A New Approach That Selects a Single Hyperplane from the Optimal Pairwise Linear Classifier

Luis Rueda*

School of Computer Science, University of Windsor,
401 Sunset Avenue, Windsor, ON, Canada, N9B 3P4.
lrueda@uwindsor.ca

Abstract. In this paper, we introduce a new approach to selecting the *best hyperplane classifier (BHC)* from the optimal pairwise linear classifier is given. We first propose a procedure for selecting the BHC, and analyze the conditions in which the BHC is selected. In one of the cases, it is formally shown that the BHC and *Fisher's classifier (FC)* are coincident. The empirical and graphical analysis on synthetic data and real-life datasets from the UCI machine learning repository, which involves the optimal quadratic classifier, the BHC, the optimal pairwise linear classifier, and FC.

1 Introduction

Linear classifiers have been extensively studied because of their classification speed and their simplicity in the implementation. We consider two classes, c_1 and c_2, which are represented by two normally distributed d-dimensional random vectors, $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Thus, the statistical information about the classes is determined by the mean vectors, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and the covariance matrices, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. We assume that these parameters are already known, or estimated by using a conventional estimation method, such as the *maximum likelihood estimate (MLE)*, *the Bayesian estimate* [4,14], etc. We also assume that the *a priori* probabilities of the two classes are equal. When dealing with two normally distributed random vectors, the general form of the optimal Bayesian classifier is quadratic. In special cases, the quadratic function can be factored as a product of two linear functions, as follows:

$$g_1(\mathbf{x})g_2(\mathbf{x}) \underset{c_2}{\overset{c_1}{\lessgtr}} 0\,, \tag{1}$$

where $g_1(\mathbf{x}) = \mathbf{w}_1^t\mathbf{x} + w_1$ and $g_2(\mathbf{x}) = \mathbf{w}_2^t\mathbf{x} + w_2$.

This is possible when the necessary and sufficient conditions hold [11,12]. Although (1) is optimal, and it achieves high classification accuracy, it requires two linear algebraic operations to classify a single object. We will see later in this paper that using the best of these two hyperplanes leads to nearly optimal classification.

* Member, IEEE. Partially supported by NSERC, the Natural Science and Engineering Research Council of Canada.

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 521–528, 2003.

Various schemes that yield linear classifiers have been reported in the literature, including *Fisher's classifier* [4,6,13], the *perceptron algorithm* (the basis of the back propagation *neural network* learning algorithms) [9], *piecewise recognition models* [7], *random search optimization* [8], *removal classification structures* [1], *adaptive linear dimensionality reduction* [5] (which outperforms Fisher's classifier for some data sets), *linear constrained distance-based classifier analysis* [3] (an improvement to Fisher's approach designed for hyperspectral image classification), and *recursive Fisher's discriminant* [2].

Rueda and Oommen [11,12] have recently shown that the optimal classifier between two normally distributed classes can be linear even when the covariance matrices are not equal. They showed that although the optimal classifier for normally distributed random vectors is a second-degree polynomial, this polynomial degenerates to be either a single hyperplane or a pair of hyperplanes. In this paper, we introduce a novel approach to selecting the *best hyperplane classifier* (BHC) in the framework of optimal pairwise linear classifiers.

2 Optimal Pairwise Linear Classifiers

Let $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ be two normally distributed random vectors. The three cases and the conditions in which the optimal classifier is a pair of hyperplanes are listed below.

Case I: Suppose that[1]

$$\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = [\mu_1, \ldots, \mu_d]^t,\ \boldsymbol{\Sigma}_1 = \mathbf{I}, \text{ and } \boldsymbol{\Sigma}_2 = \text{diag}(a_1^{-1}, \ldots, a_d^{-1}). \quad (2)$$

The optimal classifier is a pair of hyperplanes if and only if any of the following conditions is satisfied.

(*i*) $0 < a_i < 1$, $a_j > 1$, $a_k = 1$, $\mu_k = 0$, for all $k = 1, \ldots, d$, $i \neq j$, $k \neq i$, $k \neq j$, with

$$a_i(1-a_j)\mu_i^2 + a_j(1-a_i)\mu_j^2 - \frac{1}{4}(a_i a_j - a_i - a_j + 1)\log(a_i a_j) = 0\,. \quad (3)$$

(*ii*) $a_i \neq 1$, $a_j = 1$, $\mu_j = 0$, for all $j \neq i$ ..

(*iii*) $a_i = 1$, for all $i = 1, \ldots, d$.

When $d = 2$, and the parameters have the form of:

$$\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = [r, s]^t,\ \boldsymbol{\Sigma}_1 = \mathbf{I},\ \boldsymbol{\Sigma}_2 = \text{diag}(a^{-1}, b^{-1}), \quad (4)$$

the condition of (4) is instead:

$$a(1-b)r^2 + b(1-a)s^2 - \frac{1}{4}(ab - a - b + 1)\log(ab) = 0\,, \quad (5)$$

[1] In this paper, $\text{diag}(a_1, \ldots, a_d)$ represents a $d \times d$ diagonal matrix, whose diagonal elements are $a_1, \ldots, a_d$ respectively.

Case II: Suppose that

$$\boldsymbol{\mu}_1 = [\mu_1, \ldots, \mu_i, \ldots, \mu_j, \ldots, \mu_d]^t,$$
$$\boldsymbol{\mu}_2 = [\mu_1, \ldots, \mu_{i-1}, -\mu_i, \mu_{i+1}, \ldots, \mu_{j-1}, -\mu_j, \mu_{j+1}, \ldots, \mu_d]^t \quad (6)$$
$$\boldsymbol{\Sigma}_1 = \text{diag}(a_1^{-1}, \ldots, a_i^{-1} \ldots, a_j^{-1}, \ldots, a_d^{-1}), \text{ and}$$
$$\boldsymbol{\Sigma}_2 = \text{diag}(a_1^{-1}, \ldots, a_j^{-1} \ldots, a_i^{-1}, \ldots, a_d^{-1}) .. \quad (7)$$

The optimal classifier is a pair of hyperplanes if and only if $\mu_i^2 = \mu_j^2$. When $d = 2$, and the parameters are of the form:

$$\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = [r, s]^t,\ \boldsymbol{\Sigma}_1 = \text{diag}(a^{-1}, b^{-1}), \text{ and } \boldsymbol{\Sigma}_2 = \text{diag}(b^{-1}, a^{-1}), \quad (8)$$

the necessary and sufficient condition is $r^2 = s^2$.

Case III: Suppose that the covariance matrices have the form of (7), and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Then, the classifier is always a pair of hyperplanes.

3 Selecting the Best Hyperplane

First of all, we introduce the following definition, which will be fundamental in the criteria for selecting the BHC.

Definition 1. *Let $g(\mathbf{x})$ be the value resulting from classifying a vector $\mathbf{x}$. The* sign *of $g(\mathbf{x})$, $sgn(g, \mathbf{x})$, is defined as follows:*

$$sgn(g, \mathbf{x}) = \begin{cases} -1 \text{ if } g(\mathbf{x}) < 0 \\ 0 \text{ if } g(\mathbf{x}) = 0 \\ 1 \text{ if } g(\mathbf{x}) > 0 \end{cases} \quad (9)$$

In other words, a new sample falls in the "negative" side, $\text{sgn}(g, \mathbf{x}) = -1$, or in the "positive" side, $\text{sgn}(g, \mathbf{x}) = 1$. Ties are resolved arbitrarily, where we assign 0 to $\text{sgn}(g, \mathbf{x})$. The criteria for selecting the BHC is based on the result of classifying the two means, and uses Definition 1 to evaluate the sign resulting from the classification.

Rule 1 *Let $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ be two normally distributed random vectors, and $g_1(\mathbf{x})g_2(\mathbf{x})$ be the optimal pairwise linear classifier. The BHC is selected as per the following rule.*

Select:

- *g_1, if $sgn(g_1, \boldsymbol{\mu}_1) \neq sgn(g_1, \boldsymbol{\mu}_2)$,*
- *g_2, if $sgn(g_2, \boldsymbol{\mu}_1) \neq sgn(g_2, \boldsymbol{\mu}_2)$, or*
- *g_1 and g_2, if $sgn(g_1, \boldsymbol{\mu}_1) = sgn(g_1, \boldsymbol{\mu}_2) = 0$.* □

In other words, the BHC is the hyperplane that separates the space into two regions when the mean vectors are different. One region contains $\boldsymbol{\mu}_1$ and the other contains $\boldsymbol{\mu}_2$. When the mean vectors are coincident, both g_1 and g_2 are the best classifiers, and hence both must be selected.

We now analyze the conditions for selecting the BHC for the case in which the covariance matrices are the identity and a diagonal matrix respectively (Case I). The formal proof of the result can be found in [10]

Theorem 1. *Let* $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *be two normally distributed random vectors, where the parameters have the form of (4), and*

$$y_+ = \alpha(a-1)x + \alpha(a+1)r - \beta s\,, \text{ and} \tag{10}$$

$$y_- = \alpha(1-a)x - \alpha(a+1)r - \beta s\,, \tag{11}$$

be the linear functions (in their explicit form) composing the optimal pairwise linear classifier, where $\alpha = \frac{\sqrt{\frac{1-b}{a-1}}}{b-1}$ *and* $\beta = \frac{b+1}{b-1}$.

The linear function $g_1(\mathbf{x})$ *is selected as per Rule 1, if:*

$$\kappa \in (\frac{a}{b}r, r) \text{ when } g_1(\boldsymbol{\mu}_1) > 0 \text{ and } g_1(\boldsymbol{\mu}_2) < 0, \text{ or} \tag{12}$$

$$\kappa \in (r, \frac{a}{b}r) \text{ when } g_1(\boldsymbol{\mu}_1) < 0 \text{ and } g_1(\boldsymbol{\mu}_2) > 0, \tag{13}$$

where $\kappa = s\sqrt{\frac{a-1}{1-b}}$.

Conversely, $g_2(\mathbf{x})$ *is selected when* κ *is outside the intervals.*

The extension of Theorem 1 to d-dimensional normally distributed random vectors, where $d > 2$, is straightforward. The conditions for which the BHC is selected are similar to those of the two-dimensional case. The formal proof for the result can be found in [10]

We now analyze another case (Case II) in which the mean vectors are in opposite directions and the diagonal covariance matrices have the two elements of their diagonal switched.

Theorem 2. *Let* $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *be two normally distributed random vectors whose parameters are of the form of (8). The BHC is always:*

$$g_1(\mathbf{x}) = x + y = 0 \;\; \textit{if } r = s, \textit{ and} \tag{14}$$

$$= x - y = 0 \;\; \textit{if } r = -s\,.. \tag{15}$$

The formal proof of this theorem can be found in [10]. The extension to d-dimensional normally distributed random vectors, where $d > 2$, can be derived by replicating the steps of the proof of Theorem 2, and substituting r and s for μ_i and μ_j respectively. The formalization of the result is stated and proved in [10].

We now show that for the case discussed above, i.e. when the two distributions have mean vectors of the form of (6), and covariance matrices of the form of (7), the BHC is identical to Fisher's classifier. In the theorem below [10], we show the result for d-dimensional normally distributed random vectors, where $d \geq 2$.

Theorem 3. *Let* $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *be two normally distributed random vectors whose mean vectors and covariance matrices have the form of (6) and (7) respectively. The BHC is identical to Fisher's classifier.*

The third case that we consider is when we deal with two normally distributed random vectors whose covariance matrices have the form of (7), and their mean vectors are coincident. This case is the generalized Minsky's paradox for the perceptron. The result for two-dimensional normally distributed random vectors is stated as follows, and the proof is available in [10].

Theorem 4. *Let* $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *be two normally distributed random vectors whose covariance matrices have the form of (8), and whose mean vectors have the form* $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. *The BHC is composed of two linear functions:*

$$g_1(\mathbf{x}) = -x - y + (r + s), \text{ and } g_2(\mathbf{x}) = x - y + (s - r)\,. \tag{16}$$

The generalization of the result above for d-dimensional normally distributed random vectors, where $d > 2$, follows the same steps of the proof of Theorem 4. The details of the proof are found in [10]. The result of Theorem 4 is quite useful in deciding which linear function should be selected as the BHC, a single hyperplane or the pair of hyperplane composing the optimal pairwise linear classifier. Indeed, the case in which the distributions have coincident means rarely occurs in real-life scenarios.

The extension of the BHC classifier for more than two classes is straightforward. It can be achieved by deriving the BHC for each pair of classes. Then, the classification is performed by using the Voronoi diagram constructed using all the "inter-class" BHC classifiers. How this framework works in real-life scenarios is a problem that we are currently investigating.

4 Classification Accuracy and Speed

In order to test the accuracy and speed of the BHC and other two linear classifiers, we have performed some simulations for the different cases discussed in Section 3. We chose the dimensions $d = 2$ and $d = 3$ and trained our classifier using 100 randomly generated training samples, each sample represented by a two or three dimensional vector. For each case, we considered two classes, c_1 and c_2, which are represented by two normally distributed random vectors, $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \mathbf{I})$ and $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ respectively, where $\mathbf{I}$ is the identity.

The first case that we analyze consists of two examples that instantiate two-dimensional normally distributed random vectors, 2DD-1 and 2DD-2, whose mean vectors and covariance matrices satisfy the conditions of (4). The parameters are $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 \approx [0.747, 1.914]^t$, $\boldsymbol{\Sigma}_2 \approx \text{diag}(0.438, 5.827)$, $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 \approx [-1.322, -1.034]^t$, and $\boldsymbol{\Sigma}_2 \approx \text{diag}(2.126, 0.205)$ respectively.

The second case tested in our simulations is an example of two three-dimensional normally distributed random vectors, 3DD-1, whose covariance matrices and mean vectors, which satisfy the constraints of (2), are $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 \approx [0.855, 1.776, 0]^t$ and $\boldsymbol{\Sigma}_2 \approx \text{diag}(0.562, 3.842, 1)$.

In each case, the OPLC was obtained using the methods described in [11,12]. The BHC was obtained by invoking Rule 1 introduced in Section 3, and Fisher's classifier (FC) was obtained using the method described in [4].

To test the classifiers, we then generated ten sets each containing 100 random samples for each class using the original parameters. The results obtained after testing the three classifiers are shown in Table 1. The classification accuracy was computed as the average of the percentage of testing samples that were correctly classified for each of the ten data sets. Besides for each individual data set, the average between the classification accuracies for classes c_1 and c_2 was computed. The classification speed represents the average number of CPU seconds taken to classify 100 testing samples.

Table 1. Classification accuracy and speed obtained after testing three linear classifiers, OPLC, BHC and FC, on randomly generated data sets.

Example	OPLC		BHC		FC	
	Accuracy	Speed	Accuracy	Speed	Accuracy	Speed
2DD-1	92.15	4.48	91.85	1.75	91.15	1.80
2DD-2	96.20	4.38	96.15	1.73	95.15	1.79
3DD-1	93.40	4.57	93.30	2.11	92.35	1.82

For the first two examples, 2DD-1 and 2DD-2, the classification accuracy of the BHC is very close to that of the OPLC, and higher than that of FC. For the three-dimensional example, 3DD-1, we again observe the superiority of the BHC over FC. We also see that the BHC attains nearly optimal classification – just 0.1% less than the optimal classifier, OPLC. The BHC and FC are more than twice as fast as the OPLC, and both the BHC and FC achieve comparable speed rates. In Figure 1 the BHC, the OPLC, FC, and the samples of one of the testing data sets for each class, are plotted. It is clearly seen that FC misclassified objects which are in a region where the samples are more likely to occur. Similar plots for the other two examples are available in [10].

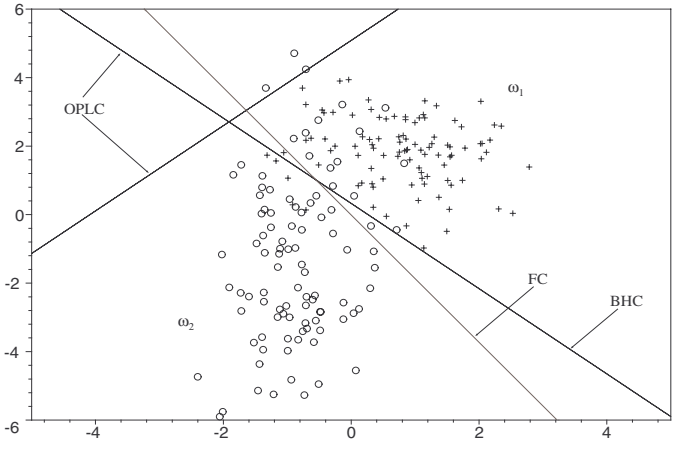

Fig. 1. Testing samples and the corresponding classifiers for two-dimensional normally distributed random vectors whose parameters are those of Example 2DD-1.

We also conducted experiments on real-life datasets. For the training and classification tasks we have composed 10 data subsets with all possible pairs of features obtained from the first five numeric features. For each of the pairs we composed the training set and the testing set by drawing samples without replacement from the original datasets.

The OQC and FC have been trained by invoking the traditional maximum likelihood method (MLE) [4,14]. The OPLC and the BHC have been trained by following the procedure described in [10], thus yielding the approximated pairwise linear classifier, and subsequently the best hyperplane, for each subset.

The classification of each object was performed using the classifiers mentioned above, and invoking a voting scheme, which assigns the class in which the sample yielded a positive result for the majority of voters. Ten voting rounds were invoked (one for each pair of features), and thus, the majority for class c_1 was chosen to be *five* or more voters. From the WDBC dataset, we randomly selected, without replacement, 100 samples for training, and 100 samples for the testing phase for each class. The classification accuracy obtained from testing the OQC, the OPLC, the BHC and FC are shown in Table 2. The results on the table show that using the voting scheme, as expected, the OQC is more accurate than the other classifiers. We also observe that the OPLC and the BHC (both achieving the same classification accuracy) lead to higher classification accuracy than FC. When considering the pair-based classification, the averages on the fifth column show that the OQC was the most accurate classifier. In this scheme, the BHC outperformed the OPLC, and FC was the least accurate classifier. We also observe that on the WDBC, the OPLC and the BHC achieve nearly optimal classification. Similar results that show the efficiency of the BHC, and a graphical analysis on real-life date are available in [10].

Table 2. Classification accuracy obtained from testing the classifiers on the WDBC data set.

Classifier	Benign	Malignant	Avg.(voting)	Avg.(pair)
OQC	96.00	87.00	91.50	88.45
OPLC	95.00	86.00	90.50	87.85
BHC	95.00	86.00	90.50	88.05
FC	93.00	85.00	89.00	86.30

5 Conclusions

In this paper, we presented an approach that selects the best hyperplane classifier (BHC) from the optimal pairwise linear classifier (OPLC). We first introduced the criteria for selecting the BHC given the OPLC. We then formalized the conditions for selecting the BHC for three cases. In the second case (the most general scenario for multi-dimensional random vectors), we have shown that the BHC is identical to Fisher's classifier (FC).

The efficiency of the BHC, the OPLC and FC has been evaluated in terms of classification accuracy and speed. In terms of accuracy, we have shown that the BHC is nearly optimal, and in some cases, it achieves the same accuracy as FC. The empirical results on real-life datasets show that the OPLC and the BHC attained similar classification accuracy, and that the BHC is superior to FC in the WDBC datasets. The graphical analysis corroborates this relation.

The extension of the BHC for d-dimensional random vectors, where $d > 2$, is far from trivial, as it involves to derive an MLE method for the constrained pairwise linear classifier. How this MLE is designed, and how the corresponding BHC is derived is a problem that is currently being undertaken.

References

1. M. Aladjem. Linear Discriminant Analysis for Two Classes Via Removal of Classification Structure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):187–192, 1997.
2. T. Cooke. Two Variations on Fisher's Linear Discriminant for Pattern Recognition. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 24(2):268–273, 2002.
3. Q. Du and C. Chang. A Linear Constrained Distance-based Discriminant Analysis for Hyperspectral Image Classification. *Pattern Recognition*, 34(2):361–373, 2001.
4. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, NY, 2nd edition, 2000.
5. R. Lotlikar and R. Kothari. Adaptive Linear Dimensionality Reduction for Classification. *Pattern Recognition*, 33(2):185–194, 2000.
6. W. Malina. On an Extended Fisher Criterion for Feature Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:611–614, 1981.
7. A. Rao, D. Miller, K. Rose, , and A. Gersho. A Deterministic Annealing Approach for Parsimonious Design of Piecewise Regression Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(2):159–173, 1999.
8. S. Raudys. On Dimensionality, Sample Size, and Classification Error of Nonparametric Linear Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):667–671, 1997.
9. S. Raudys. Evolution and Generalization of a Single Neurone: I. Single-layer Perception as Seven Statistical Classifiers. *Neural Networks*, 11(2):283–296, 1998.
10. L. Rueda. Selecting the Best Hyperplane in the Framework of Optimal Pairwise Linear Classifiers. Technical Report 02-009, School of Computer Science, University of Windsor, Windsor, Canada, 2002.
11. L. Rueda and B. J. Oommen. On Optimal Pairwise Linear Classifiers for Normal Distributions: The Two-Dimensional Case. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 24(2):274–280, February 2002.
12. L. Rueda and B. J. Oommen. On Optimal Pairwise Linear Classifiers for Normal Distributions: The d-Dimensional Case. *Pattern Recognition*, 36(1):13–23, January 2003.
13. R. Schalkoff. *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley and Sons, Inc., 1992.
14. A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, N.York, second edition, 2002.