S_b + S_w$,

$$\begin{aligned} S_b g_x^{cx} &= \lambda S_t g_x^{cx}, \\ S_b g_x^{cx} &= \lambda \left(S_b + S_w\right) g_x^{cx}, \\ S_b g_x^{cx} &= \frac{\lambda}{1-\lambda} S_w g_x^{cx}. \end{aligned}$$

Accordingly, the vector g_x^{cx} from CCA(X,C) becomes the solution obtained by LDA(X) (Sun and Chen, 2007).

3.3. Equivalence of DCCA(X, Y) and LDA(X) + LDA(Y). Combining the discussions from Sections 3.1 and 3.2, we can conclude that the projective directions obtained from DCCA(X,Y) are equal to the projective directions by LDA performed on X and Y with respect to orthogonal transformations.

For a comparison of computational complexities, major computational procedures in DCCA(X,Y) and LDA(X) + LDA(Y) are the computation of inverse matrices and singular value decomposition. The computation of inverse matrices is needed for both DCCA(X,Y) and LDA(X) + LDA(Y). While DCCA(X,Y) requires only the SVD of H in (12), LDA(X) and LDA(Y) have to perform the SVD of $S_w^{-1} S_b$ separately. Hence DCCA is computationally more efficient than LDA(X) + LDA(Y).

4. Improved method

The scatterness of data points within a class is calculated using deviations from the class mean. When class distributions have normal distributions, class means can well represent data points in a class. But when the shape of a

Table 1. 2D artificial data.

Feature set	feature	class 1 (blue o)	class 2 (green x)	class 3 (red △)
X	x_1	[0, 0.7]	[0, 2]	[0.5, 1.2]
	x_2	[-1, 0]	[0.5, 1.5]	[-0.5, 0.3]
Y	y_1	[0.5, 1.5]	[0, 0.5]	[0.4, 0.8]
	y_2	[0.5, 1.5]	[0, 0.5]	[0, 0.2]

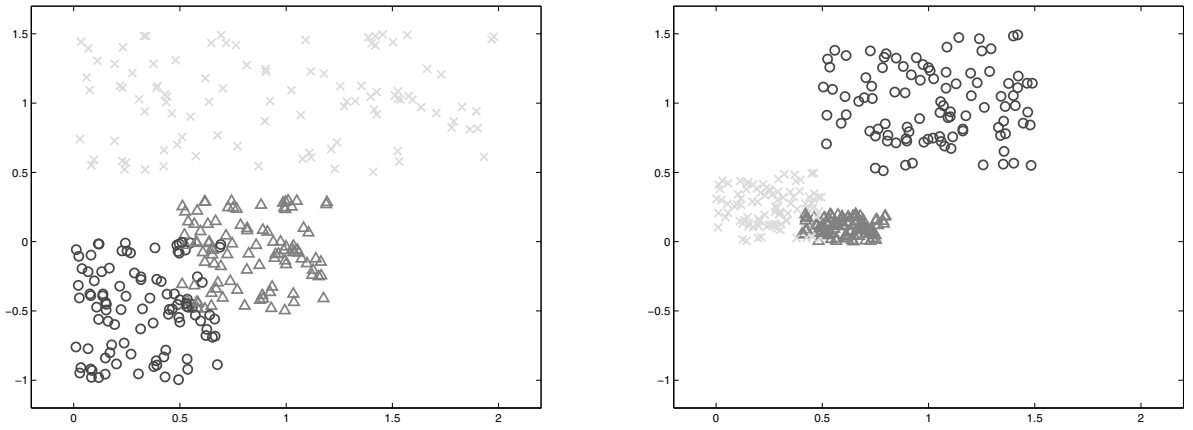


Fig. 1. Visualization of artificially generated data. The left panel displays the data samples by the features in X and the right one draws them using the features in Y.

class is complex, this may fail. In this section, we propose an improved method for feature extraction which is especially effective for data with non-normal class distributions.

We consider using *within-class nearest neighborhood scatter*, denoted as S_{nw}, that is a measure of within-class distribution using nearest neighbors (Fukunaga and Mantock, 1983). S_{nw} is defined as

$$S_{nw} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(x_j^i - \overline{z_{x_j^i}} \right) \left(x_j^i - \overline{z_{x_j^i}} \right)^T, \qquad (16)$$

where $z_{x_j^i}^1, \ldots, z_{x_j^i}^k$ is the k-nearest neighbors of x_j^i among the data points in the same class as x_j^i and

$$\overline{z_{x_j^i}} = \frac{1}{k} \sum_{l=1}^{k} z_{x_j^i}^l.$$

Within-class nearest neighborhood scatters in the projection space of X and Y become $g_x^T S_{nw}^X g_x$, $g_y^T S_{nw}^Y g_y$, when S_{nw}^X and S_{nw}^Y are the within-class nearest neighborhood scatters in X and Y, respectively. These scatters should become smaller for the improvement of classification performance. Hence the objective function of DCCA can be modified as the problem of maximizing

$$f(g_x, g_y, \lambda) = g_x^T C_w g_y - \frac{\lambda}{2} \left(g_x^T S_{nw}^X g_x + g_y^T S_{nw}^Y g_y \right).$$

By taking the derivatives of $f(g_x, g_y, \lambda)$ with respect to g_x and g_y and setting them as zeros, we obtain

$$\frac{\partial f}{\partial g_x} = C_w g_y - \lambda S_{nw}^X g_x = 0, \qquad (17)$$

$$\frac{\partial f}{\partial g_y} = C_w^T g_x - \lambda S_{nw}^Y g_y = 0. \qquad (18)$$

From (18), we get

$$g_y = \frac{1}{\lambda} \left(S_{nw}^Y \right)^{-1} C_w^T g_x. \qquad (19)$$

Using g_y with (19) in (17), we obtain

$$C_w \left(S_{nw}^Y \right)^{-1} C_w^T g_x = \lambda^2 S_{nw}^X g_x.$$

By applying the same procedure for g_y, we have eigenvalue equations such that

$$C_w \left(S_{nw}^Y \right)^{-1} C_w^T g_x = \lambda^2 S_{nw}^X g_x,$$
$$C_w^T \left(S_{nw}^X \right)^{-1} C_w g_y = \lambda^2 S_{nw}^Y g_y.$$

Solving the above equations, $r-1$ pairs of eigenvectors (g_x, g_y) corresponding to the largest eigenvalues give the solution for the proposed method. By minimizing the within-class scatter using nearest neighbors, the proposed method can make more powerful performance in classification problems with complex class structures.

5. Experimental results

5.1. DCCA(X, Y) and LDA(X) + LDA(Y). In order to illustrate the relationship between DCCA(X,Y) and LDA(X) + LDA(Y), we generated artificial data. They are composed of three classes, each of which has 100 data samples. The data are described by two feature sets X and Y which have two features, respectively. For each feature, data samples were uniformly distributed in the ranges described in Table 1. The left part of Fig. 1 displays the data samples by the feature set X and the right one draws them using the features in Y. We performed DCCA(X,Y) and LDA(X) + LDA(Y) for the data in Table 1. Since the number of classes is three, $\{g^d_{x1}, g^d_{x2}\}$ and $\{g^d_{y1}, g^d_{y2}\}$ from DCCA(X,Y), $\{g^x_{x1}, g^x_{x2}\}$ from LDA(X), and $\{g^y_{y1}, g^y_{y2}\}$ from LDA(Y) can be computed. As we proved in Section 3, we can show the equivalence relationship between $\{g^d_{x1}, g^d_{x2}\}$ and $\{g^x_{x1}, g^x_{x2}\}$ for an orthogonal matrix Q_1 such that

$$\begin{bmatrix} g^d_{x1} & g^d_{x2} \end{bmatrix} = Q_1 \begin{bmatrix} g^x_{x1} & g^x_{x2} \end{bmatrix},$$

$$\begin{bmatrix} -0.0369 & -0.1457 \\ -0.0685 & 0.0709 \end{bmatrix} = \begin{bmatrix} 0.9858 & 0.1678 \\ -0.1678 & 0.9858 \end{bmatrix} \begin{bmatrix} -0.0119 & -0.1499 \\ 0.0794 & 0.0585 \end{bmatrix}.$$

Similarly, an orthogonal matrix Q_2 completes the relation between $\{g^d_{y1}, g^d_{y2}\}$ and $\{g^y_{y1}, g^y_{y2}\}$ such that

$$\begin{bmatrix} g^d_{y1} & g^d_{y2} \end{bmatrix} = Q_2 \begin{bmatrix} g^y_{y1} & g^y_{y2} \end{bmatrix},$$

$$\begin{bmatrix} 0.1375 & -0.1472 \\ 0.0290 & 0.1686 \end{bmatrix} = \begin{bmatrix} -0.8827 & -0.4699 \\ -0.4699 & 0.8827 \end{bmatrix} \cdot \begin{bmatrix} -0.0523 & -0.1946 \\ -0.1048 & 0.1352 \end{bmatrix}.$$

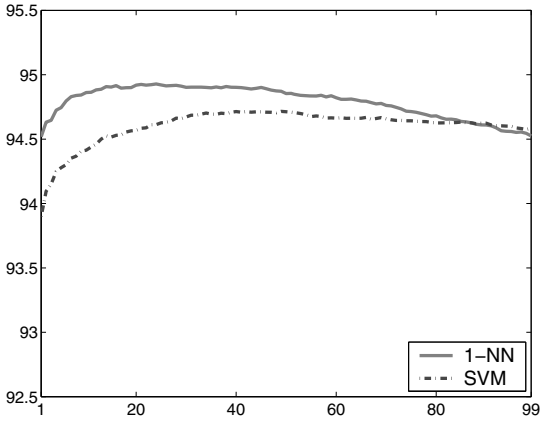


Fig. 2. Sensitivity for the parameter k. The value of k is varied from 1 to 99 on the x-axis and the average prediction accuracies by new DCCA are shown on the y-axis.

5.2. Performance comparison. Using both artificial and real data, we compared the performances of the proposed method denoted as new DCCA with CCA, DCCA and PLS (Wegelin, 2000; Garthwaite, 1994), which is a feature fusion method by partial least squares. The test is proceeded through in three steps as follows:

1. Compute projection vectors $\{g^d_{x1}, \ldots, g^d_{x(r-1)}\}$ and $\{g^d_{y1}, \ldots, g^d_{y(r-1)}\}$ using the training set by DCCA(X, Y), where r is the number of classes.

2. The training data are projected onto each projection vectors, and by concatenating the projected vectors as

$$\{[g^d_{x1}, \ldots, g^d_{x(r-1)}]^T X, [g^d_{y1}, \ldots, g^d_{y(r-1)}]^T Y\}$$

   $2(r-1)$-dimensional vectors are obtained. For test data, they are projected in the same way.

3. The test data are classified in the projected space. We applied a 1-nearest neighbor classifier and a linear support vector machine with the regularization parameter $C = 1$.

For CCA, PLS, and new DCCA, similar steps were followed.

The first test was performed using a multiple feature data set from the UCI machine learning repository (`http://mlearn.ics.uci.edu/MLRepository.html`). The dataset has handwritten digit data that consist of 6 feature sets. It has 10 classes and each class contains 200 data samples. Detailed information of the 6 feature sets is given in Table 2. For the experiments, any 2 feature sets are paired to construct X and Y, and therefore there are 15 pairs of different combinations. For each combination, 100 data samples per class are randomly selected for training and the remaining are used for testing. The splitting to training and testing sets is repeated 10 times and the average prediction accuracies over 10 times running are reported in Table 3. We set the number of neighbors k as 10 for the within-class nearest neighbor scatter in new DCCA. However, Fig. 2 shows that the number of neighbors did not make great impacts on classification performance. In Fig. 2, when k was ranged from 10 to 50, average prediction accuracies by new DCCA did not differ greatly. As shown in Table 3, the proposed method is superior to other methods compared in most cases, while DCCA utilizing class label information gives higher prediction accuracies than the unsupervised feature fusion method CCA.

The data for the second experiment were constructed by using several data sets from UCI, which are summarized in Table 4. The goal of the Isolet and Letimg data sets is to recognize alphabet letters by either speech or letter image, and the other data sets are for the recognition of handwritten digits. In order to construct X and Y, we

Table 2. Description of the multiple feature data set.

feature name	feature #	description
fac	216	profile correlations
fou	76	Fourier coefficients of character shapes
kar	64	Karhunen–Love coefficients
mor	6	morphological features
pix	240	pixel averages in 2 x 3 windows
zer	47	Zernike moments

Table 3. Comparison of prediction accuracies (%) for the multiple feature data set.

		by using the 1-NN classifier					by using SVM			
X	Y	CCA	PLS	DCCA	new DCCA		CCA	PLS	DCCA	new DCCA
fac	fou	81.3	93.7	98.2	98.6		93.0	96.4	97.1	98.4
fac	kar	93.5	93.6	97.7	97.9		95.9	96.4	97.1	97.7
fac	mor	75.9	88.0	93.8	97.9		80.8	93.0	91.4	97.6
fac	pix	83.1	93.7	97.4	97.8		97.1	96.7	96.8	97.4
fac	zer	85.9	95.4	97.7	97.7		88.9	96.9	97.6	97.6
fou	kar	90.2	96.9	97.2	98.3		96.4	95.9	95.0	97.3
fou	mor	75.9	43.4	81.6	84.3		79.2	74.1	80.4	83.6
fou	pix	73.8	97.4	96.7	98.4		89.3	96.7	94.8	97.3
fou	zer	80.0	80.9	84.9	86.2		87.2	82.2	84.1	86.6
kar	mor	81.9	62.5	92.6	97.2		84.5	94.6	89.1	96.6
kar	pix	91.4	97.4	94.2	96.6		94.5	96.8	93.5	95.1
kar	zer	91.6	82.6	96.0	96.4		94.7	92.5	95.5	95.9
mor	pix	75.0	71.2	90.9	97.2		79.7	94.5	87.9	96.8
mor	zer	73.7	71.8	81.1	81.7		79.1	79.0	81.1	82.7
pix	zer	82.3	83.6	95.3	96.7		87.2	93.4	95.0	95.9
average		82.4	83.5	93.0	94.9		88.5	91.9	91.7	94.4

Table 4. Description of letters or digits recognition data sets from UCI.

name	feature #	description
Isolet	617	isolated letter speech recognition
Letimg	16	letter image recognition
Pendigit	16	pen-based recognition of the handwritten digits data set
Optdigit	64	optical recognition of the handwritten digits data set
Semdigit	256	Semeion handwritten digit data set

Table 5. Comparison of prediction accuracies (%) for the pairs constructed from the data in Table 4.

		by using the 1-NN classifier					by using SVM			
X	Y	CCA	PLS	DCCA	new DCCA		CCA	PLS	DCCA	new DCCA
Isolet	Letimg	74.2	96.2	95.7	98.1		80.6	97.5	90.5	96.0
Pendigit	Optdigit	97.1	97.1	98.4	98.6		97.9	98.0	97.2	97.8
Pendigit	Semdigit	91.4	95.6	96.3	97.3		94.0	94.3	94.3	95.9
Optdigit	Semdigit	80.9	96.6	97.2	98.5		95.6	96.6	96.6	98.7

paired any 2 among the data sets having the same goals, making combinations of 4 pairs as in Table 5. For a pair of Isolet and Letimg, we selected 100 data samples for training and 100 data samples for testing randomly from each alphabet class of two data sets. The test was repeated 10 times as in the previous experiment. The same process was performed for the other pairs of data sets. The accuracies are shown in Table 5. It also shows that the proposed method improved the classification accuracy of DCCA.

The last experiment is to test the performance of the proposed method for the data with complex class structures. The data were generated artificially so as to show the difference between DCCA and the new DCCA. The two top row parts of Fig. 3 represent two feature sets X and Y, respectively, where some classes have multiple modes. The projective directions obtained by DCCA

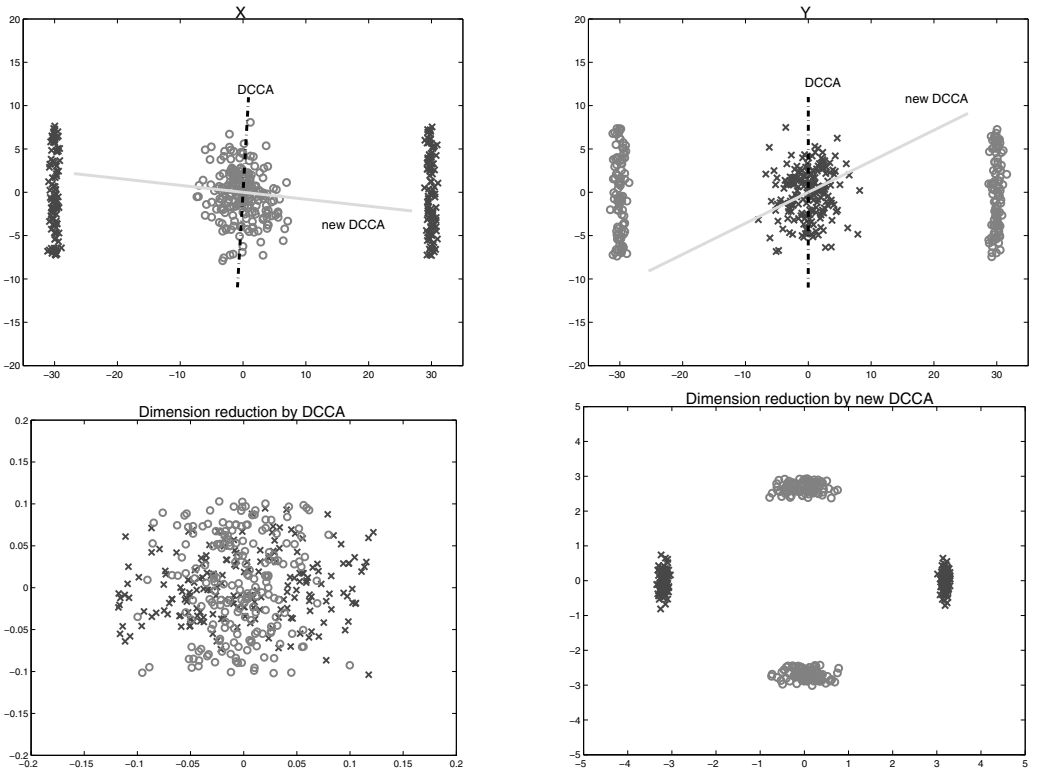


Fig. 3. Comparison of the proposed method with DCCA for the data with complex class structures.

were drawn by black dash-dot lines and the projective directions by new DCCA were drawn by green solid lines. The two bottom row parts of the figure show data points in the reduced dimensional space by DCCA and new DCCA. New DCCA was able to find discriminant directions, since it tried to minimize within-class scatter using local neighborhoods.

6. Conclusion

In this paper, we analyzed the relation between DCCA and LDA, showing that DCCA for two feature sets X and Y can be obtained by applying LDA to X and Y separately up to an orthonormal transformation. We also proposed a new feature extraction method for data sets with nonnormal class distributions. The experimental results demonstrate that the classification performance can be improved by the proposed method.

Acknowledgment

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0003530).

References

Anton, H. and Busby, R. (2003). *Contemporary Linear Algebra*, John Wiley and Sons, Denver, CO.

Baudat, G. and Anouar, F. (2000). Generalized discriminant analysis using a kernel approach, *Neural Computation* **12**(10): 2385–2404.

Billings, S. and Lee, K. (2002). Nonlinear fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm, *Neural Networks* **15**(2): 263–270.

Chen, L., Liao, H., M.Ko, Lin, J. and Yu, G. (2000). A new LDA-based face recognition system which can solve the small sample size problem, *Pattern Recognition* **33**(10): 1713–1726.

Duda, R., Hart, P. and Stork, D. (2001). *Pattern Classification*, Wiley Interscience, New York, NY.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd Edn., Academic Press, San Diego, CA.

Fukunaga, K. and Mantock, J. (1983). Nonparametric discriminant analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5**(6): 671–678.

Garthwaite, P. (1994). An interpretation of partial least squares, *Journal of the American Statistical Society* **89**(425): 122–127.

He, X. and Niyogi, P. (2003). Locality preserving projections, *Proceedings of the Advances in Neural Information Pro-*

cessing Systems Conference, Vancouver, Canada, pp. 153–160.

Hotelling, H. (1936). Relations between two sets of variates, *Biometrika* **28**(3): 321–377.

Hou, C., Nie, F., Zhang, C. and Wu, Y. (2009). Learning an orthogonal and smooth subspace for image classification, *IEEE Signal Processing Letters* **16**(4): 303–306.

Howland, P. and Park, H. (2004). Generalizing discriminant analysis using the generalized singular value decomposition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(8): 995–1006.

Jolliffe, I. (1986). *Principal Component Analysis*, Springer, New York, NY.

Nie, F., Xiang, S. and Zhang, C. (2007). Neighborhood minmax projections, *Proceedings of the International Joint Conference on Artificial Intelligence*, Hyderabad, India, pp. 993–998.

Pardalos, P. and Hansen, P. (2008). *Data Mining and Mathematical Programming*, CRM Proceedings & Lecture Notes, Vol. 45, American Mathematical Society, Montreal.

Park, C. and Park, H. (2008). A comparison of generalized linear discriminant analysis algorithms, *Pattern Recognition* **41**(3): 1083–1097.

Roweis, S.T. and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**(5500): 2323–2326.

Sugiyama, M. (2006). Local fisher discriminant analysis for supervised dimensionality reduction, *Proceedings of the IEEE International Conference on Machine Learning, Pittsburgh, PA, USA*, pp. 905–912.

Sun, Q., Zeng, S., Liu, Y., Heng, P. and Xia, D. (2005). A new method of feature fusion and its application in image recognition, *Pattern Recognition* **38**(12): 2437–2448.

Sun, T. and Chen, S. (2007). Class label versus sample label-based CCA, *Applied Mathematics and Computation* **185**(1): 272–283.

Sun, T., Chen, S., Yang, J. and Shi, P. (2008). A supervised combined feature extraction method for recognition, *Proceedings of the IEEE International Conference on Data Mining, Pisa, Italy*, pp. 1043–1048.

Tenenbaum, J.B., de Silva, V. and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science* **290**(5500): 2319–2323.

Wegelin, J. (2000). A survey of partial least squares (PLS) methods, with emphasis on the two block case, *Technical report*, Department of Statistics, University of Washington, Seattle, WA.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q. and Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionlity reduction, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(1): 40–51.

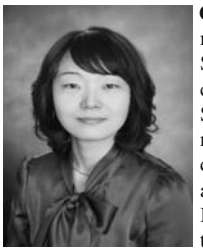
Yang, J. and Yang, J.-Y. (2003). Why can LDA be performed in PCA transformed space?, *Pattern Recognition* **36**(2): 563–566.

Yang, J., Yang, J., Zhang, D. and Lu, J. (2003). Feature fusion: Parallel strategy vs. serial strategy, *Pattern Recognition* **36**(6): 1369–1381.

Ye, J. (2005). Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, *Journal of Machine Learning Research* **6**(4): 483–502.

Yu, H. and Yang, J. (2001). A direct LDA algorithm for high-dimensional data-with application to face recognition, *Pattern Recognition* **34**(10): 2067–2070.

Yong Joon Shin received the B.Eng. degree in computer engineering from Chungnam National University, Korea, in 2009. Now, he is an M.Sc. course student in the Department of Computer Science and Engineering, Chungnam National University. His current research interest is in data mining, pattern recognition, and machine learning.



Cheong He Park received her Ph.D. in mathematics from Yonsei University, Korea, in 1998. She received the M.Sc. and Ph.D. degrees in computer science at the Department of Computer Science and Engineering, University of Minnesota, in 2002 and 2004, respectively. She is currently in the Department of Computer Science and Engineering, Chungnam National University, Korea, as an assistant professor. Her research interests include pattern recognition, data mining, bioinformatics, and machine learning.

Received: 12 May 2010
Revised: 22 October 2010
Re-revised: 5 January 2011