# Relationships Between Support Vector Classifiers and Generalized Linear Discriminant Analysis on Support Vectors[*]

Hyunsoo Kim, Barry L. Drake, Haesun Park [†]

**Abstract.** The linear discriminant analysis based on the generalized singular value decomposition (LDA/GSVD) has been introduced to circumvent the nonsingularity restriction inherent in the classical LDA. The LDA/GSVD provides a framework in which a dimension reducing transformation can be effectively obtained for undersampled problems. In this paper, relationships between support vector machines (SVMs) and the generalized linear discriminant analysis applied to the support vectors are studied. Based on the GSVD, the weight vector of the hard-margin SVM is proved to be equivalent to the dimension reducing transformation vector generated by LDA/GSVD applied to the support vectors of the binary class. We also show that the dimension reducing transformation vector and the weight vector of soft-margin SVMs are related when a subset of support vectors are considered. These results can be generalized when kernelized SVMs and the kernelized LDA/GSVD called KDA/GSVD are considered. Through these relationships, it is shown that support vector classification is related to data reduction as well as dimension reduction by LDA/GSVD.

**Key words.** classifier for binary classes, dimension reduction, generalized SVD, linear discriminant analysis, support vector machines

**AMS subject classifications.** 15A18, 15A23, 68T05

[*] This material is based upon work supported in part by the National Science Foundation Grants CCR-0204109 and ACI-0305543. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Manuscript date: April 27, 2006

[†] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA (hskim@cc.gatech.edu, bldrake@cc.gatech.edu, hpark@cc.gatech.edu).

**1. Introduction.** Support vector machines (SVMs) [28, 29] construct an optimal separating hyperplane that maximizes the margin (i.e. the distance between the hyperplane and the nearest data point of each class) by mapping the input space into a high dimensional feature space. This mapping is implicitly determined by a kernel function. Training with SVMs has crucial advantages including the ability to find the problem formulation as a quadratic convex function minimization that is easier to solve [3, 2, 16, 28]. SVMs have demonstated state-of-the-art performance in numerous application areas.

In Fisher's Discriminant Analysis (FDA), a linear transformation is found which maximizes the trace of the between-class scatter matrix and minimizes the trace of the within-class scatter matrix [6, 5]. Although LDA is conceptually simple and has been used in many application areas, it has a critical limitation: it requires the within-class scatter matrix to be nonsingular. To overcome the nonsingularity restriction, recently, linear discriminant analysis based on the generalized singular values decomposition (LDA/GSVD) has been introduced [9, 10]. The generalized singular value decomposition (GSVD) has been studied extensively in the numerical linear algebra literature [26, 19, 11], and numerical algorithms for computing the GSVD have been widely investigated [27, 18, 4, 1]. The GSVD has been applied to a wide variety of interesting problems including signal processing [7, 24, 22, 15], the positive definite generalized eigenvalue problem [23], the generalized total least squares problem [25], the least squares problem with Tikhonov regularization [8], the constrained least squares problems [7], and the generalized linear model regression problem [17].

In this paper, a mathematical relationship between the hard-margin SVM and LDA/GSVD applied to the support vectors is illustrated. The relationship between the $L_1$-norm soft-margin SVM and LDA/GSVD on a subset of the support vectors is also presented. These results can be generalized when kernelized SVMs and the kernelized LDA/GSVD called KDA/GSVD are considered. Through these relationships, it is shown that support vector classification is related to data reduction as well as dimension reduction by LDA/GSVD.

The following notations will be used in the paper. For a matrix $A \in \mathbb{R}^{m\times n}$, range$(A) = \{\mathbf{y} \in \mathbb{R}^m \mid \mathbf{y} = A\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^n\}$ and null$(A)$ = $\{\mathbf{z} \in \mathbb{R}^n \mid A\mathbf{z} = 0\}$. Assume that we are given a data set $A \in \mathbb{R}^{m\times n}$ with $p$ classes,

$$A = \left( \mathbf{a}_1 \cdots \mathbf{a}_n \right) = \left( A_1 \cdots A_p \right) \in \mathbb{R}^{m\times n},$$

where the $j$th column $\mathbf{a}_j$ of the matrix $A$ denotes the $j$th data item, the columns of submatrix $A_i$ belong to class $i$, for $i = 1 \cdots p$. In addition, let $N_i$ be the set of data items that belong to class

$i$, $n_i$ the number of items in class $i$, $\mathbf{c}_i$ the centroid vector which is the average of all the data in class $i$, and $\mathbf{c}$ the global centroid vector.

**2. Linear and Kernel Discriminant Analysis based on the GSVD and Support Vector Machines.** The goal of linear discriminant analysis (LDA) is to find a dimension reducing transformation that minimizes the scatter within each class and maximizes the scatter between classes. The within-class scatter matrix $S_w$, the between-class scatter matrix $S_b$, and the mixture scatter matrix $S_m$ are defined as

$$\begin{aligned} S_w &= \sum_{i=1}^{p} \sum_{j \in N_i} (\mathbf{a}_j - \mathbf{c}_i)(\mathbf{a}_j - \mathbf{c}_i)^T, \\ S_b &= \sum_{i=1}^{p} n_i (\mathbf{c}_i - \mathbf{c})(\mathbf{c}_i - \mathbf{c})^T, \\ S_m &= \sum_{i=1}^{n} (\mathbf{a}_i - \mathbf{c})(\mathbf{a}_i - \mathbf{c})^T. \end{aligned}$$

In addition,

$$\mathrm{trace}(S_w) = \sum_{i=1}^{p} \sum_{j \in N_i} \|\mathbf{a}_j - \mathbf{c}_i\|_2^2 \quad \text{and} \quad \mathrm{trace}(S_b) = \sum_{i=1}^{p} n_i \|\mathbf{c}_i - \mathbf{c}\|_2^2.$$

Defining the matrices

$$H_w = [A_1 - \mathbf{c}_1 \mathbf{e}_1^T, \ldots, A_p - \mathbf{c}_p \mathbf{e}_p^T] \in \mathbf{R}^{m \times n}, \tag{2.1}$$

where $\mathbf{e}_i = [1, \ldots, 1]^T \in \mathbf{R}^{n_i \times 1}$,

$$H_b = [\sqrt{n_1}(\mathbf{c}_1 - \mathbf{c}), \ldots, \sqrt{n_p}(\mathbf{c}_p - \mathbf{c})] \in \mathbb{R}^{m \times p}, \tag{2.2}$$

and

$$H_m = [\mathbf{a}_1 - \mathbf{c}, \mathbf{a}_2 - \mathbf{c}, \ldots, \mathbf{a}_n - \mathbf{c}] \in \mathbb{R}^{m \times n}, \tag{2.3}$$

we have

$$S_w = H_w H_w^T, \;\; S_b = H_b H_b^T, \;\; \text{and} \;\; S_m = H_m H_m^T.$$

Assume that $W \in \mathbb{R}^{m \times l}$ denotes the transformation that maps a vector in an $m$ dimensional space to a vector in an $l$ dimensional space. In the reduced dimensional space obtained by the

TABLE 2.1

*Generalized eigenvalues $\lambda_i$'s and eigenvectors $x_i$'s from the GSVD. The superscript $c$ denotes the complement.*

| | $\eta_i$ | $\zeta_i$ | $\lambda_i = \frac{\eta_i^2}{\zeta_i^2}$ | $x_i$ belongs to |
|---|---|---|---|---|
| $1 \le i \le \mu$ | 1 | 0 | $\infty$ | null($S_w$) $\cap$ null($S_b$)$^c$ |
| $\mu+1 \le i \le \mu+\tau$ | $1 > \eta_i > 0$ | $0 < \zeta_i < 1$ | $\infty > \lambda_i > 0$ | null($S_w$)$^c$ $\cap$ null($S_b$)$^c$ |
| $\mu+\tau+1 \le i \le t$ | 0 | 1 | 0 | null($S_w$)$^c$ $\cap$ null($S_b$) |
| $t+1 \le i \le n$ | any value | any value | any value | null($S_w$) $\cap$ null($S_b$) |

dimension reducing transformation $W$, a scatter matrix $S$ becomes $W^TSW$. When $S_w$ is nonsingular, minimizing trace$(W^TS_wW)$ and maximizing trace$(W^TS_bW)$ is commonly approximated by maximizing

$$J_1(W) = \text{trace}((W^TS_wW)^{-1}(W^TS_bW)). \tag{2.4}$$

It is well known that the columns of $W$ that maximizes $J_1(W)$ are the leading $l$ eigenvectors of $S_w^{-1}S_b$, where $l = \text{rank}(H_b) = \text{rank}(S_w^{-1}S_b) \le p-1$ [6] and this provides the foundation of the classfical LDA.

In Fisher's discriminant analysis (FDA) which is a special case of LDA for two-class problems, when $S_w$ is singular, the dimension reducing transformation $W$ is a vector $\mathbf{w}$ which is given as

$$\mathbf{w} = S_w^{-1}(\mathbf{c}_1 - \mathbf{c}_2).$$

When the number of features is larger than the number of the data points ($m > n$), the classical FDA cannot be applied since $S_w$ is singular. Linear discriminant analysis with the generalized singular value decomposition (LDA/GSVD) [9] circumvents this nonsingularity restriction so that it can effectively reduce dimension even when $m > n$ which we call *undersampled* for multi-class problems, and it also provides the solution for under-sampled binary class problems as well.

We briefly review the generalized singular decomposition (GSVD).

THEOREM 2.1 (C.C. Paige and M.A. Saunders [19]: Generalized Singular Value Decomposition (GSVD)). *Suppose two matrices with the same number of columns, $K_b \in \mathbb{R}^{p_1 \times q}$ and*

*$K_w \in \mathbb{R}^{p_2 \times q}$, are given. Then there exist orthogonal matrices $U_b \in \mathbb{R}^{p_1 \times p_1}$ and $U_w \in \mathbb{R}^{p_2 \times p_2}$, and a nonsingular matrix $X \in \mathbb{R}^{q \times q}$ such that*

$$U_b^T K_b^T X = [\Sigma_b \;\; 0] \;\; \textit{and} \;\; U_w^T K_w^T X = [\Sigma_w \;\; 0]\,,$$

*where*

$$\Sigma_b = \begin{array}{l} \left[\begin{array}{ccc} I_b & & \\ & D_b & \\ & & 0_b \end{array}\right] \begin{array}{l} \}\mu \\ \}\tau \\ \}p_1 - \mu - \tau \end{array} \\ \;\;\underbrace{\qquad}_{\mu}\underbrace{\qquad}_{\tau}\underbrace{\qquad}_{t-\mu-\tau} \end{array}$$

*and*

$$\Sigma_w = \begin{array}{l} \left[\begin{array}{ccc} 0_w & & \\ & D_w & \\ & & I_w \end{array}\right] \begin{array}{l} \}p_2 - t + \mu \\ \}\tau \\ \}t - \mu - \tau \end{array} \\ \;\;\underbrace{\qquad}_{\mu}\underbrace{\qquad}_{\tau}\underbrace{\qquad}_{t-\mu-\tau} \end{array} \quad ,$$

$$t = \mathrm{rank}(\left[\begin{array}{c} K_b \\ K_w \end{array}\right]), \quad \mu = t - \mathrm{rank}(K_w) \; and \; \tau = \mathrm{rank}(K_b) + \mathrm{rank}(K_w) - t,$$

$$D_b = diag(\eta_{\mu+1}, \ldots, \eta_{\mu+\tau}) \quad \textit{and} \quad D_w = diag(\zeta_{\mu+1}, \ldots, \zeta_{\mu+\tau})$$

*for which*

$$\begin{array}{c} 1 > \eta_{\mu+1} \geq \cdots \geq \eta_{\mu+\tau} > 0, \\ 0 < \zeta_{\mu+1} \leq \cdots \leq \zeta_{\mu+\tau} < 1, \end{array}$$

*and*

$$\eta_i^2 + \zeta_i^2 = 1 \; for \; i = 1, \ldots, \mu + \tau.$$

*Proof.* See Paige and Saunders [19]. □

In the LDA/GSVD algorithm, the GSVD of the matrix $(H_b, H_w)^T$ is computed, where $H_w$ and $H_b$ are defined in Eqn. (2.1) and Eqn. (2.2), respectively. Then from the GSVD, the generalized singular vectors $X$ are obtained which satisfy

$$X^T S_b X = X^T H_b H_b^T X \; = \begin{bmatrix} I_\mu & & & \\ & D_b^T D_b & & \\ & & 0_{t-\mu-\tau} & \\ & & & 0_{m-t} \end{bmatrix} \tag{2.5}$$

and

$$X^T S_w X = X^T H_w H_w^T X = \begin{bmatrix} 0_\mu & & & \\ & D_w^T D_w & & \\ & & I_{t-\mu-\tau} & \\ & & & 0_{m-t} \end{bmatrix}, \tag{2.6}$$

where the subscripts in $I$ and 0 denote the order of identity and zero matrices, respectively, and the order of each matrix is denoted by its subscript. From Eqn. (2.5) and Eqn. (2.6), the generalized eigenvalues and eigenvectors obtained by the GSVD are classified as shown in Table 2.1 [20]. The LDA/GSVD takes the leading $p-1$ generalized eigenvectors and these include all the eigenvectors in null($S_w$) $\cap$ null($S_b$)$^c$ and null($S_w$)$^c$ $\cap$ null($S_b$)$^c$.

To discuss the relationship between linear and nonlinear discriminant analysis and SVMs, let us first review the hard-margin SVM. The decision rule for binary classification in a hard margin SVM is given by $sign(f(\mathbf{x}))$ with

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \beta \tag{2.7}$$

based on the following optimization problelm. The training data $(\mathbf{a}_i, y_i)$ with $y_i \in \{-1, +1\}$ for $1 \le i \le n$, $f(\mathbf{x})$ is obtained by solving an optimization problem called the hard-margin SVM

$$\begin{aligned} &\min_{\mathbf{w},b} \; \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &s.t. \;\; y_i\left[\mathbf{w}^T\mathbf{a}_i + \beta\right] \ge 1, \; i = 1, \ldots, n. \end{aligned} \tag{2.8}$$

The dual formulation of Eqn. (2.8) is

$$\begin{aligned} &\max_{\alpha} \;\; \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{a}_i^T\mathbf{a}_j \\ &s.t. \;\; \sum_{i=1}^{n} \alpha_i y_i = 0, \; \alpha_i \ge 0, \; i = 1, \ldots, n. \end{aligned} \tag{2.9}$$

Then, the weight vector of the hard-margin SVMs is

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{a}_i, \tag{2.10}$$

where $\alpha_i$, $1 \leq i \leq n$, are solutions of Eqn. (2.9). Hence, in the expression for the weight vector only data points that have corresponding nonzero $\alpha$ are involved, and these are called the support vectors.

**3. Relationship between Hard-margin Support Vector Classifier and Generalized Linear Discriminant Analysis on Support Vectors.** In the rest of this paper, let us assume that the data set has only two classes in studying the relationships between SVM and LDA/GSVD applied to the support vectors. Dsenote the data set of support vectors as matrix $B$:

$$B = \left( \begin{array}{c} \mathbf{b}_1 \cdots \mathbf{b}_s \end{array} \right) = \left( \begin{array}{cc} B_1 & B_2 \end{array} \right) \in \mathbb{R}^{m \times s},$$

where $\mathbf{b}_j$ denotes the $j$th column of the matrix $B$, $N_i$ the set of column indices that belong to the class $i$, $s_i \geq 1$ the number of columns in $B_i$, $i = 1, 2$, and columns of submatrices $B_1$ and $B_2$ are support vectors which belong to classes 1 and 2, respectively. We also assume that the two centroid vectors are linearly independent.

Now we apply the LDA/GSVD to the set of support vectors, $B$. Then the dimension reducing transformation $\mathbf{w} \in \mathbb{R}^{n \times 1}$ will be obtained from applying the GSVD to the matrix pair $(H_b, H_w)$, where

$$H_w = [\mathbf{b}_1 - \mathbf{c}_1, \ldots, \mathbf{b}_{s_1} - \mathbf{c}_1, \mathbf{b}_{s_1+1} - \mathbf{c}_2, \ldots, \mathbf{b}_s - \mathbf{c}_2] \in \mathbb{R}^{m \times s} \tag{3.1}$$

and

$$H_b = [\sqrt{s_1}(\mathbf{c}_1 - \mathbf{c}), \sqrt{s_2}(\mathbf{c}_2 - \mathbf{c})] \in \mathbb{R}^{m \times 2}, \tag{3.2}$$

where $\mathbf{c}_1$ and $\mathbf{c}_2$ are the centroid vectors and $\mathbf{c}$ a global centroid vector for all support vectors.

Following the notation of Theorem 2.1, we have

$$\mathrm{rank}(H_w) = t - \mu \;\text{ and }\; \mathrm{rank}(H_b) = \mu + \tau, \tag{3.3}$$

where

$$t = \mathrm{rank}\left( \left[ \begin{array}{c} H_b^T \\ H_w^T \end{array} \right] \right),$$

$\mu$ is the number of infinite generalized singular values, and $\tau$ is the number of finite nonzero generalized singular values.

The vectors in null$(S_w) \cap$ null$(S_b)$ do not convey any discriminant information among the classes and those in null$(S_w)^c \cap$ null$(S_b)$ make the within class relationship more remote without changing the trace of the between class scatter. Therefore, $\mu + \tau$ leading generalized eigenvectors that belong to null$(S_w) \cap$ null$(S_b)^c$ or null$(S_w)^c \cap$ null$(S_b)^c$ are chosen in LDA/GSVD. We already know the following inequality,

$$\mu + \tau = \mathrm{rank}(H_b) \le p - 1,$$

where $p$ is the number of classes. Since we assume that the two centroid vectors are linearly independent, we have

$$\mu + \tau = \mathrm{rank}(H_b) = p - 1.$$

Note that for binary class problems, the LDA/GSVD solution is the single leading generalized singular vector.

Since $\mathrm{rank}(H_b) = 1$ and $\mu + \tau = 1$, one of the following two cases are possible: If $\mu = 0$, the solution vector $\mathbf{w}$ from LDA/GSVD comes from null$(S_w)^c \cap$ null$(S_b)^c$. If $\mu \neq 0$, then since $\mu = 1$ and $\tau = 0$, the solution vector $\mathbf{w}$ from LDA/GSVD comes from null$(S_w) \cap$ null$(S_b)^c$. In the following, we show that the solution vector $\mathbf{w}$ always belongs to null$(S_w) \cap$ null$(S_b)^c$. Let us denote the number of *different* support vectors as $s$. The results are shown by considering the following two cases:

$$\begin{cases} \text{Case I}: & m \ge 2 \ \text{ and } \ s \ge 3 \\ \text{Case II}: & m = 1 \ \text{ or } \ s = 2. \end{cases} \tag{3.4}$$

First, let us consider Case I: $m \ge 2$ and $s \ge 3$. The rank of $H_b$ is 1 since the vectors $\mathbf{c}_1 - \mathbf{c}$ and $\mathbf{c}_2 - \mathbf{c}$ are linearly dependent for two class problems, dim$(range(S_b))$ = $\mathrm{rank}(S_b) = \mathrm{rank}(H_b H_b^T) = 1$. For the rank of $H_w$, the following theorem can be found.

LEMMA 3.1. *[Rank of $H_w$ on the Support Vectors] For $H_w \in \mathbf{R}^{m \times s}$ shown in Eqn. (3.1) and defined on the support vectors $\mathbf{b}_i$, for $i = 1, \dots, s$ which are the training vectors for which $\alpha_i \neq 0$ in the hard-margin SVM shown in Eqn. (2.9) in a binary classification problem, we have*

$$\mathrm{rank}(H_w) \le \min(s - 1, m - 1). \tag{3.5}$$

*Proof.* If $s < m$, then $\text{rank}(H_w) \leq s - 1$ since the columns of $H_w$ are linearly dependent. If $s \geq m$, $\text{rank}(S_w) = \text{rank}(H_w H_w^T) = \text{rank}(H_w) \leq m - 1$ since all support vectors lie on the hyperplanes, i.e.

$$\mathbf{w}^T \mathbf{b}_i + \beta = \pm 1. \tag{3.6}$$

□

Now the following result regarding rank$((H_b, H_w)^T)$ can be obtained.

LEMMA 3.2 (Relationship between Rank of $H_w$ and Rank of $(H_b, H_w)^T$ defined on the Support Vectors). *For $H_w$ and $H_b$ shown in Eqn. (3.1) and Eqn. (3.2), respectively, and defined for the support vectors $\mathbf{b}_i$, for $i = 1, \ldots, s$, for which the corresponding $\alpha_i \neq 0$ in the hard-margin SVM, shown in Eqn. (2.9), we have*

$$t = \text{rank}\left(\begin{bmatrix} H_b^T \\ H_w^T \end{bmatrix}\right) = \text{rank}(H_w) + 1. \tag{3.7}$$

*Proof.* By Lemma 3.1, $H_w$ is rank deficient. Therefore, we only need to show that appending $H_b$ increses the rank by one. Suppose $\mathbf{c} - \mathbf{c_1} \in Range(H_w)$, i.e., there are scalars $\delta_i$ for which

$$\mathbf{c_1} - \mathbf{c} = \sum_{i \in N_1} \delta_i (\mathbf{b_i} - \mathbf{c_1}) + \sum_{i \in N_2} \delta_i (\mathbf{b_i} - \mathbf{c_2}). \tag{3.8}$$

From Eqn. (3.6), we have

$$\mathbf{w}^T \mathbf{b_i} = -1 - \beta \;\text{ for }\; i \in N_1 \;\text{ and }\; \mathbf{w}^T \mathbf{b_i} = 1 - \beta \;\text{ for }\; i \in N_2.$$

In addition,

$$\mathbf{w}^T \mathbf{c_1} = -1 - \beta \;\text{ and }\; \mathbf{w}^T \mathbf{c_2} = 1 - \beta$$

Using the fact $\mathbf{c_1} - \mathbf{c} = s_2(\mathbf{c_1} - \mathbf{c_2})/s$ and multiplying $\mathbf{w}^T$ to Eqn. (3.8), we obtain $2s_2/s = 0$ which is a contraction. Therefore, $\mathbf{c_1} - \mathbf{c}$ is linearly independent from the columns of $H_w$. The same result holds for $\mathbf{c_2} - \mathbf{c}$ and since rank$(H_b)$ = 1, the Lemma holds. □

From Lemma 3.2 and Theorem 2.1, we have

$$\mu = t - \text{rank}(H_w) = 1 \;\text{ and }\; \tau = \text{rank}(H_b) + \text{rank}(H_w) - t = 1 - \mu = 0. \tag{3.9}$$

Therefore, the solution vector $\mathbf{w}$ from LDA/GSVD belongs to null$(S_w) \cap$ null$(S_b)^c$ since $\mu \neq 0$ in Eqn. (2.6). Accordingly, we have

$$X^T S_b X = \begin{bmatrix} 1 & & \\ & 0_{t-1} & \\ & & 0_{m-t} \end{bmatrix} \tag{3.10}$$

and

$$X^T S_w X = \begin{bmatrix} 0 & & \\ & I_{t-1} & \\ & & 0_{m-t} \end{bmatrix}, \tag{3.11}$$

where the subscripts in $I$ and 0 denote the order of identity and zero matrices. From Eqn. (3.10) and Eqn. (3.11), the generalized eigenvalues and eigenvectors obtained by GSVD in two-class problems are classified as shown in Table 3.1. The table illustrates that the single generalized eigenvector for binary class problems belongs to null($S_w$) $\cap$ null($S_b$)$^c$.

Next, let us consider the Case II, i.e. $m = 1$ or $s = 2$. When $m = 1$, the input data items are scalars and the result is obvious. When $s = 2$, there is only one representative support vector for each class. Hence, the centroid vectors of two classes, $\mathbf{c}_1$ and $\mathbf{c}_2$, are the two different support vectors among all possibly duplicated support vectors. By definition of $H_w$ and $H_b$ in Eqn. (2.1) and Eqn. (2.2), respectively, $H_w$ will be an $m$-by-$s$ zero matrix. Again, the solution vector $\mathbf{w}$ from LDA/GSVD belongs to null($S_w$) $\cap$ null($S_b$)$^c$ since $\mu \neq 0$ in Eqn. (2.5)-Eqn. (2.6). The following relations can be obtained:

$$X^T S_b X = X^T H_b H_b^T X = \begin{bmatrix} 1 & \\ & 0_{m-1} \end{bmatrix} \tag{3.12}$$

and

$$X^T S_w X = X^T H_w H_w^T X = \begin{bmatrix} 0 & \\ & 0_{m-1} \end{bmatrix}, \tag{3.13}$$

where the subscripts in 0 denote the order of the zero matrix. It is shown that the solution vector $\mathbf{w}$ obtained by applying LDA/GSVD to the support vectors is the only vector which belongs to null($S_w$) $\cap$ null($S_b$)$^c$ for both Case I and Case II. The above results are summarized in the following theorem.

THEOREM 3.3 (Dimension Reduction by LDA/GSVD on Support Vectors of Binary Classes). *Consider only the set of support vectors* $\mathbf{b}_i$*, for* $i = 1, \ldots, s$*, for which the corresponding* $\alpha_i \neq 0$ *in the hard-margin SVM, Eqn. (2.9), for a binary classification problem. Then, null(*$S_w$*)* $\cap$ *null(*$S_b$*)*$^c$ = $\{\mathbf{w}\}$*, where* $\mathbf{w}$ *is the solution vector obtained by LDA/GSVD on these support vectors.*

*Proof.* By Lemma 3.2, $\mu = t - \text{rank}(H_w) = 1$ and $\tau = \text{rank}(H_b) - \mu = 1 - 1 = 0$, since $\text{rank}(H_b) = 1$. Therefore, the solution vector $\mathbf{w}$ obtained by LDA/GSVD is the only vector which belongs to null($S_w$) $\cap$ null($S_b$)$^c$. □

TABLE 3.1

*Generalized eigenvalues $\lambda_i$'s and eigenvectors $x_i$'s from the GSVD when $S_w$ and $S_b$ are built on support vectors in binary class problems for Case I, i.e. $m > 1$ and $s > 2$. See Section 3 for the notations. The superscript $c$ denotes the complement.*

| | $\eta_i$ | $\zeta_i$ | $\lambda_i = \frac{\eta_i^2}{\zeta_i^2}$ | $x_i$ belongs to |
|---|---|---|---|---|
| $i = 1$ | 1 | 0 | $\infty$ | null($S_w$) $\cap$ null($S_b$)$^c$ |
| $2 \leq i \leq t$ | 0 | 1 | 0 | null($S_w$)$^c$ $\cap$ null($S_b$) |
| $t + 1 \leq i \leq n$ | any value | any value | any value | null($S_w$) $\cap$ null($S_b$) |

We illustrate a simple case of $m = 2$ and $s > 2$. Then, $\text{rank}(H_w) = 1$, $t = \text{rank}(H_w)+1 = 2$, $\mu = 2 - 1 = 1$, $\tau = 1 - \mu = 0$, and

$$X^T S_b X = X^T H_b H_b^T X = \begin{bmatrix} 1 & \\ & 0 \end{bmatrix} \tag{3.14}$$

$$X^T S_w X = X^T H_w H_w^T X = \begin{bmatrix} 0 & \\ & 1 \end{bmatrix}. \tag{3.15}$$

Clearly the leading generalized eigenvector belongs to null($S_w$) $\cap$ null($S_b$)$^c$.

In the following theorem, it is shown that the weight vector $\mathbf{w}$ from the hard-margin SVM belongs to null($S_w$) $\cap$ null($S_b$)$^c$. Therefore, it provides a relationship between the hard-margin SVM and the LDA/GSVD on the support vectors.

THEOREM 3.4 (Relationship between the Hard-margin SVM and Generalized Linear Discriminant Analysis on Support Vectors). *Consider the set of support vectors $\mathbf{b}_i$, for $i = 1, \dots, s$, for which the corresponding $\alpha_i \neq 0$ in the hard-margin SVM, Eqn. (2.9), for a binary classification problem. Then, the weight vector $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{b}_i$ is the single vector in null($S_w$) $\cap$ null($S_b$)$^c$ where $S_w$ and $S_b$ are the scatter matrices for the support vectors.*

*Proof.* Consider the support vectors, $\mathbf{b}_i \in \mathbb{R}^{n \times 1}$, $1 \leq i \leq s$. Then they lie on the hyperplanes, i.e.

$$\mathbf{w}^T \mathbf{b}_i + \beta = \pm 1,$$

where $\mathbf{w} = \sum_{i=1}^{s} \alpha_i y_i \mathbf{b}_i$, $\alpha_i > 0$, $1 \leq i \leq s$, is the solution to the quadratic programming problem of Eqn. (2.9), and $\beta = (-s_1 + s_2)/s - \mathbf{w}^T \mathbf{c}$. Define a matrix $B = [B_1 \;\; B_2] \in \mathbb{R}^{m \times s}$

whose columns $\mathbf{b}_i$ are the support vectors and $B_1$ and $B_2$ contain support vectors for class 1 and class 2, respectively. Let $\mathbf{c}$ denote the global centroid vector for all support vectors, and $\mathbf{c}_1$ and $\mathbf{c}_2$ the centroid vectors for classes 1 and 2 respectively. Then, the weight vector $\mathbf{w}$ satisfies the following equations:

$$B^T\mathbf{w} = \mathbf{y} - \beta\mathbf{e}, \tag{3.16}$$

where $\mathbf{e}$ is the vector of 1's and

$$\mathbf{y} = [\underbrace{-1, \ \cdots, \ -1}_{s_1}\underbrace{1, \ \cdots, \ 1}_{s_2}]^T \in \mathbb{R}^{s\times 1}.$$

Then,

$$S_m = BB^T - s\mathbf{c}\mathbf{c}^T \ \text{ and } \ S_w = H_w H_w^T, \tag{3.17}$$

where $H_w = [\mathbf{b}_1 - \mathbf{c}_1, \ldots, \mathbf{b}_{s_1} - \mathbf{c}_1, \mathbf{b}_{s_1+1} - \mathbf{c}_2, \ldots, \mathbf{b}_{s_1+s_2} - \mathbf{c}_2] \in \mathbb{R}^{m\times s}$. Now, we will show $S_m\mathbf{w} = (S_w + S_b)\mathbf{w} = S_b\mathbf{w}$. Applying $B$ to Eqn. 3.16, we obtain

$$\begin{aligned} BB^T\mathbf{w} &= B\mathbf{y} - \beta B\mathbf{e} \\ &= (-s_1\mathbf{c}_1 + s_2\mathbf{c}_2) - \beta s\mathbf{c} \\ &= -s_1\mathbf{c}_1 + s_2\mathbf{c}_2 - (s_2 - s_1)\mathbf{c} + s\mathbf{c}\mathbf{c}^T\mathbf{w}. \end{aligned} \tag{3.18}$$

Therefore,

$$\begin{aligned} S_m\mathbf{w} &= (BB^T - s\mathbf{c}\mathbf{c}^T)\mathbf{w} \\ &= s_1(\mathbf{c} - \mathbf{c}_1) - s_2(\mathbf{c} - \mathbf{c}_2) \\ &= 2s_1s_2(\mathbf{c}_2 - \mathbf{c}_1)/s. \end{aligned} \tag{3.19}$$

Applying $S_b$ to $\mathbf{w}$ and using

$$\mathbf{c}_1^T\mathbf{w} = -1 - \beta \ \text{ and } \ \mathbf{c}_2^T\mathbf{w} = 1 - \beta,$$

we obtain

$$\begin{aligned} S_b\mathbf{w} &= s_1(\mathbf{c}_1 - \mathbf{c})(\mathbf{c}_1 - \mathbf{c})^T\mathbf{w} \\ &\quad + s_2(\mathbf{c}_2 - \mathbf{c})(\mathbf{c}_2 - \mathbf{c})^T\mathbf{w} \\ &= -2s_1s_2(\mathbf{c}_1 - \mathbf{c})/s + 2s_1s_2(\mathbf{c}_2 - \mathbf{c})/s \\ &= 2s_1s_2(\mathbf{c}_2 - \mathbf{c}_1)/s. \end{aligned} \tag{3.20}$$

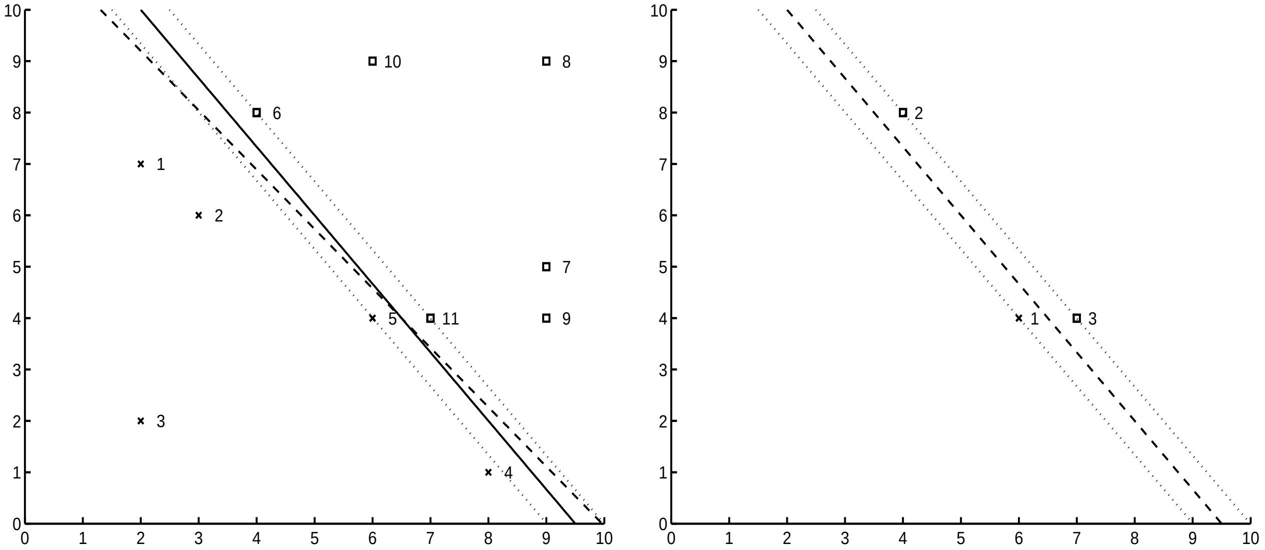

FIG. 3.1. *Classification results of HARD data set (left) and HARD-SVs data set that consists of only the support vectors of HARD data set (right). The dotted line presents the boundaries, i.e.* $\mathbf{w}^T\mathbf{x} + \beta = \pm 1$. *The solid line presents a separation line of the hard-margin SVM and the dashed line presents that of MLDC/GSVD. In the figure shown on the right side, the solid line is (not shown) identical to the dashed line.*

Since $S_m\mathbf{w} = (S_w + S_b)\mathbf{w} = S_b\mathbf{w} = 2s_1s_2(\mathbf{c}_2 - \mathbf{c}_1)/s$ and $\mathbf{c}_1 \neq \mathbf{c}_2$, $S_w\mathbf{w} = 0$ and $S_b\mathbf{w} \neq 0$. The weight vector $\mathbf{w}$ of the hard-margin SVM belongs to null$(S_w) \cap$ null$(S_b)^c$. By Theorem 3.3, the weight vector $\mathbf{w}$ of the hard-margin SVM is equivalent to the dimension reducing transformation vector generated by LDA/GSVD applied to the support vectors. This relationship holds for $\varrho \cdot \mathbf{w}$, where $\varrho \neq 0$ is a scale factor. □

In the following Section, a relationship between the $L_1$-norm soft-margin SVM and LDA/GSVD applied to a subset of the support vectors is presented.

**4. Relationship between the $L_1$-norm Soft-margin SVM and Generalized Linear Discriminant Analysis on a Subset of Support Vectors.** In the primal formulation of the $L_1$-norm soft-margin SVM [28, 29], the margin is maximized and the training error is minimized simultaneously by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{w},\xi_i,b} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i \\ &s.t. \ \ y_i\left[\mathbf{w}^T\mathbf{a}_i + \beta\right] \geq 1 - \xi_i, \ \xi_i \geq 0, \ \ i = 1, \ldots, m, \end{aligned} \tag{4.1}$$

where $\mathbf{a}_i$ represents an input vector, $y_i = \pm 1$ according to whether $\mathbf{a}_i$ is in the positive or negative class, $m$ is the size of the training data, and $C$ is a parameter that controls the trade-off between margin and classification error represented by slack variables $\xi_i$'s.

The corresponding dual quadratic programming problem can be written as

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{a}_i^T \mathbf{a}_j, \\ s.t. \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \;\; 0 \leq \alpha_i \leq C, \;\; i = 1, \ldots, n, \end{aligned} \tag{4.2}$$

where $\alpha_i$ are the solutions of the dual formulation. This formulation shows that the influence of a single training example is limited by $C$. Then, the decision rule is given by $sign(f(\mathbf{x}))$ with

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + \beta, \tag{4.3}$$

where $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{a}$, $\beta$ is chosen so that $y_i f(\mathbf{a}_i) = 1$ for all $i$ with $0 < \alpha_i < C$. According to the Karush-Kuhn-Tucker (KKT) conditions, optimal solutions $\alpha$ and $(\mathbf{w}, \xi, \beta)$ satisfy

$$\begin{aligned} \alpha_i [y_i (\mathbf{w}^T \mathbf{a}_i + \beta) - 1 + \xi_i] &= 0 \quad \text{and} \\ \xi_i (\alpha_i - C) &= 0, \quad i = 1, \ldots, m. \end{aligned} \tag{4.4}$$

These conditions can be rewritten as

$$\begin{cases} \text{Case 1:} \quad y_i f(\mathbf{a}_i) \geq 1, \; if \; \alpha_i = 0 \\ \text{Case 2:} \quad y_i f(\mathbf{a}_i) = 1, \; if \; 0 < \alpha_i < C \\ \text{Case 3:} \quad y_i f(\mathbf{a}_i) \leq 1, \; if \; \alpha_i = C. \end{cases} \tag{4.5}$$

The second and third cases occur when $\xi_i = 0$ and $\xi_i > 0$ respectively. The slack variable can have a non-zero value only when $\alpha_i = C$. If $\alpha_i = 0$ (Case 1), then $\mathbf{a}_i$ is not a support vector. If $0 < \alpha_i < C$ (Case 2), then $\mathbf{a}_i$ is the support vector with $\xi_i = 0$. If $\alpha_i = C$ (Case 3), then $\mathbf{a}_i$ is the support vector with $\xi_i > 0$.

The following theorem shows that the solution $\mathbf{w}$ of soft-margin SVMs is equivalent to the dimension reducing transformation vector generated by LDA based on the generalized singular value decomposition.

THEOREM 4.1 (Relationship between $L_1$-norm Soft-margin SVM and Generalized Linear Discriminant Analysis on a Subset of Support Vectors). *Consider a subset of support vectors* $\mathbf{b}_i$, *for* $i = 1, \ldots, s$, *for which* $0 < \alpha_i < C$ *in the* $L_1$*-norm soft-margin SVM, Eqn. (4.2), for a binary* *classification problem. The weight vector* $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{b}_i$ *belongs to null(*$S_w$*)* $\cap$ *null(*$S_b$*)*$^c$ *where* $S_w$ *and* $S_b$ *are scatter matrices defined for the subset of the support vectors.*

*Proof.* Since all of the support vectors $\mathbf{b}_i$ with $0 < \alpha_i < C$ lie on the boundaries, they satisfy $\mathbf{w}^T\mathbf{b}_i + \beta = \pm 1$ and with $\alpha_i \neq 0$ we have $\mathbf{w} = \sum_{i=1}^{s} \alpha_i y_i \mathbf{b}_i$. Also $\{\alpha_i\}$, $1 \leq i \leq s$ and $0 < \alpha_i < C$, are the solution to the QP problem Eqn. (4.2). Therefore, the arguments used in the proof of Theorem 3.3 hold for this subset of support vectors. □

Nonlinear extension of LDA/GSVD using kernel functions was also developed in [21]. The kernelized discriminant analysis based on the generalized singular value decomposition (KDA/GSVD) can reduce the data dimension significantly allowing extraction of nonlinear features regardless of the nonsingularity of the scatter matrix in either the input space or feature space. Consider a nonlinear feature mapping $\phi$ that maps the input data to a feature space where the mapped data may be linearly separable. The superscript $\phi$ denotes the corresponding vector or matrix computed in the feature space. The between-class scatter matrix and within-class scatter matrix in the feature space are defined as

$$S_b^\phi = H_b^\phi (H_b^\phi)^T \quad \text{and} \quad S_w^\phi = H_w^\phi (H_w^\phi)^T,$$

respectively, where

$$H_w^\phi = [\phi(\mathbf{b}_1) - \mathbf{c}_1^\phi, \ldots, \phi(\mathbf{b}_{s_1}) - \mathbf{c}_1^\phi, \phi(\mathbf{b}_{s_1+1}) - \mathbf{c}_2^\phi, \ldots, \phi(\mathbf{b}_s) - \mathbf{c}_2^\phi]$$

and

$$H_b^\phi = [\sqrt{s_1}(\mathbf{c}_1^\phi - \mathbf{c}^\phi), \sqrt{s_2}(\mathbf{c}_2^\phi - \mathbf{c}^\phi)],$$

$\mathbf{c}_1^\phi$ and $\mathbf{c}_2^\phi$ are the centroid vectors for each class and $\mathbf{c}^\phi$ is a global centroid vector for the subset of support vectors in the feature space.

Let $\varphi$ represent a vector in $range(\phi(A)) = range((\phi(\mathbf{a}_1) \cdots \phi(\mathbf{a}_n))$, i.e., $\varphi = \sum_i^n \alpha_i \phi(\mathbf{a}_i)$ for some $\alpha_i$, $i = 1, \cdots, n$. Then,

$$(H_b^\phi)^T \varphi = (K_b^\phi)^T \tilde{\alpha} \ \text{ and } \ (H_w^\phi)^T \varphi = (K_w^\phi)^T \tilde{\alpha},$$

where $\tilde{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$,

$$(K_b^\phi)_{ij} = \frac{\sqrt{n_j}}{n} \sum_{s \in N_j} \mathbf{k}(\mathbf{a}_s^T, \mathbf{a}_i) - \frac{\sqrt{n_j}}{n} \sum_{q=1}^{n} \mathbf{k}(\mathbf{a}_q^T, \mathbf{a}_i), \ \text{ for} 1 \leq i \leq n, \ \ 1 \leq j \leq p$$

$$(K_w^\phi)_{ij} = \mathbf{k}(\mathbf{a}_i^T, \mathbf{a}_j) - \frac{1}{n_f} \sum_{t \in N_f} \mathbf{k}(\mathbf{a}_t^T, \mathbf{a}_i), \ \text{ for} 1 \leq i \leq n, \ \ 1 \leq j \leq n,$$

The $n_j$ and $n_f$ are the number of data points in the class $j$ and $f$ respectively. By computing the GSVD of the matrix pair $(K_b^\phi, K_w^\phi)$, we can obtain the generalized singular vectors $W^\phi \in \mathbb{R}^{m\times(p-1)}$. Then the reduced representation of the a data point $\mathbf{x}$ can be computed by

$$\mathbf{x}^r = (W^\phi)^T \begin{bmatrix} \mathbf{k}(\mathbf{a}_1^T, \mathbf{x}) \\ \vdots \\ \mathbf{k}(\mathbf{a}_n^T, \mathbf{x}) \end{bmatrix}.$$

The reduced representation of all the training data points can be computed by

$$Y = (W^\phi)^T \mathbf{k}(A^T, A),$$

where $\mathbf{k}(A^T, A)$ is a $n \times n$ kernel matrix. In particular, if $\mathbf{x}$ and $\mathbf{y}$ are column vectors in $\mathbb{R}^m$, then $\mathbf{k}(\mathbf{x}^T, A) = \mathbf{k}(A^T, \mathbf{x})$ is a row vector in $\mathbb{R}^n$, and $\mathbf{k}A^T, A$ is an $n \times n$ matrix.

Now, we discuss the relationship between the nonlinear $L_1$-norm soft-margin SVM and kernelized discriminant analysis based on the generalized singular value decomposition (KDA/GSVD) [21]. Consider a nonlinear feature mapping $\phi(\mathbf{x})$ that maps the input data to a feature space where the mapped data may have a linearly separable structure. Without knowing the feature mapping $\phi(\mathbf{x})$ or the feature space explicitly, we can work on the feature space through a kernel function, as long as the problem formulation depends only on the inner products between data points in the feature space and not on the data points themselves. The dual formulation with a kernel function, i.e.

$$\mathbf{k}(\mathbf{a}_i^T, \mathbf{a}_j) = \phi(\mathbf{a}_i)^T \phi(\mathbf{a}_j), \tag{4.6}$$

is

$$\begin{aligned} \max_{\alpha} \ & \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j \mathbf{k}(\mathbf{a}_i^T, \mathbf{a}_j) \\ s.t. \ & \sum_{i=1}^n \alpha_i y_i = 0, \ 0 \le \alpha_i \le C, \ i = 1, \dots, n. \end{aligned} \tag{4.7}$$

Then the decision rule is given by $sign(f(\mathbf{x}))$ with

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i \mathbf{k}(\mathbf{a}_i^T, \mathbf{x}) + \beta, \tag{4.8}$$

where $\beta$ is chosen so that $y_i f(\mathbf{a}_i) = 1$ for all $i$ with $0 < \alpha_i < C$. The Karush-Kuhn-Tucker (KKT) conditions in Eqn. (4.5) can be applied. Then, the relationship between the nonlinear

$L_1$-norm soft-margin SVM and kernelized discriminant analysis on a subset of support vectors follows.

THEOREM 4.2 (Relationship between the Nonlinear $L_1$-norm Soft-margin SVM and Kernelized Discriminant Analysis on Subset of Support Vectors). *Consider the subset of support vectors* $\mathbf{b}_i$, *for* $i = 1, \ldots, s$, *for which* $0 < \alpha_i < C$ *in the nonlinear* $L_1$*-norm soft-margin SVM, Eqn. (4.7), for a binary classification problem. The weight vector* $\tilde{\mathbf{w}} = \sum_i \alpha_i y_i \phi(\mathbf{b}_i)$ *of the nonlinear* $L_1$*-norm soft-margin SVM, where* $\phi(\cdot)$ *is a nonlinear feature mapping, belongs to null(*$S_w^\phi$*)* $\cap$ *null(*$S_b^\phi$*)*$^c$ *where* $S_w^\phi$ *and* $S_b^\phi$ *are the scatter matrices for the subset of support vectors in the feature space.*

*Proof.* Consider the subset of support vectors for which $0 < \alpha_i < C$. Note that all $s$ data points $\mathbf{a} \in \mathbb{R}^n$ lie on the boundary, i.e. $\tilde{\mathbf{w}}^T\phi(\mathbf{x}) + \beta = \pm 1$ where $\tilde{\mathbf{w}} = \sum_{i=1}^s \alpha_i y_i \phi(\mathbf{b}_i)$ and a nonlinear feature mapping $\phi(\cdot)$ that maps the input data to a feature space. The $\{\alpha_i\}$, $i = 1, \ldots, s$ and $0 < \alpha_i < C$, are the solution to the QP problem Eqn. (4.7). The same procedure as that in the proof of Theorem 4.1 is applied considering only the subset of support vectors that satisfy $0 < \alpha_i < C$ in the feature space. Since $S_m^\phi \tilde{\mathbf{w}} = (S_w^\phi + S_b^\phi)\tilde{\mathbf{w}} = S_b^\phi \tilde{\mathbf{w}} = 2s_1 s_2(\mathbf{c}_2^\phi - \mathbf{c}_1^\phi)/s$ and $\mathbf{c}_1^\phi \neq \mathbf{c}_2^\phi$, $S_w^\phi \tilde{\mathbf{w}} = 0$ and $S_b^\phi \tilde{\mathbf{w}} \neq 0$. The weight vector $\tilde{\mathbf{w}}$ of the nonlinear $L_1$-norm soft-margin SVM belongs to null($S_w^\phi$) $\cap$ null($S_b^\phi$)$^c$. $\square$

**5. Results and Discussion.** To illustrate the relationships shown in this paper, a small artificial classification problem is used. The perfectly separable data set (HARD) consists of eleven two-dimensional data points

$$A = \begin{pmatrix} 2 & 3 & 2 & 8 & 6 & 4 & 9 & 9 & 9 & 6 & 7 \\ 7 & 6 & 2 & 1 & 4 & 8 & 5 & 9 & 4 & 9 & 4 \end{pmatrix} \in \mathbb{R}^{2\times 11}$$

for which the class index vector $\mathbf{y}$ is

$$\mathbf{y} = \begin{pmatrix} -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}^T \in \mathbb{R}^{11\times 1}.$$

With the hard-margin SVM, it can be found that the support vectors are the 5th, 6th, and 11th data points. The HARD-SVs data set consists of the support vectors and the corresponding class index. A SOFT data set is also prepared by adding a data point $(4, 4)$ to the positive class of the HARD data set, which now becomes a non-linearly separable classification problem. Using the $L_1$-norm soft-margin SVM, the support vectors are found to be the 1st, 2nd, 4th, 6th, and 12th data points when $C = 10.0$. Among those support vectors, the 1st, 2nd, 4th and 6th data points lie on the boundary, i.e. $\mathbf{w}^T\mathbf{x} + \beta = \pm 1$ where $\mathbf{w} = \sum_{i=1}^s \alpha_i y_i \mathbf{a}_i$, $0 < \alpha_i < C$. These data points

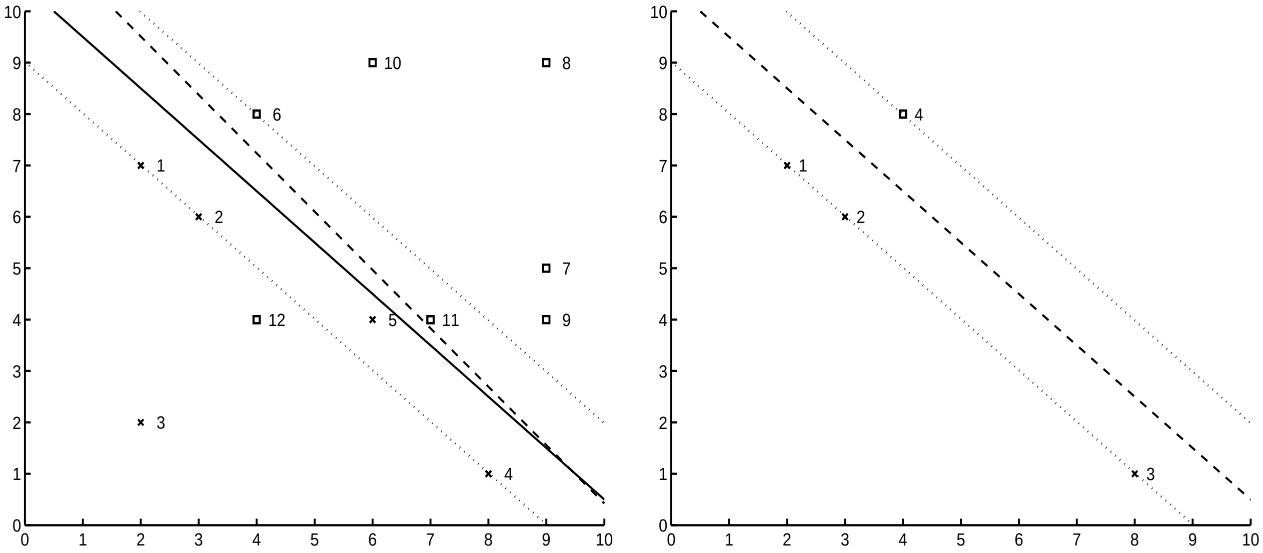

FIG. 4.1. *Classification results of SOFT data set (left) and SOFT-SVs data set that contains the subset of support vectors that satisfy* $0 < \alpha < C$ *(right). The dotted line presents the boundaries, i.e.* $\mathbf{w}^T\mathbf{x} + \beta = \pm 1$*. The regularization parameter* $C$ *is set to 10.0. The solid line presents a separation line of the* $L_1$*-norm soft-margin SVM and the dashed line presents that of MLDC/GSVD. In the figure shown on the right side, the solid line is (not shown) identical to the dashed line.*

constitute SOFT-SVs data set that contains the subset of support vectors that satisfy $0 < \alpha < C$. For the 12th data point, $\alpha_{12} = C$.

The above results show the relationships of the weight vector from SVMs and the LDA/GSVD solution. To further visualize the separating hyperplane obtained by the LDA/GSVD, the bias term $\beta$ also needs to be determined. For determining $\beta$, we used the marginal linear discriminant classifier based on the GSVD, called MLDC/GSVD [14]. In MLDC/GSVD, a negative class is defined as a class that has smaller mean value in the projected space between two classes, i.e. $\mathbf{c}_-^r < \mathbf{c}_+^r$. The centroid vectors of the training set in the reduced dimensional space are computed by

$$\mathbf{c}_-^r = \frac{1}{n_-}\sum_{i \in N_-} \mathbf{w}^T\mathbf{a}_i, \;\; \mathbf{c}_+^r = \frac{1}{n_+}\sum_{i \in N_+} \mathbf{w}^T\mathbf{a}_i,$$

where $N_-$ and $N_+$ are the sets of the indices of the data set $A$ that belong to the negative class and positive class, and $n_-$ and $n_+$ the number of data items in the negative class and the positive class, respectively. Then, the parameter $b$ is determined from

$$\beta = -\frac{1}{2}(\max_{i \in N_-}(\mathbf{w}^T \mathbf{a}_i) + \min_{i \in N_+}(\mathbf{w}^T \mathbf{a}_i)). \tag{5.1}$$

In summary, a classifier to discriminate two classes is obtained by $sign(\mathbf{w}^T \mathbf{x} + \beta)$, where $\mathbf{w}$ comes from LDA/GSVD and $\beta$ from Eqn. (5.1). When the projected points of two classes are overlapped such that

$$\max_{i \in N_-}(\mathbf{w}^T \mathbf{a}_i) > \min_{i \in N_+}(\mathbf{w}^T \mathbf{a}_i),$$

$\beta = -\mathbf{w}^T \mathbf{c}$ is used.

Fig. 3.1 shows the relationship between the hard-margin SVM and the LDA/GSVD applied on the support vectors. Fig. 4.1 visualizes a relationship between $L_1$-norm soft-margin SVM and MLDC/GSVD on the subset of support vectors that satisfy $0 < \alpha < C$.

For nonlinear classification by a kernel function, a marginal kernelized discriminant classifier based on the generalized singular value decomposition (MKDC/GSVD) is used [12]. A threshold value of the kernelized discriminant is

$$b^* = -\frac{1}{2}(\max_{i \in N_-}(\mathbf{k}(\mathbf{a}_i^T, A)\mathbf{z}^*) + \min_{i \in N_+}(\mathbf{k}(\mathbf{a}_i^T, A)\mathbf{z}^*)),$$

where $\mathbf{k}(\mathbf{a}_u^T, A)$ is a $1 \times m$ kernel vector, $\mathbf{z}^*$ is the nonlinear dimension reducing transformation vector obtained by the kernel discriminant analysis based on the generalized singular value decomposition (KDA/GSVD) [21]. The negative class has smaller mean value in the projected space between the two classes. Then, the class of a new test point $\mathbf{x}$ can be assigned by

$$f^*(\mathbf{x}) = \mathbf{k}(\mathbf{x}^T, A)\mathbf{z}^* + \beta^*. \tag{5.2}$$

Notice that there is no condition $\sum_i z_i = 0$ in the decision function of MKDA/GSVD. Even though the results presented in Theorem 4.2 hold, the decision boundary ($f^*(\mathbf{x}) = 0$) may not match with the decision boundary ($f(\mathbf{x}) = 0$) of the nonlinear $L_1$-norm soft-margin SVM.

Fig. 4.2 visualizes the relationship between the nonlinear $L_1$-norm soft-margin SVM and MKDC/GSVD on the support vectors that satisfy $\epsilon < \alpha < C - \epsilon$. The tolerance for support vectors $\epsilon$ was set to $C * 10^{-6}$. The radial basis function (RBF) kernel $\mathbf{k}(\mathbf{a}_i^T, \mathbf{a}_j) = \exp(-\gamma \|\mathbf{a}_i - \mathbf{a}_j\|^2)$ was used. The solid contour that presents a decision boundary of the nonlinear $L_1$-norm soft-margin SVM approximately matches the dashed contour from MKDC/GSVD. Since the separating hyperplane $f(\mathbf{x}) = \sum_{i=1}^m \alpha_i y_i \mathbf{k}(\mathbf{a}_i^T, \mathbf{x}) + \beta$ is affected by the data points with the condition $\alpha_i = C$ as well as those with $0 < \alpha_i < C$, the same separating hyperplanes may not be produced. In Fig. 4.2 and Fig. 4.3, different radial basis function (RBF) kernel parameters and the regularization parameters were tested. It can be recognized that the decision boundary ($f^*(\mathbf{x}) = 0$) may not match with the decision boundary ($f(\mathbf{x}) = 0$) of the nonlinear $L_1$-norm soft-margin SVM in the bottom right graph of Fig. 4.3. However, the decision boundaries ($f^*(\mathbf{x}) = 0$) that were shown in Figs. 4.2 and 4.3 are very similar to ($f(\mathbf{x}) = 0$) or reasonably close when the subset of support vectors are used.

**6. Conclusions and Discussion.** We have shown the mathematical relationship that the parameter $w$, which determines the separating hyperplane in the hard margin support vector classifier, is the same as the dimension reducing transformation $w$ obtained when the generalized linear discriminant analysis, based on the generalized singular value decomposition (LDA/GSVD), is applied to the support vectors. In addition, we have also shown that the parameter $w$ from the soft margin $L_1$ norm support vector classifier is the same as the dimension reducing transformation $w$ obtained when the LDA/GSVD is applied to a certain subset of the support vectors. These results can also be generalized when a kernel function is introduced into the SVM and the LDA/GSVD formulations, which allows the methods to be applied to nonlinear problems. By extending the LDA/GSVD to compute the parameter $\beta$ that provides the separating hyperplane $\mathbf{w}^x + \beta$ with $w$, the results provide an interesting relationship between the Support Vector Classification methods and Linear Discriminant Analysis, which is a dimension reduction method.

If we can predict the support vectors in the hard margin SVM or the subset of the of support vectors with $0 < \alpha < C$, the weight vector $\mathbf{w}$ of the support vector machines can be estimated by LDA/GSVD. The *pseudo support vectors* obtained by boundary data point hunting algorithms presented in [13] may produce a dimension reducing transformation vector which is similar to the weight vector of the $L_1$-norm soft-margin SVM. It is possible to interpret a classification problem as a problem to search a subset of data points that can yield an accurate decision boundary, i.e. a *data reduction problem*. This study may provide new insights regarding the generalized LDA and provide a new direction in designing a classifier and data reduction method based on the generalized LDA.

**Acknowledgments.** The work of Haesun Park has been performed while at the NSF and was partly supported by IR/D from the National Science Foundation (NSF).

## REFERENCES

[1] Z. Bai and H. Zha, *A new preprocessing algorithm for the computation of the generalized singular value decomposition*, SIAM J. Sci. Comp., 14 (1993), pp. 1007–1012.

[2] C. J. C. Burges, *A tutorial on support vector machines for pattern recognition*, Data Mining and Knowledge Discovery, 2 (1998), pp. 121–167.

[3] C. J. C. Burges and B. Schölkopf, *Improving the accuracy and speed of support vector learning machines*, in Advances in Neural Information Processing Systems, M. Mozer, M. Jordan, and T. Petsche, eds., vol. 9, Cambridge, MA, 1997, MIT Press, pp. 375–381.

[4] B. de Moor and P. van Dooren, *Generalizing the singular value and QR decompositions*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 993–1014.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-interscience, New York, 2001.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition, second edition*, Academic Press, Boston, 1990.

[7] G. H. Golub and C. F. van Loan, *Matrix Computations, third edition*, Johns Hopkins University Press, Baltimore, 1996.

[8] P. C. Hansen, *Regularization, GSVD and truncated GSVD*, BIT, 29 (1989), pp. 491–504.

[9] P. Howland, M. Jeon, and H. Park, *Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 165–179.

[10] P. Howland and H. Park, *Generalizing discriminant analysis using the generalized singular value decomposition*, IEEE Trans. Pattern Anal. Machine Intell., 26 (2004), pp. 995–1006.

[11] B. Kågström, *The generalized singular value decomposition and the general $A - \lambda B$ problem*, BIT, 24 (1985), pp. 568–583.

[12] H. Kim, B. Drake, and H. Park, *Multiclass classifiers based on dimension reduction with generalized lda algorithms*, 2004. submitted for publication.

[13] H. Kim and H. Park, *Data reduction in support vector machines by a kernelized ionic interaction model*, in Proceedings of the 4th SIAM International Conference on Data Mining (SDM04), M. W. Berry, U. Dayal, C. Kamath, and D. Skillicorn, eds., 2004, pp. 507–511.

[14] ——, *Gene selection by LDA based on generalized singular value decomposition*, in Proceedings of SIAM Bioinformatics Workshop at the 4th SIAM International Conference

on Data Mining (SDM04), Z. Obradovic and J. Komorowski, eds., 2004, pp. 36–41.

[15] C. V. LOAN, *A unitary method for esprit direction-or-arrival estimation algorithm*, in Proceedings of SPIE Advanced Algorithms and Architectures for Signal Processing II, SPIE, 1987, pp. 170–176.

[16] E. OSUNA, R. FREUND, AND F. GIROSI, *Support vector machines: Training and applications*, Tech. Report AI Memo 1602, MIT A.I. Lab, 1997.

[17] C. C. PAIGE, *The general linear model and generalized singular value decomposition*, Lin. Alg. Appl., 70 (1985), pp. 269–284.

[18] ——, *Computing the generalized singular value decomposition*, SIAM J. Sci. and Stat. Comp., 7 (1986), pp. 1126–1146.

[19] C. C. PAIGE AND M. A. SAUNDERS, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.

[20] C. H. PARK AND H. PARK, *A comparison of generalized LDA algorithms for undersampled problems*, Tech. Report 03-048, Department of Computer Science and Engineering, University of Minnesota, 2003.

[21] ——, *Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition*, SIAM Journal on Matrix Analysis and Applications, (2005), pp. 98–102.

[22] H. PARK, *Esprit direction-of-arrival estimation in the presence of spatially correlated noise*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 185–193.

[23] W. SHOUGEN AND Z. SHUQUIN, *An algorithm for $Ax = \lambda Bx$ with symmetric positive definite $A$ and $B$*, SIAM J. Matrix Anal. Appl., 2 (1991).

[24] J. M. SPEISER AND C. F. VAN LOAN, *Signal processing computations using the generalized singular value decomposition*, in Proc. SPIE Vol.495, Real Time Signal Processing VII, 1984, pp. 47–55.

[25] S. VAN HUFFEL AND J. VANDEWALLE, *Analysis and properties of the generalized total least squares problem $AX \approx B$ when some or all columns in $A$ are subject to error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 294–315.

[26] C. F. VAN LOAN, *Generalizing the singular value decomposition*, SIAM J. Num. Anal., 13 (1976), pp. 76–83.

[27] ——, *Computing the CS and generalized singular value decomposition*, Numer. Math., 46 (1985), pp. 479–492.

[28] V. VAPNIK, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[29] ——, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

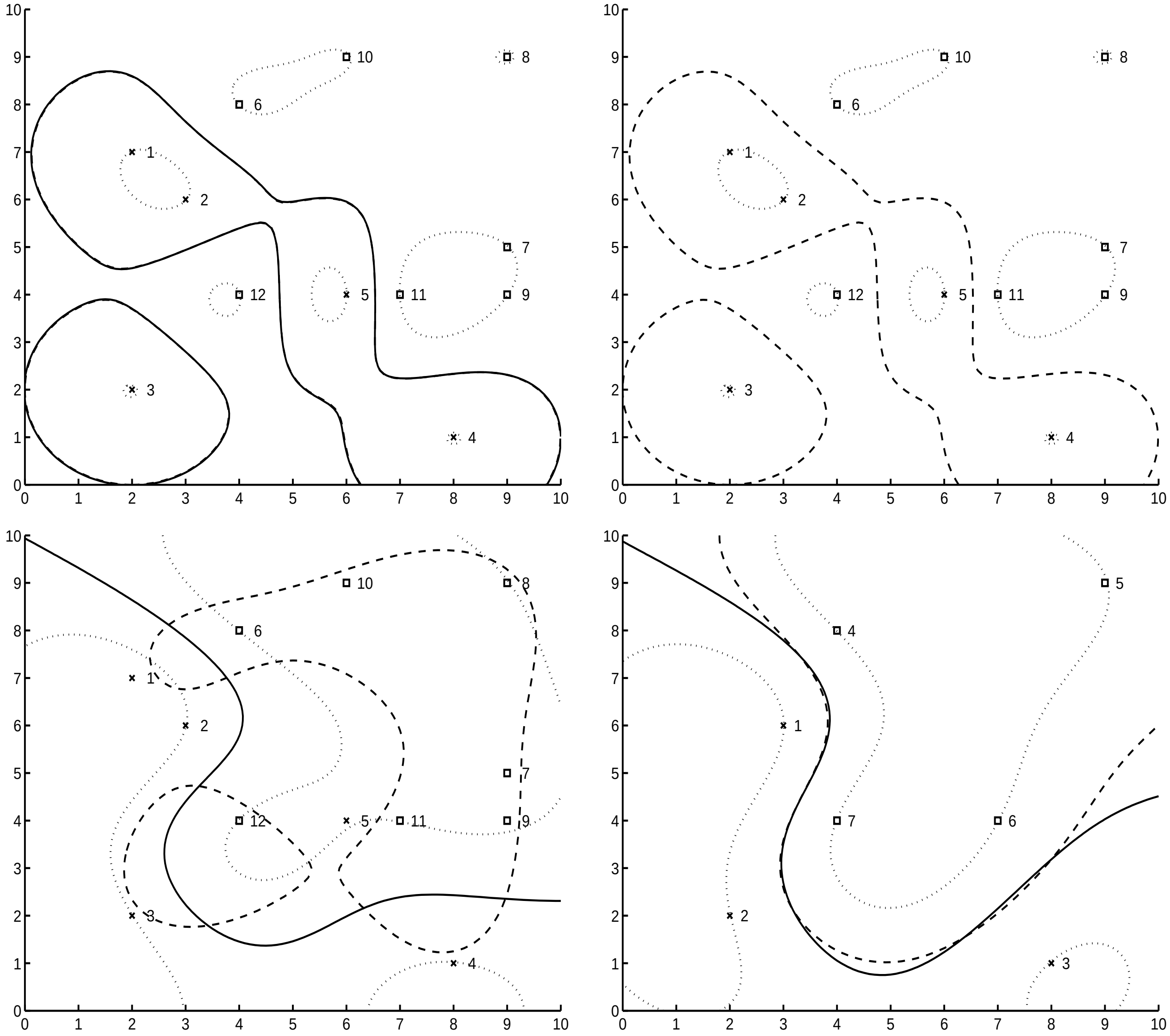

FIG. 4.2. *Classification results of SOFT data set (left side) and SOFT-SVs data set that contains the subset of support vectors that satisfy* $0 < \alpha < C$ *(right side). The solid contour presents a decision boundary of the nonlinear* $L_1$*-norm soft-margin SVM and the dashed contour presents that of MKDC/GSVD. The dotted line presents the boundaries, i.e.* $\sum_{i=1}^{n} \alpha_i y_i \mathbf{k}(\mathbf{a}_i^T, \mathbf{x}) + \beta = \pm 1$. *The radial basis function (RBF) kernel parameter and the regularization parameter are* $\gamma = 0.5$ *and* $C = 10.0$ *for upper figures and* $\gamma = 0.1$ *and* $C = 10.0$ *for lower figures, respectively.*

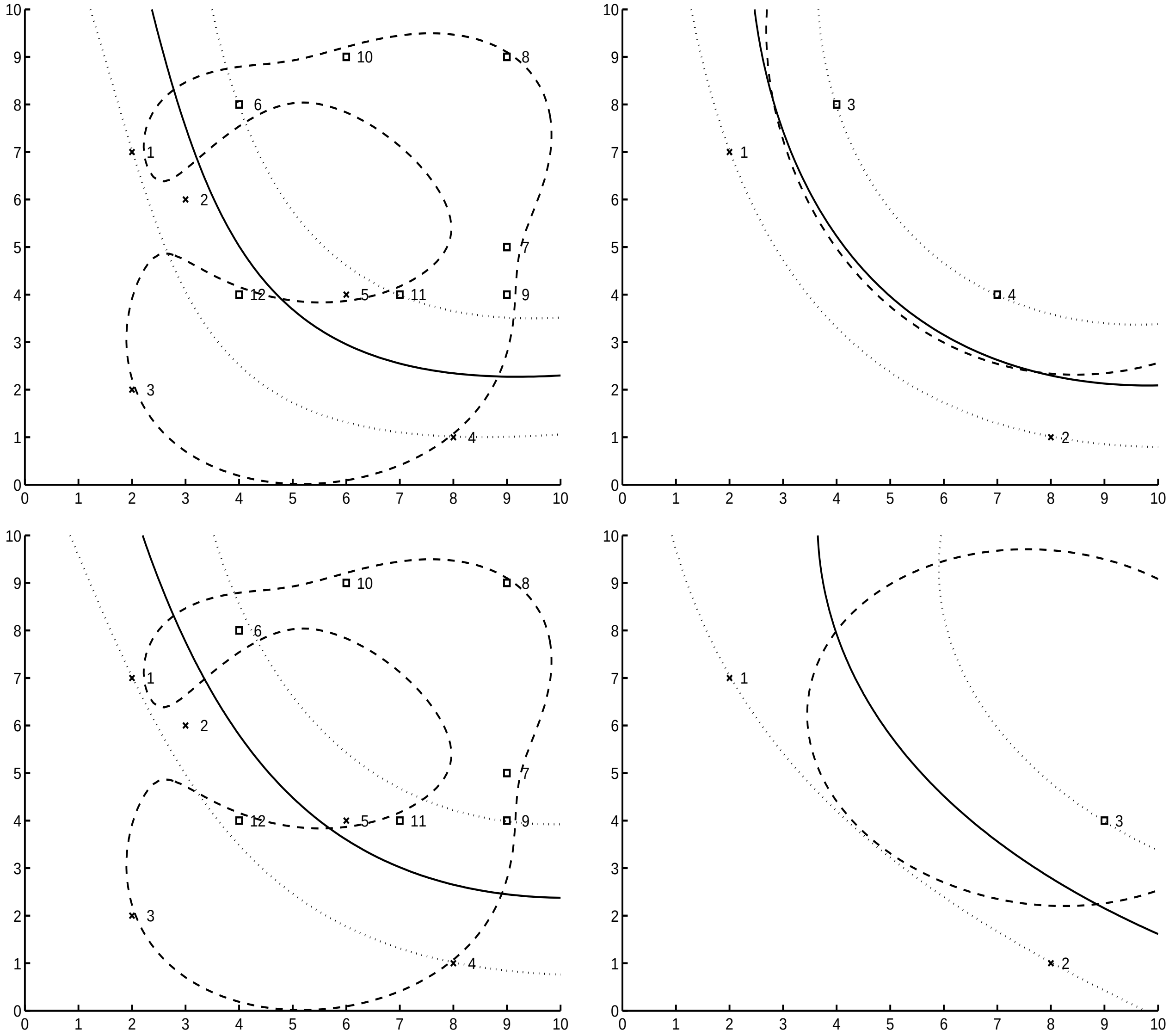

FIG. 4.3. *Classification results of SOFT data set (left side) and SOFT-SVs data set that contains the subset of support vectors that satisfy* $0 < \alpha < C$ *(right side). The solid contour presents a decision boundary of the nonlinear* $L_1$*-norm soft-margin SVM and the dashed contour presents that of MKDC/GSVD. The dotted line presents the boundaries, i.e.* $\sum_{i=1}^{n} \alpha_i y_i \mathbf{k}(\mathbf{a}_i^T, \mathbf{x}) + \beta = \pm 1$*. The radial basis function (RBF) kernel parameter and the regularization parameter are* $\gamma = 0.01$ *and* $C = 50.0$ *for upper figures and* $\gamma = 0.01$ *and* $C = 20.0$ *for lower figures, respectively.*