Biometrika (2012), **99**, 4, *pp.* 851–864
© 2012 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/ass029
Advance Access publication 24 July 2012

Bidirectional discrimination with application to data visualization

By HANWEN HUANG

Center for Clinical and Translational Sciences, University of Texas Health Science Center at Houston, Houston, Texas 77030, U.S.A.

hanwen.huang@uth.tmc.edu

YUFENG LIU and J. S. MARRON

Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

yfliu@email.unc.edu marron@email.unc.edu

Summary

Linear classifiers are very popular, but can have limitations when classes have distinct subpopulations. General nonlinear kernel classifiers are very flexible, but do not give clear interpretations and may not be efficient in high dimensions. We propose the bidirectional discrimination classification method, which generalizes linear classifiers to two or more hyperplanes. This new family of classification methods gives much of the flexibility of a general nonlinear classifier while maintaining the interpretability, and much of the parsimony, of linear classifiers. They provide a new visualization tool for high-dimensional, low-sample-size data. Although the idea is generally applicable, we focus on the generalization of the support vector machine and distance-weighted discrimination methods. The performance and usefulness of the proposed method are assessed using asymptotics and demonstrated through analysis of simulated and real data. Our method leads to better classification performance in high-dimensional situations where subclusters are present in the data.

Some key words: Asymptotics; Classification; High-dimensional data; Initial value; Iteration; Optimization; Visualization.

1. Introduction

In statistical machine learning, the objective of linear classification is to make a decision based on the value of a linear combination of the characteristics. Examples of linear classification algorithms include Fisher's (1936) linear discriminant, the support vector machine (Vapnik, 1995; Cristianini & Taylor, 2000; Hastie et al., 2001) and distance-weighted discrimination (Marron et al., 2007).

Although linear classifiers are very widely used, they can be improved upon when each class contains diverse subpopulations. For example, in microarray analysis, within each class of interest, e.g., disease versus control, auxiliary differences such as male versus female can lead to diverse subpopulations. A toy example illustrating this given in Fig. 1 includes two classes, each divided into two subclusters. Linear methods for classification cannot capture the class differences effectively in this case, which motivates us to find a more general hypersurface that can do so.

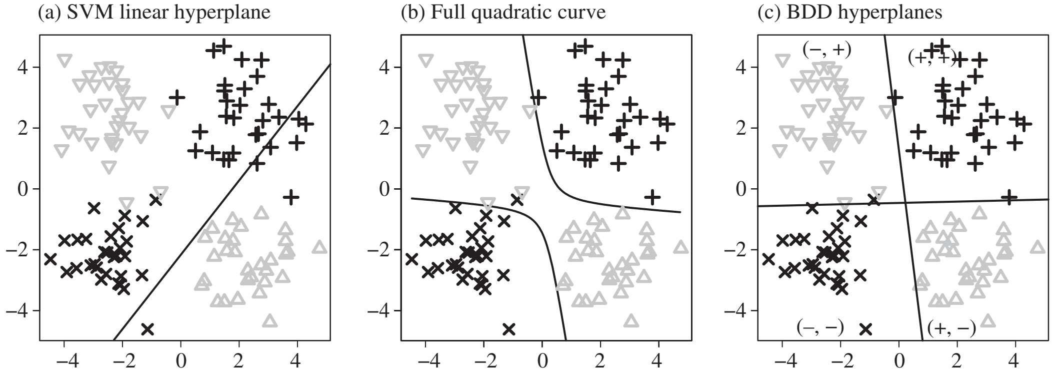

Fig. 1. The scatter-plot of a toy data example in two dimensions. Black indicates the positive class and grey indicates the negative class. Different symbols in the same class represent different subclusters. The linear support vector machine decision boundary is denoted by the solid line in (a). The decision boundaries from the two nonlinear methods are denoted by the curves in (b) and (c). SVM, support vector machine; BDD, bidirectional discrimination.

The extension of support vector machine and distance-weighted discrimination methods from the linear case to the nonlinear case is quite straightforward using the so-called kernel trick (Aizerman et al., 1964; Boser et al., 1992); for an overview, see Hastie et al. (2001). The solid curves shown in Fig. 1(b) are the nonlinear decision boundary implemented using the full quadratic kernel support vector machine method (Vapnik, 1995; Burges, 1998). Its performance is clearly much better than that of the linear classifier.

Although nonlinear classification methods can have low error rates, they do not easily provide intuitive interpretations of class differences relative to the linear ones. To achieve a new balance between the relative strengths of linear and nonlinear methods, we develop a new classification method, bidirectional discrimination, which lies between linear and full nonlinear kernel methods. This employs two or more linear hyperplanes to separate the two classes. The bidirectional discrimination decision boundary, shown in Fig. 1(c), does an intuitively appealing job of separating the two classes, since it not only provides good between-class separation but also divides each class into two subclusters. Relative to linear methods, it has the flexibility to tackle problems with a complex structure. In contrast to general nonlinear methods, it has a simpler functional form and thus retains most of the parsimony of the linear methods.

Linear methods have good interpretability in the sense that they seek to find a direction that can explain the biggest difference between the two classes. The loadings, i.e., the entries of this direction vector, explain the importance of each variable in the model. The score of each data point can be evaluated by projecting it onto this direction. Bidirectional discrimination inherits these appealing interpretation properties.

Another important feature of bidirectional discrimination is that its two hyperplanes automatically provide a visualization tool for high-dimensional low-sample-size data. In particular, visualization of the scores is usually very informative.

Many statistical methods suffer from overfitting, especially in high-dimensional situations. Regularization schemes have been introduced to mitigate against severe overfitting, but as the dimension grows, with the signal in the data fixed, any method will eventually break down. It is interesting to compare the methods on the basis of when this breakdown occurs. If the dimension is d, the number of parameters included in the bidirectional discrimination method will be $2d$, far fewer than the number used in the quadratic kernel method, $d(d+1)/2$. We will see in § 3 that this dependence on fewer parameters, together with a more effective use of the data, gives bidirectional discrimination superior breakdown properties, relative to the full quadratic and Gaussian kernel methods.

Another important use of bidirectional discrimination is for datasets that include subpopulations, where subclusters within each class can be discovered, as shown in § 3·2.

In this article, we focus on the bidirectional method. We have also generalized bidirectional discrimination to multiple directions as discussed in the Supplementary Material. Here explicit focus is on the support vector machine and distance-weighted discrimination methods. Hence, these are used as examples to illustrate how the bidirectional discrimination method works. The fundamental concept is more general and can also be applied to other linear classifiers.

2. Bidirectional discrimination framework

2·1. *Review*

Suppose that the training dataset consists of n d-vectors $x_i = (x_{i1}, \ldots, x_{id})$ with corresponding class indicators $y_i \in \{+1, -1\}$ $(i = 1, \ldots, n)$, which are distributed according to some unknown probability distribution function $\mathrm{pr}(x, y)$. The main idea behind the classical one-directional classification problem is to find the separating hyperplane with maximum separation between the two classes, assuming that there is a plane that perfectly separates them. One important goal is prediction: if we choose $w \in \mathcal{R}^d$ as the normal vector for our hyperplane and $\beta \in \mathcal{R}$ to determine its position, the sign of $f = x^{\mathrm{T}} w + \beta$ can be used for the prediction of class labels for new inputs x.

Both the one-directional support vector machine and the distance-weighted discrimination approaches can be represented in terms of optimization problems. They depend on the signed distance from each data point to the decision boundary, which is defined as

$$r_i = y_i(x_i^{\mathrm{T}} w + \beta) + \xi_i, \tag{1}$$

where the slack variable $\xi_i \geqslant 0$ is added to make sure that all residuals are positive (Cortes & Vapnik, 1995). The one-directional support vector machine classifier solves the regularization problem

$$\min_{\{w,\beta\}} \left(\frac{1}{2} \|w\|^2 + C_{1\mathrm{SVM}} \sum_{i=1}^{n} \xi_i \right),$$

subject to $y_i(x_i^{\mathrm{T}} w + \beta) + \xi_i \geqslant 1$ and $\xi_i \geqslant 0$, where the penalty parameter $C_{1\mathrm{SVM}} > 0$ balances the separation and the amount of violation of the constraints, and $\|w\|$ denotes the Euclidean norm of w.

The optimization task of the one-directional distance-weighted discrimination is to solve

$$\min_{\{w,\beta\}} \sum_i \left(\frac{1}{r_i} + C_{1\mathrm{DWD}} \xi_i \right), \tag{2}$$

subject to $r_i = y_i(x_i^{\mathrm{T}} w + \beta) + \xi_i \geqslant 0$, $\|w\|^2 = 1$ and $\xi_i \geqslant 0$, where $C_{1\mathrm{DWD}} > 0$ is the penalty parameter. The optimization formula (2) can be reparameterized as a second-order cone programming problem. There exist many well-established algorithms for solving such problems (Alizadeh et al., 2001). For a more detailed description of the one-directional distance-weighted discrimination, see Marron et al. (2007).

2·2. *Bidirectional discrimination*

In the bidirectional case, we have two hyperplanes represented by parameters (w_1, β_1) and (w_2, β_2). Let $f_1 = x^{\mathrm{T}} w_1 + \beta_1$ and $f_2 = x^{\mathrm{T}} w_2 + \beta_2$ be classification functions representing each of the two separating hyperplanes, defined as $f_1 = 0$ and $f_2 = 0$. As shown in Fig. 1(c), we denote by $(+, +)$ the region that satisfies $f_1 > 0$ and $f_2 > 0$ and denote the other three regions likewise. Data from the positive class tend to be located on the upper-right and lower-left regions, with labels $(+, +)$ and $(-, -)$ while those from the negative class tend to lie in the upper-left and lower-right regions, with labels $(-, +)$ and $(+, -)$. Thus, $\mathrm{sign}(f_1 f_2)$ is used as the prediction rule in the bidirectional setting. A natural way of generalizing linear classifiers is to replace the signed distance r_i of the ith data point (1) with

$$s_i = y_i f_1 f_2 + \xi_i. \tag{3}$$

Once s_i are given, the optimization problem solved by the bidirectional support vector machine can be stated as

$$\min_{w,\beta,\xi} \frac{1}{2}(\|w_1\|^2 + \|w_2\|^2) + C_{\mathrm{SVM}} \sum_{i=1}^{n} \xi_i,$$

subject to

$$s_i = y_i(x_i^{\mathrm{T}} w_1 + \beta_1)(x_i^{\mathrm{T}} w_2 + \beta_2) + \xi_i \geqslant 1, \quad \xi_i \geqslant 0. \tag{4}$$

Similarly, the optimization problem solved by the bidirectional distance-weighted discrimination can be stated as

$$\min_{w,\beta,\xi} \sum_i \left(\frac{1}{s_i} + C_{\mathrm{DWD}} \xi_i \right),$$

subject to

$$\begin{gathered} s_i = y_i(x_i^{\mathrm{T}} w_1 + \beta_1)(x_i^{\mathrm{T}} w_2 + \beta_2) + \xi_i \geqslant 0, \quad \xi_i \geqslant 0, \\ \|w_1\|^2 + \beta_1^2 = 1, \quad \|w_2\|^2 + \beta_2^2 = 1. \end{gathered} \tag{5}$$

To meet the uniqueness requirement, here we use the constraints $\|w_j\|^2 + \beta_j^2 = 1$ instead of $\|w_1\|^2 = \|w_2\|^2 = 1$, as used in the one-directional method. We choose this type of constraint to ensure that the optimization problem can be described in second-order cone programming terms.

The multiplicative form of the s_i in (3) poses great optimization challenges and makes it difficult to solve simultaneously for (w_1, β_1) and (w_2, β_2) in (4) and (5). However, provided one of the two hyperplanes is given, the other can be obtained using methods similar to the one-directional problem. This suggests that iterative algorithms be used, so we propose to solve the bidirectional minimization problem by minimizing a sequence of one-directional subproblems, as follows: first, propose initial values for $\{w_1^{(0)}, \beta_1^{(0)}\}$; obtain $\{w_2^{(0)}, \beta_2^{(0)}\}$ by solving the revised one-directional problems with y_i replaced by $\hat{y}_i = y_i(x_i^{\mathrm{T}} w_1^{(0)} + \beta_1^{(0)})$; based on $\{w_2^{(0)}, \beta_2^{(0)}\}$, obtain $\{w_1^{(1)}, \beta_1^{(1)}\}$ and repeat this process until convergence of both parameters. In all cases we have considered, this algorithm has converged in at most ten steps.

2·3. *Starting points*

The solutions of (4) and (5), based on the iterative algorithm described above, strongly depend on the choice of the initial values, especially in high-dimensional situations. Our next goal is to propose some appropriate ways to choose good initial values. We have considered a full quadratic

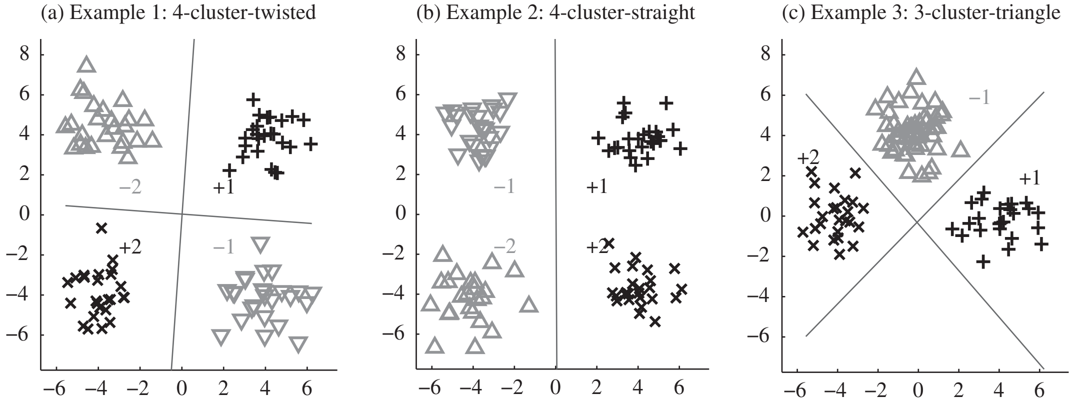

Fig. 2. Illustration of different subcluster structures. Black indicates the positive class. Grey indicates the negative class. Different symbols in the same class represent different subclusters. These show different levels of challenge to linear discriminant methods, all of which are resolvable using bidirectional discrimination.

projection approach that chooses two initial hyperplanes to be those whose product is closest to the hypersurface solved using the full quadratic kernel method. Because of relatively poor performance, further discussion of this approach appears only in the Supplementary Material. Now we will introduce approaches based on within-class clustering.

As one of the motivations of our method comes from the fact that there might be further subclusters within each class, we can illustrate our approaches using three bidimensional examples shown in Fig. 2. Three subcluster structures are considered, with each subcluster sampled from a shifted standard bivariate normal distribution determined by a common parameter μ, taken as $\mu = 5^{1/2}$.

Example 1. Four-cluster-twisted is shown in Fig. 2(a), which includes four clusters, two for each class. The clusters $+1$, $+2$, -1, -2 are shifted by (μ, μ), $(-\mu, -\mu)$, $(\mu, -\mu)$ and $(-\mu, \mu)$, respectively. Each cluster has sample size 25. This is particularly challenging for linear discrimination methods.

Example 2. Four-cluster-straight is shown in Fig. 2(b), which includes four clusters, two for each class. The clusters $+1$, $+2$, -1, -2 are shifted by (μ, μ), $(\mu, -\mu)$, $(-\mu, \mu)$ and $(-\mu, -\mu)$, respectively. Each cluster has sample size 25. Linear methods can be expected to perform well here.

Example 3. Three-cluster-triangle is shown in Fig. 2(c), which includes three subclusters, two for the positive class and one for the negative class. The clusters $+1$, $+2$, -1 are shifted by $(\mu, 0)$, $(-\mu, 0)$ and $(0, \mu)$, respectively. Sample sizes $n_{+1} = n_{+2} = n_{-1}/2 = 25$. This is also challenging for linear methods.

The three examples given above represent settings with at most two clusters per class. For more complex settings, e.g., classes with more than two subpopulations, we find that treating them as two subpopulations will frequently give better performance than a single linear method. For those cases where the subpopulations cannot be combined in a useful way, we can handle them by generalizing bidirectional discrimination to more directions as shown in the Supplementary Material.

For Example 1, the ideal choice for the initial hyperplane will be the one-directional hyperplane that separates groups $(+1, -1)$ and $(+2, -2)$ or the one that separates groups

$(+1, -2)$ and $(+2, -1)$. Therefore, our cluster-2-2 method first uses the 2-means clustering algorithm to divide the positive class into two clusters labelled as c_{+1} and c_{+2} and similarly divides the negative class into two clusters labelled as c_{-1} and c_{-2}. Then we choose the initial hyperplane as the usual one-directional hyperplane that either separates between groups (c_{+1}, c_{-1}) and (c_{+2}, c_{-2}) or separates between groups (c_{+1}, c_{-2}) and (c_{+2}, c_{-1}).

Similarly, cluster-1-1 and cluster-1-2 methods are motivated by Examples 2 and 3, respectively. We will see that each method for finding initial values has a situation for which it works the best. Typically, there is no prior knowledge as to the subcluster structure of the dataset. Therefore, we propose to implement all of these proposed initial values and take our solution to be the one that gives the minimum value of the objective function.

3. Visualization and numerical data

3·1. *Simulated high-dimensional examples*

We now investigate the performance of the proposed method using simulated data. We have tried simulations for both low- and high-dimensional situations. Since our main focus is on high-dimensional low-sample-size settings, low-dimensional results are given in the Supplementary Material. We set the sample sizes of training and test data as 100 and 1000, respectively. We generated the test data from the same distributions as the training data. In this paper, bidirectional discrimination is implemented using both the distance-weighted discrimination and support vector machine methods. As the results were similar, we focus here on distance-weighted discrimination. However, for the final comparisons, the support vector machine is included.

It is of interest to compare both methods with the kernel support vector machine, kernel distance-weighted discrimination and random forest methods (Breiman, 2001). The two kernels used here are the full quadratic and the Gaussian. The bandwidth parameter in the Gaussian kernel support vector machine was tuned via crossvalidation. Crossvalidation was also used to tune the penalty parameters in distance-weighted discrimination and in the support vector machine. For random forest, we used the R package (R Development Core Team, 2012) randomForest and took the default value 500 for the number of trees to grow.

Consider a typical high-dimensional low-sample-size context. Let $d = 1000$. We simulated three types of examples. The first two dimensions are generated using distributions similar to Examples 1–3. We maintain an appropriate signal-to-noise ratio by taking $\mu = d^{1/2}/8$ instead of the constant $\mu = 5^{1/2}$. The rest of the $d - 2$ dimensions are pure noise, i.e., all sampled from the standard normal distribution. The four different initialization options considered in § 2·3 are used. The combined solution is determined from the one that gives the minimum objective function value among the four options.

The visualization results of the simulated training data for the four-cluster-twisted high-dimensional example are shown in Fig. 3. The visualization results for other examples are shown in the Supplementary Material. From Fig. 3, cluster-2-2 seems to find the right structure. The combination of one-directional distance-weighted discrimination and orthogonal first principal component directions can separate the four clusters, but the structure is twisted in contrast to the original one, as shown in the upper-left plot. No structure of this type is part of the underlying signal in the data, so we conclude that it is a noise artefact. All the other bidirectional discrimination initialization methods exhibit artefacts that suggest overfitting. Cluster-1-2 attempts to divide the data into three clusters and gives an apparently reasonable separation of the negative class into two clusters. Cluster-1-1 attempts to divide the data into only two clusters. Even cluster-2-2 seems to show some overfitting, as the clusters are better separated than in the raw data.

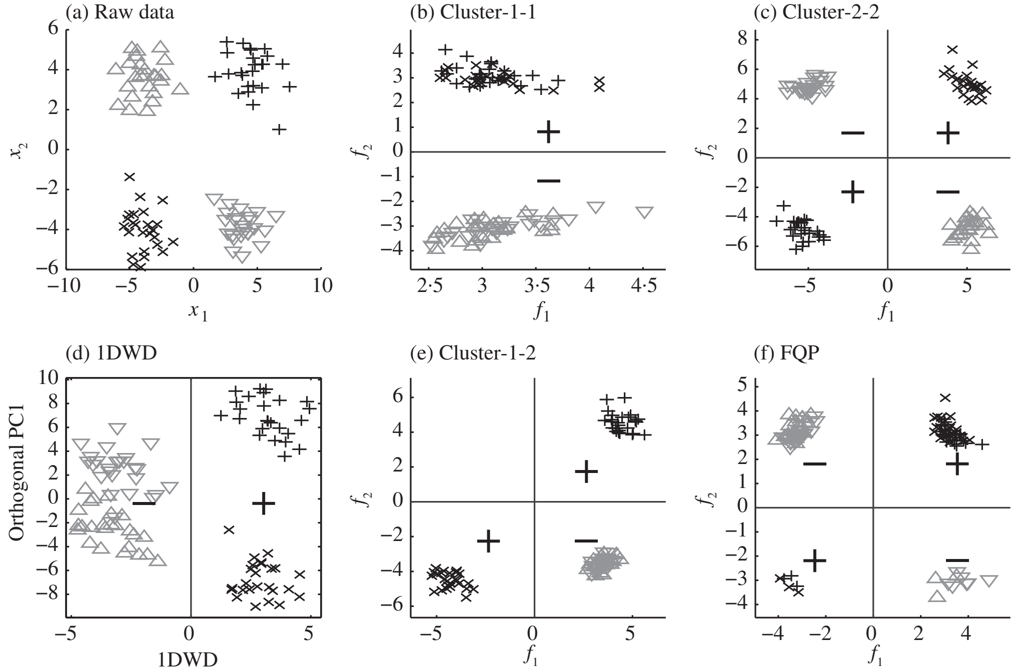

Fig. 3. Application to the four-cluster-twisted high-dimensional simulated dataset. (a) Raw data projected onto the first two directions. (d) Projections onto the one-directional distance-weighted discrimination and an orthogonal principal component direction. (b), (c), (e) and (f) Projections onto the f_1, f_2 directions. 1DWD, one-directional distance-weighted discrimination; PC1, the first principal component; FQP, full quadratic projection.

Table 1. *Performance summary, average error rates in percentage over* 100 *simulations, of the application of the one-directional and the bidirectional distance-weighted discrimination methods to three high-dimensional simulated examples*

		Bidirectional distance-weighted discrimination				
	1DWD	Cluster-2-2	Cluster-1-2	Cluster-1-1	FQP	Combined
Example 1	50·2	0·2	21·9	50·1	39·7	0·2
Example 2	0·1	50·3	43·9	0·1	39·4	0·1
Example 3	14·9	26	6·1	26·4	30·4	6·1

1DWD, one-directional distance-weighted discrimination; FQP, full quadratic projection. The largest standard error for the numbers in the table is 0·3.

To analyse which methods have found reproducible structure in the data, we repeat the simulation 100 times; see Table 1. For Example 1, not surprisingly, the cluster-2-2 method works the best. The cluster-1-1 method is no better than random choice. The performances of the cluster-1-2 and the full quadratic projection methods are in between these. For Example 2, the cluster-1-1 method works best, as expected, and the cluster-2-2 method is no better than random choice. For Example 3, the cluster-1-2 method works best, as expected. For all three examples, the combined method always chooses the best among the four initialization methods. On the other hand, the standard one-directional distance-weighted discrimination method is best only in Example 2. It is no better than random choice in Example 1 and gives moderate performance in Example 3. For all examples, the full quadratic projection method typically was far from the best performance, because it works in a space with much higher dimension than the original one and thus is more prone to overfitting.

Table 2. *Comparison of test errors (%) among different methods on three high-dimensional simulated datasets*

Method	Example 1	Example 2	Example 3
Linear support vector machine	50	0·2	16
Linear distance-weighted discrimination	49	0·1	15
Quadratic kernel support vector machine	4	44	26
Quadratic kernel distance-weighted discrimination	3	44	24
Gaussian kernel support vector machine	16	0·1	16
Gaussian kernel distance-weighted discrimination	46	0·1	14
Bidirectional support vector machine	0·4	0·2	16
Bidirectional distance-weighted discrimination	0·2	0·1	6
Random forest	50·1	4	28

The largest standard error for the numbers in the table is 0·3.

Table 2 summarizes the comparison of one-directional linear, full quadratic kernel, Gaussian kernel, and bidirectional discrimination methods based on support vector machine and distance-weighted discrimination implementations for the three simulated high-dimensional datasets. Results based on the random forest method are also included. For Example 1, one-directional linear methods were no better than random choice, quadratic kernel methods give much improvement but bidirectional discrimination methods are the best. Random forest is also no better than random. For Example 2, all methods work well except the quadratic kernel methods. For Example 3, bidirectional distance-weighted discrimination gives the best performance. From Table 2, we can see that each kernel method works well in some situations, and the special strength of bidirectional discrimination comes from its ability to frequently mimic the performance of any of linear, quadratic or Gaussian kernel methods, in situations where each is the best. The random forest works well in certain situations where the two classes can be well separated by the linear methods. In situations where the two classes cannot be separated by the linear methods, such as Example 1, the performance of the random forest is poor.

The relatively poor performance of the regularized kernel methods may be somewhat surprising. Certainly, regularization is critical to avoiding overfitting in such high-dimensional settings. In the Supplementary Material, it is carefully checked that this is not simply an artefact of poor tuning. As the dimension increases, any method will eventually break down. Our results indicate that bidirectional discrimination has better breakdown properties in this sense than other methods considered in Table 2. It would be interesting to see a more explicit study of this phenomenon in future work.

To study interpretability, we show relative contributions of each variable for the one-directional linear and bidirectional support vector machine methods shown in Fig. 4. Recall that the target direction is nonzero only in the first two entries. The bidirectional method correctly picks the first two variables, whereas the one-directional method assigns roughly equal importance to all variables. Thus, in this example, bidirectional results give much better interpretability in terms of directions found than linear methods.

3·2. *Real data*

In this section, we apply our method to a real glioblastoma dataset. In this example, the subcluster labels for each class are unknown and we also want to see whether or not our method can discover some subclusters within each class.

Glioblastoma multiforme is the most common form of malignant brain cancer in adults. For the purposes of the current analysis, we selected a cohort of patients with glioblastoma cancer whose

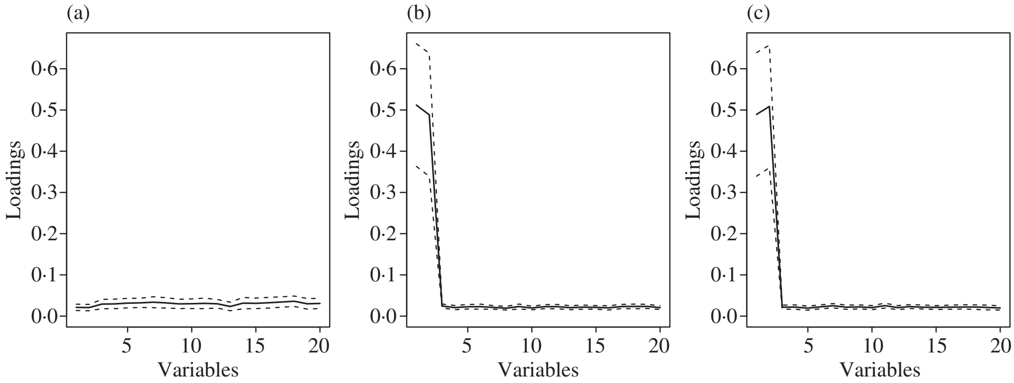

Fig. 4. The averages (solid lines) with ±2 standard deviations (dashed lines), over 100 replications of the absolute values of the first 20 entries, i.e., loadings, of the normal direction vectors from applying the support vector machine to Example 1. (a) Based on the one-directional method. (b) Based on the f_1 function from the bidirectional method. (c) Based on the f_2 function from the bidirectional method.

Table 3. *Average crossvalidation errors* (%) *over* 100 *replications for the glioblastoma microarray dataset*

	Bidirectional distance-weighted discrimination				
1DWD	Cluster-2-2	Cluster-1-2	Cluster-1-1	FQP	Combined
3·41 (0·34)	9·89 (0·73)	2·84 (0·30)	3·22 (0·35)	4·73 (0·51)	2·73 (0·26)

Standard errors are shown in parentheses. 1DWD, one-directional distance-weighted discrimination; FQP, full quadratic projection.

brain samples were assayed to obtain the corresponding gene expression data. Several clinically relevant subtypes were identified using integrated genomic analysis discussed in Verhaak et al. (2010). After filtering the genes using the ratio of the sample standard deviation and sample mean of each gene, the dataset contains 186 patients with 2727 genes. Our analysis focused on mesenchymal and neural subtypes because there was an impression that the latter might have two subclasses. There are 117 mesenchymal samples and 69 neural samples.

We consider the classification problem that treats mesenchymal as the positive class and neural as the negative class. Due to the limited sample size, we study the generalization properties of our method using crossvalidation. The dataset is split into 80 and 20% for a training set and a test set. We further split the training set with 80 and 20% to give crossvalidation for tuning parameter selection. We use a stratified sampling scheme; splits are random subject to the constraint that we keep the original proportion of each class in the training set and the test set. The division of the training data is randomly repeated 100 times and the tuning parameter is chosen to be the one that gives the lowest average crossvalidation error. The test error is calculated based on this parameter.

The crossvalidation errors for the classification problem, listed in Table 3, are computed on the basis of 100 random splits of the dataset. The crossvalidation errors show that the cluster-1-2 method gives the best performance among the methods considered, although the difference between the cluster-1-2 and one-directional methods is not highly significant.

Projection of the data onto the f_1 and f_2 directions from the cluster-1-2 method is shown in Fig. 5. The data naturally fall into three clusters: one for the mesenchymal class and two for the neural class. Our analysis provides new evidence for the notion of two subclusters in the neural class. To confirm whether or not these clusters represent potentially important new cancer

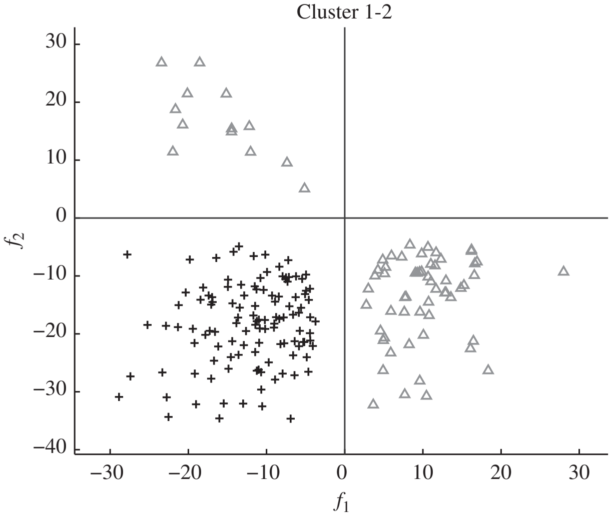

Fig. 5. Visualization of the glioblastoma dataset using two directions from the cluster-1-2 method. Black plus signs denote mesenchymal and grey triangles denote neural.

Table 4. *Crossvalidated error comparison of different methods for the glioblastoma data*

Linear support vector machine	3·03 (0·19)
Linear distance-weighted discrimination	3·41 (0·28)
Quadratic kernel support vector machine	10·1 (0·34)
Quadratic kernel distance-weighted discrimination	9·30 (0·60)
Gaussian kernel support vector machine	3·35 (0·23)
Gaussian kernel distance-weighted discrimination	3·51 (0·34)
Bidirectional support vector machine	2·90 (0·19)
Bidirectional distance-weighted discrimination	2·73 (0·28)
Random forest	4·17 (0·34)

Standard errors are shown in parentheses.

subclasses, the statistical significance of the clustering method (Liu et al., 2008) was performed to evaluate the significance of the split of the neural class into two clusters. The resulting p-value is less than 0·001, so we conclude that there are further subclusters within the neural samples that are worth deeper biological investigation.

Two subsets of genes were selected based on the 200 biggest absolute values of the loadings, i.e., coefficients in f_1 and f_2 from the cluster-1-2 method. The number of common genes in the two subsets is 26. Let Neural-1 denote the neural samples that satisfy $f_2 < 0$, and Neural-2 denote those that satisfy $f_1 < 0$, see Fig. 5. The Supplementary Material contains heatmap visualizations, which give deeper insights into the driving genes for these subclasses. Thus, our bidirectional discrimination method not only improves the classification performance but also provides an effective tool for feature selection.

Table 4 summarizes the crossvalidation comparison of linear, full quadratic kernel, Gaussian kernel, and bidirectional discrimination methods implemented through both support vector machine and distance-weighted discrimination for the glioblastoma data. The result for the random forest method is also included. For this example, two lowest errors are obtained by the bidirectional distance-weighted discrimination and support vector machine methods.

The quadratic kernel methods give the worst performance. The performances of the linear, Gaussian kernel and random forest methods are in between. The improved crossvalidation errors of the bidirectional methods over the one-directional methods seem to be due to the distinct subclusters in the neural class. This example shows how further consideration of the subcluster structure can improve the classification error rates. We have also studied another real example of lung cancer data in the Supplementary Material where the subcluster structures are known and it is shown that the bidirectional discrimination methods correctly identify them without using the cluster labels.

4. High-dimensional low-sample-size asymptotics

4·1. *Four clusters case*

To gain further insight into bidirectional discrimination, in this section we study some of its theoretical properties. We consider asymptotics of the method for $d \to \infty$ with the sample size n fixed. Hall et al. (2005) first demonstrated the insight available from such asymptotics. They showed that, under some conditions, each data point in a sample of size n tends to lie near a vertex of a regular n-simplex and all the randomness in the data appears in the form of a random rotation of this simplex. This data structure yields new insight into the binary classification problem. In practice, data points from the positive class of size n^+ and those from the negative class of size n^- can be viewed as an n^+-simplex and an n^--simplex, respectively. This gave direct results on completely perfect and completely imperfect classifications.

The regularity conditions for the geometric representation in Hall et al. (2005) require that the entries of the data vector satisfy a ρ-mixing condition. Ahn et al. (2007) gave a milder condition using asymptotic properties of the sample covariance. A more general and even milder set of conditions for the result of Hall et al. (2005) is given in Jung & Marron (2009) and Qiao et al. (2010).

To illustrate the principles underlying bidirectional discrimination, we consider two examples. The first includes four clusters labelled as $+1$, $+2$, -1 and -2. Assume that data points from clusters $+1$ and $+2$ belong to the positive class and those from clusters -1 and -2 belong to the negative class.

We use the regularity conditions of Qiao et al. (2010). Consider the $+1$ cluster consisting of data vectors $x_1^{+1}(d), \ldots, x_{n_{+1}}^{+1}(d)$ with d variables, where $x_j^{+1}(d) = (x_{1j}^{+1}, \ldots, x_{dj}^{+1})^{\mathrm{T}} \in R^d$ $(j = 1, \ldots, n_{+1})$. Assume that these vectors are independent and identically distributed from a d-dimensional multivariate distribution. Concatenate these into a $d \times n_{+1}$ data matrix $X_d^{+1} = [x_1^{+1}(d), \ldots, x_{n_{+1}}^{+1}(d)]$.

For a fixed n_{+1}, consider a sequence of random data matrices $X_1^{+1}, \ldots, X_d^{+1}, \ldots$, indexed by the number of rows d. Assume that each X_d^{+1} comes from a d-dimensional multivariate distribution with covariance matrix Σ_d^{+1}. Let $\lambda_{1,d}^{+1} \geqslant \cdots \geqslant \lambda_{d,d}^{+1}$ be the eigenvalues of Σ_d^{+1}. Assume the following.

Assumption 1. The fourth moments of each entry of each column of X_d^{+1} are uniformly bounded.

Assumption 2. The entries of $Z_d^{+1} = (\Sigma_d^{+1})^{(-1/2)} X_d^{+1}$ are independent.

Assumption 3. The eigenvalues of Σ_d^{+1} are sufficiently diffuse, in the sense that

$$\epsilon_d^{+1} = \frac{\sum_{j=1}^{d} (\lambda_{j,d}^{+1})^2}{\left(\sum_{j=1}^{d} \lambda_{j,d}^{+1}\right)^2} \to 0, \quad d \to \infty.$$

Assumption 4. The sum of the eigenvalues of $\Sigma_{+1,d}$ is of the same order as d, in the sense that $\sum_{j=1}^{d} \lambda_{j,d}^{+1} = O(d)$ and $1/\sum_{j=1}^{d} \lambda_{j,d}^{+1} = O(1/d)$.

Define the scaled variance $(\sigma_d^{+1})^2 = d^{-1} \sum_{j=1}^{d} \lambda_{j,d}^{+1}$. Under Assumptions 1–4, as $d \to \infty$, these n_{+1} data vectors tend to form a regular n_{+1}-simplex in R^d with the side length $2^{1/2} d\sigma_d^{+1}$. Assume that the other three independent data samples X_d^{+2}, X_d^{-1} and X_d^{-2} also satisfy Assumptions 1–4. Define σ_d^{+2}, σ_d^{-1} and σ_d^{-2} similarly to σ_d^{+1}. Then the four clusters can be viewed asymptotically as four simplices in R^d with side lengths $2^{1/2} d\sigma_d^{+1}$, $2^{1/2} d\sigma_d^{+2}$, $2^{1/2} d\sigma_d^{-1}$ and $2^{1/2} d\sigma_d^{-2}$, respectively. Based on this geometric representation, our next goal is to develop conditions under which the bidirectional method is better than the usual one-directional method.

In general, the population mean positions of the four clusters lie in a three-dimensional hyperplane in R^d. They are located at the vertices of a tetrahedron. In order to illustrate the basic idea of when the bidirectional method is preferred, we consider a simple setting here. Given two sequences of between-class distances $l_{+,d} \geqslant 0, l_{-,d} \geqslant 0$ and a sequence of within-class distances $l_{0,d} \geqslant 0$, the mean positions of the four clusters in the three-dimensional space are $C_{+1,d} = (l_{+,d}/2, 0, l_{0,d}/2)$, $C_{+2,d} = (-l_{+,d}/2, 0, l_{0,d}/2)$, $C_{-1,d} = (0, l_{-,d}/2, -l_{0,d}/2)$, $C_{-2,d} = (0, -l_{-,d}/2, -l_{0,d}/2)$, respectively. These mean positions can also be parameterized in terms of variance shifts and rotations, but this form makes the main idea most clear. The geometries of this setting are fully characterized by three sequences $l_{+,d}$, $l_{-,d}$ and $l_{0,d}$. Here for simple understanding of the main ideas, we assume that the sample sizes and variances of the two clusters within each class are the same.

Assumption 5. The sample sizes satisfy $n_{+1} = n_{+2} = n_+/2$, $n_{-1} = n_{-2} = n_-/2$.

Assumption 6. For given constants $\sigma_+, \sigma_- > 0$, $\sigma_{+1,d} = \sigma_{+2,d} = \sigma_{+,d} \to \sigma_+$, $\sigma_{-1,d} = \sigma_{-2,d} = \sigma_{-,d} \to \sigma_-$, as $d \to \infty$.

For some distance orders $\alpha_\pm \geqslant 0$ and $\alpha_0 \geqslant 0$, we study the asymptotic behaviours of the one- and bidirectional classifiers as the within-class distances $l_{\pm,d}$ grow at the rate of $d^{\alpha_\pm}$ and the between-class distance $l_{0,d}$ grows at the rate of d^{α_0}, in the sense that $l_{\pm,d}/d^{\alpha_\pm} \to \mu_\pm$ and $l_{0,d}/d^{\alpha_0} \to \mu_0$ for some $\mu_\pm > 0$ and $\mu_0 > 0$. To test the performance of the classification methods, we need to add a new random point to a d-variate space which is independent of the data in $X_d^{+1} \cup X_d^{+2} \cup X_d^{-1} \cup X_d^{-2}$ and has the distribution of any of the four clusters.

THEOREM 1. *Without loss of generality, assume that $\sigma_+^2/n_+ > \sigma_-^2/n_-$; if need be, interchange $+$ and $-$ to achieve this. Under Assumptions 1–6, we have*:

(i) *the one-directional support vector machine gives either completely correct or incorrect classification as follows*: (a) *If* $\lim_{d\to\infty}(\mu_0^2 d^{2\alpha_0 - 1}) > \sigma_+^2/n_+ - \sigma_-^2/n_-$, *then the probability that the usual one-directional hyperplane gives correct classification of new points converges to* 1 *as* $d \to \infty$. (b) *If* $\lim_{d\to\infty}(\mu_0^2 d^{2\alpha_0 - 1}) < \sigma_+^2/n_+ - \sigma_-^2/n_-$, *then with probability converging to* 1 *as* $d \to \infty$ *a new datum from either population will be classified by the usual one-directional hyperplane as belonging to the positive population.*

(ii) *bidirectional discrimination gives completely correct classification as follows*: *If either* $\lim_{d\to\infty}(\mu_0^2 d^{2\alpha_0 - 1}) > \sigma_+^2/n_+ - \sigma_-^2/n_-$ *or* $\lim_{d\to\infty}(\mu_\pm d^{\alpha_\pm - 1/2}) > 0$, *the probability that the bidirectional classifier gives correct classification of new points converges to* 1 *as* $d \to \infty$.

Theorem 1(i) says that the one-directional support vector machine gives an asymptotically correct classification of a new point when the between-class distance is large enough, in the sense that either $\alpha_0 > 1/2$ or $\alpha_0 = 1/2$ and $\mu_0^2 > \sigma_+^2/n_+ - \sigma_-^2/n_-$. When the between-class distance is small enough, in the sense that either $\alpha_0 < 1/2$ or $\alpha_0 = 1/2$ and $\mu_0^2 < \sigma_+^2/n_+ - \sigma_-^2/n_-$, the one-directional method will fail regardless of the size of the within-class distances. Theorem 1(ii) shows that the bidirectional discrimination method works as well as the one-directional method when either $\alpha_0 > 1/2$ or $\alpha_0 = 1/2$ and $\mu_0^2 > \sigma_+^2/n_+ - \sigma_-^2/n_-$. More importantly, the major improvement available from bidirectional discrimination is demonstrated in the result that it will classify correctly when $\alpha_\pm \geqslant 1/2$, for any value of between-class distance.

The Supplementary Material contains an asymptotic characterization of the balanced case with $n_+ = n_-, \sigma_+ = \sigma_-$.

4·2. *Three clusters case*

The second example includes three clusters labelled as $+1$, $+2$ and -1. Thus, only the positive class contains two distinct clusters. Similar to the four clusters case, we consider a simple setting here. Given a within-class distance $l_{+,d} \geqslant 0$ and a between-class distance $l_{0,d} \geqslant 0$, the mean positions of the three clusters $+1$, $+2$, -1 in the bidimensional space are $C_{+1,d} = (l_{+,d}/2, l_{0,d}/2)$, $C_{+2,d} = (-l_{+,d}/2, l_{0,d}/2)$, $C_{-,d} = (0, -l_{0,d}/2)$. Further, we have the following assumptions.

Assumption 7. The sample sizes satisfy $n_{+1} = n_{+2} = n_+/2$.

Assumption 8. For a given constant $\sigma_+ > 0$, $\sigma_{+1,d} = \sigma_{+2,d} = \sigma_{+,d} \to \sigma_+$, as $d \to \infty$.

The following theorem characterizes the high-dimensional low-sample-size data asymptotics of the one-directional and bidirectional methods under the above setting.

Theorem 2. *Without loss of generality, assume that $\sigma_+^2/n_+ > \sigma_-^2/n_-$; if need be, interchange $+$ and $-$ to achieve this. Under Assumptions 1–4 and Assumptions 7–8, we have the following results*:

(i) *the one-directional support vector machine gives either completely correct or incorrect classification as follows*: (a) *If* $\lim_{d\to\infty}(\mu_0^2 d^{2\alpha_0-1}) > \sigma_+^2/n_+ - \sigma_-^2/n_-$, *then the probability that the usual one-directional hyperplane gives correct classification of new points converges to* 1 *as $d \to \infty$*. (b) *If* $\lim_{d\to\infty}(\mu_0^2 d^{2\alpha_0-1}) < \sigma_+^2/n_+ - \sigma_-^2/n_-$, *then with probability converging to* 1 *as $d \to \infty$ a new datum from either population will be classified by the usual one-directional hyperplane as belonging to the positive population.*

(ii) *Bidirectional discrimination gives completely correct classification as follows*: *if either* $\lim_{d\to\infty}(\mu_0^2 d^{2\alpha_0-1}) > \sigma_+^2/n_+ - \sigma_-^2/n_-$ *or* $\lim_{d\to\infty}(\mu_+^2 d^{2\alpha_+-1}) > 8\sigma_+^2/n_+$, *the probability that the bidirectional classifier gives correct classification of new points converges to* 1 *as $d \to \infty$.*

Acknowledgement

The authors thank the reviewers for many helpful suggestions. This work was supported by the National Science Foundation and National Institutes of Health, U.S.A. Yufeng Liu is also affiliated with the Carolina Center for Genome Sciences at the University of North Carolina.

Supplementary material

Supplementary material available at *Biometrika* online includes additional simulated and real examples, the extension of our method to the multi-directional case and proofs of all the theorems.

References

Ahn, J., Marron, J. S., Muller, K. M. & Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94**, 760–6.

Aizerman, A., Braverman, E. M. & Rozoner, L. I. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Auto. Remote Contr.* **25**, 821–37.

Alizadeh, F., Alizadeh, F., Goldfarb, D. & Goldfarb, D. (2001). Second-order cone programming. *Math. Prog.* **95**, 3–51.

Boser, B. E., Guyon, I. M. & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proc. 5th Ann. ACM Workshop Comp. Learn. Theory*, Ed. D. Haussler, pp. 144–52. New York: ACM Press.

Breiman, L. (2001). Random forests. *Mach. Learn.* **45**, 5–32.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining Know. Disc.* **2**, 121–67.

Cortes, C. & Vapnik, V. (1995). Support vector networks. *Mach. Learn.* **20**, 273–97.

Cristianini, N. & Taylor, S. J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–88.

Hall, P., Marron, J. S. & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B **67**, 427–44.

Hastie, T. J., Tibshirani, R. J. & Friedman, J. (2001). *The Elements of Statistical Learning*. Berlin: Springer.

Jung, S. & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37**, 4104–30.

Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *J. Am. Statist. Assoc.* **103**, 1281–93.

Marron, J. S., Todd, M. & Ahn, J. (2007). Distance-weighted discrimination. *J. Am. Statist. Assoc.* **102**, 1267–71.

Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J. & Marron, J. S. (2010). Asymptotic properties of distance-weighted discrimination. *J. Am. Statist. Assoc.* **105**, 401–14.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., et al. & Cancer Genome Atlas Research Network (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110.

[*Received February* 2011. *Revised April* 2012]