Pattern Recognition Pergamon Press 1976 Vol 8 pp 127–139 Printed in Great Britain

PATTERN RECOGNITION BY MEANS OF DISJOINT PRINCIPAL COMPONENTS MODELS

SVANTE WOLD*

Research Group for Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå Sweden

(*Received* 25 *March* 1975 *and in revised form* 22 *August* 1975)

Abstract—Pattern recognition based on modelling each separate class by a separate principal components (PC) model is discussed These PC models are shown to be able to approximate any continuous variation within a single class Hence, methods based on PC models will, provided that the data are sufficient, recognize any pattern that exists in a given set of objects In addition, fitting the objects in each class by a separate PC model will, in a simple way, provide information about such matters as the relevance of single variables, "outliers" among the objects and "distances" between different classes Application to the classical *Iris*-data of Fisher is used as an illustration

Pattern recognition Principal components Karhunen–Loeve expansion Model fitting

1 INTRODUCTION

The search for, and use of, regularities in empirical data has always been of major concern in more complex sciences such as chemistry and biology A classical example in chemistry is the periodic system, where regularities in the properties of the chemical elements occur with a period of eight when the elements are arranged after increasing atomic weight Similarly, biologists have always classified plants and animals according to regularities and patterns in the morphological properties, such as the shape of flowers and leaves of plants, and the length and width of the scull and different bones etc of animals

Methods of data analysis specifically designed to detect regularities in multivariate data—often called methods of pattern recognition—are currently finding an increased use in all branches of science The classical problem in *pattern recognition* (henceforth abbreviated *PaRC*) can be formulated as follows Given a number of classes which each is defined by a set of objects—the training or *reference set of the class*—and the values of M measurements made on each of these objects, is it possible to classify new objects on the basis of the same M measurements made on these new objects?

A large variety of methods for handling this and related problems have been proposed. Recent reviews have been given by Kanal[1] and several others[2–7]

In branches of science such as chemistry and biology, however, the scope of a data analysis is usually wider than merely obtaining a classification of unassigned objects Often, one purpose of the analysis still can be dressed as one of classification, but since one usually cannot be certain that an object does not belong to a yet unseen and therefore unknown class, one wishes to discriminate not only among the known

* Part of this work was done between 1973 and 1974 while the author was on leave to the Department of Statistics, University of Wisconsin, Madison, WI 53706, U S A

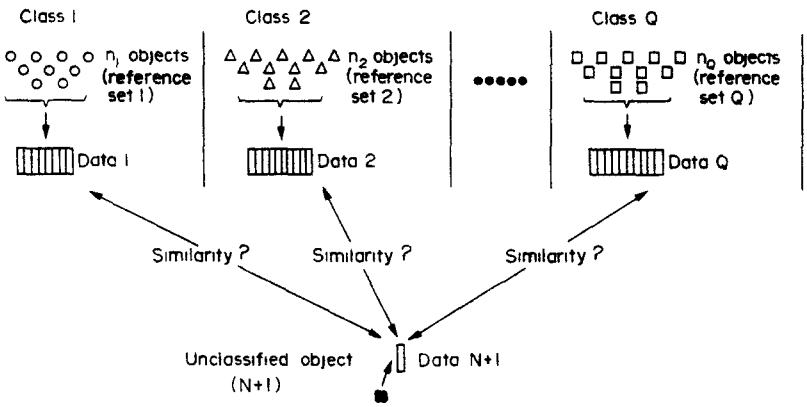

Fig 1 The Pattern Recognition problem Construct a method of assigning an object of unknown classification to the correct class on the basis of the values of the M variables measured on the object

classes but also possible unknown classes In this context of classification, it is also important that the data analytic method does not overemphasise the separation of the classes The case that the difference between two or more classes is very small or absent with respect to the variables used must be taken into account

Second, if one views the behaviour of the objects in a single class as a behaviour of "analogy", one is interested in which properties of the objects that indeed show this analogy within a class and which properties that do not. In chemistry the use of analogy models has both theoretical and practical importance[8–11] and can be seen as early chemical applications of *PaRC*.[12]

Third and perhaps most important, in chemical and biological applications, one is often interested in an empirical description of the data within a single class—obtaining an empirical model of the class This model can then be used for purposes of interpretation and of prediction, for instance for the purpose of constructing objects with desired properties

In fact, it is possible to derive rather general models for the behaviour of the measurements on similar objects, i.e. objects in a single class. These models can then be used for solving the ordinary classification problem and also for handling the other problems discussed above. The derivation of these models which is shown later in this paper is based on simple Taylor expansions. The form of the resulting models is that of principal components (PC) models. Provided that certain continuity assumptions are fulfilled for the processes underlying the data, à PC models can describe the variation of the measurements made on the objects in a single class The total model for a number of classes consequently consists of a collection of disjoint PC models, one model for each class

Disjoint PC models have been used in *PaRC* applications by Fukunaga,[13,14] Watanabe[15,16] and others[17,18] in terms of Karhunen–Loève expansions as PC models often are called in the *PaRC* literature Fukunaga *et al* point out that when the single purpose of the data analysis is classification, one might wish to abandon the model fitting approach and instead use combinations of eigenvectors as to maximise the discrimination between the classes This line has been followed up by Foley and Sammon[18] who show how to construct optimal—for the single purpose of classification—combinations of eigenvectors

Hence, the modelling approach to PaRC is not the most efficient approach to classification Some of the efficiency is sacrificed for the advantage to get an empirical model for each class and also to guard against overseparation of classes For use in modelling, PC models have the particular advantage to be able to approximate any continuous behaviour within the classes Furthermore, they have a simple representation in the measurement space—lines or hyperplanes—which make them easy to map and visualize

The purpose of the present paper is to present the use of PC models in *PaRC* with the emphasis on applications in chemistry and biology Therefore, the presentation is given in some detail in simple model fitting terms [19–21] As an illustration, the methods are applied to the classical Iris data of Fisher [22]

2. A FRAMEWORK OF PATTERN RECOGNITION IN TERMS OF CLASS MODELS

The essence of the present approach to *PaRC* is to recognize the fact that the objects in a single separate class are, by definition, in some way similar On the basis of this similarity, a mathematical model is formulated which, under rather general assumptions, describes the behaviour of the objects in a single class The total mathematical model consequently, consists of a collection of disjoint models, one for each class By means of data observed on objects with "known" classifications (these objects constitute the reference sets), the parameters in the separate similarity models are estimated (given numerical values)

Unclassified objects are then fitted to all the parametrized class models and classified according to which model they fit best The possibility that an unclassified object might be of a new kind, not fitting any of the previously known class models, should be kept in mind.

Hence, the *PaRC* framework consists of two parts, the data and the similarity models which are "calibrated" by means of these data

2.1 *Data*

The data consist of measured values (index **i**) made on a number of objects (index **k**) As an example, the classical data of Fisher[22] will be used The objects are 150 individuals of Iris The measurements are (1) sepal length, (2) sepal width, (3) petal length and (4) petal width Further, the objects are assumed

Table 1 Training set of the Iris data of Fisher[22] The 6 left columns are number, class, sepal length and width and petal length and width Then follows the results of a classification with $A = 2$ in terms of class number, residual standard deviation and "misfit number" (see text) for the closest class next closest and furthest class Finally the four rightmost columns summarize the classification by the present method with $A = 2$ and $A = 1$ a 3-nearest neighbour and a linear discriminant analysis Blank or dash means correct classification, a number indicates to which class the object was erroneously assigned by the corresponding method

1	1	51	35	14	02	1 = 0 0624	0	2 = 1 7228	0	3 = 2 0476	1	-	- - -
2	1	49	30	14	02	1 = 0 1716	0	2 = 1 4843	1	3 = 1 8613	1		
3	1	47	32	13	02	1 = 0 0441	0	2 = 1 5557	1	3 = 1 8757	1		
4	1	46	31	15	02	1 = 0 0438	0	2 = 1 3963	1	3 = 1 7381	1		

Table 1—*continued*

5	1	50	36	14	02	1 = 0 0526	0	2 = 1 7346	1	3 = 2 0363	1	
6	1	54	39	17	04	1 = 0 1326	0	2 = 1 7568	0	3 = 2 0477	1	
7	1	46	34	14	03	1 = 0 1449	0	2 = 1 5322	1	3 = 1 8088	1	
8	1	50	34	15	02	1 = 0 0307	0	2 = 1 6113	1	3 = 1 9488	1	
9	1	44	29	14	02	1 = 0 0360	0	2 = 1 3175	1	3 = 1 6596	1	
10	1	49	31	15	01	1 = 0 1202	0	2 = 1 4983	1	3 = 1 8828	1	
11	1	54	37	15	02	1 = 0 0881	0	2 = 1 8268	0	3 = 2 1606	1	- - - -
12	1	48	34	16	02	1 = 0 1260	0	2 = 1 5138	1	3 = 1 8421	1	
13	1	48	30	14	01	1 = 0 1458	0	2 = 1 4825	1	3 = 1 8639	1	
14	1	43	30	11	01	1 = 0 0493	0	2 = 1 4985	1	3 = 1 8078	1	
15	1	58	40	12	02	1 = 0 2495	0	2 = 2 1906	0	3 = 2 4975	1	
16	1	57	44	15	04	1 = 0 1148	0	2 = 2 1194	0	3 = 2 3513	12	
17	1	54	39	13	04	1 = 0 0876	0	2 = 1 9562	0	3 = 2 2109	1	
18	1	51	35	14	03	1 = 0 0588	0	2 = 1 6999	0	3 = 2 0084	1	
19	1	57	38	17	03	1 = 0 0605	0	2 = 1 8208	0	3 = 2 1707	1	
20	1	51	38	15	03	1 = 0 1214	0	2 = 1 7627	0	3 = 2 0391	1	
21	1	54	34	17	02	1 = 0 1109	0	2 = 1 6177	0	3 = 2 0100	1	- - - -
22	1	51	37	15	04	1 = 0 1232	0	2 = 1 7024	0	3 = 1 9747	1	
23	1	46	36	10	02	1 = 0 0392	0	2 = 1 8256	1	3 = 2 0549	1	
24	1	51	33	17	05	1 = 0 1827	0	2 = 1 4348	0	3 = 1 7634	1	
25	1	48	34	19	02	1 = 0 2536	0	2 = 1 3783	0	3 = 1 7327	1	
51	2	70	32	47	14	2 = 0 2215	0	1 = 0 7372	12	3 = 0 8518	0	
52	2	64	32	45	15	2 = 0 1415	0	3 = 0 6758	0	1 = 0 9258	2	
53	2	69	31	49	15	2 = 0 0293	0	3 = 0 6712	0	1 = 0 8469	2	
54	2	55	23	40	13	2 = 0 1659	0	3 = 0 4318	0	1 = 0 7957	2	
55	2	65	28	46	15	2 = 0 0884	0	3 = 0 5899	0	1 = 0 8460	2	
56	2	57	28	45	13	2 = 0 2133	0	3 = 0 4515	0	1 = 0 9544	2	
57	2	63	33	47	16	2 = 0 0228	0	3 = 0 5477	0	1 = 1 0824	2	
58	2	49	24	33	10	2 = 0 1517	0	3 = 0 6389	0	1 = 0 6587	2	
59	2	66	29	46	13	2 = 0 0855	0	1 = 0 7155	2	3 = 0 7164	0	
60	2	52	27	39	14	2 = 0 0751	0	3 = 0 4177	0	1 = 0 9911	2	
61	2	50	20	35	10	2 = 0 0789	0	3 = 0 5001	0	1 = 0 6058	2	
62	2	59	30	42	15	2 = 0 1141	0	3 = 0 5724	0	1 = 0 9668	2	
63	2	60	22	40	10	2 = 0 0386	0	1 = 0 4823	2	3 = 0 7084	0	
64	2	61	29	47	14	2 = 0 1682	0	3 = 0 4714	0	1 = 0 9489	2	
65	2	56	29	36	13	2 = 0 3329	0	3 = 0 7662	0	1 = 0 7919	2	
66	2	67	31	44	14	2 = 0 2631	0	1 = 0 7440	2	3 = 0 8474	0	
67	2	56	30	45	15	2 = 0 1688	0	3 = 0 3877	0	1 = 1 1380	2	
68	2	58	27	41	10	2 = 0 1992	0	1 = 0 6367	2	3 = 0 7160	0	
69	2	62	22	45	15	2 = 0 3404	0	3 = 0 4365	0	1 = 0 8303	2	
70	2	56	25	39	11	2 = 0 0673	0	3 = 0 6284	0	1 = 0 6621	2	
71	2	59	32	48	18	2 = 0 2306	0	3 = 0 2795	0	1 = 1 3191	2	- 3 - -
72	2	61	28	40	13	2 = 0 2195	0	1 = 0 7195	2	3 = 0 7645	0	
73	2	63	25	49	15	3 = 0 3181	0	2 = 0 3669	0	1 = 0 9060	2	3 - 3 -
74	2	61	28	47	12	2 = 0 2342	0	3 = 0 5430	0	1 = 0 8244	2	
75	2	64	29	43	13	2 = 0 1713	0	1 = 0 7109	2	3 = 0 7689	0	
101	3	63	33	60	25	3 = 0 3673	0	2 = 0 8364	0	1 = 1 8864	2	
102	3	58	27	51	19	3 = 0 0805	0	2 = 0 6021	0	1 = 1 3654	2	
103	3	71	30	59	21	3 = 0 1127	0	2 = 0 5932	1	1 = 1 3310	12	
104	3	63	29	56	18	3 = 0 1826	0	2 = 0 6226	0	1 = 1 3253	2	
105	3	65	30	58	22	3 = 0 1450	0	2 = 0 7226	0	1 = 1 5449	2	
106	3	76	30	66	21	3 = 0 0313	0	2 = 0 7993	1	1 = 1 3402	12	
107	3	49	25	45	17	3 = 0 1971	0	2 = 0 5677	0	1 = 1 3501	2	- - 2 -
108	3	73	29	63	18	3 = 0 1419	0	2 = 0 7165	1	1 = 1 1787	12	
109	3	67	25	58	18	3 = 0 0265	0	2 = 0 7591	0	1 = 1 1564	2	
110	3	72	36	61	25	3 = 0 0587	0	2 = 0 5658	1	1 = 1 6941	12	
111	3	65	32	51	20	3 = 0 2912	0	2 = 0 3023	0	1 = 1 3136	2	
112	3	64	27	53	19	3 = 0 1030	0	2 = 0 5483	0	1 = 1 2316	2	
113	3	68	30	55	21	3 = 0 1865	0	2 = 0 4994	0	1 = 1 3325	2	
114	3	57	25	50	20	3 = 0 1512	0	2 = 0 6948	0	1 = 1 3890	2	
115	3	58	28	51	24	3 = 0 2837	0	2 = 0 7652	0	1 = 1 6714	2	
116	3	64	32	53	23	3 = 0 1152	0	2 = 0 5214	0	1 = 1 5585	2	
117	3	65	30	55	18	3 = 0 1863	0	2 = 0 4792	0	1 = 1 2567	2	
118	3	77	38	67	22	3 = 0 2323	0	2 = 0 5571	1	1 = 1 5792	12	
119	3	77	26	69	23	3 = 0 3238	0	2 = 1 1153	1	1 = 1 4116	12	
120	3	60	22	50	15	3 = 0 1031	0	2 = 0 5922	0	1 = 0 9559	2	- - 2 -
121	3	69	32	57	23	3 = 0 1369	0	2 = 0 5571	0	1 = 1 4998	12	
122	3	56	28	49	20	3 = 0 0708	0	2 = 0 5582	0	1 = 1 4562	2	
123	3	77	28	67	20	3 = 0 0419	0	2 = 0 8775	1	1 = 1 2350	12	
124	3	63	27	49	18	3 = 0 2835	0	2 = 0 3762	0	1 = 1 1202	2	- - 2 -
125	3	67	33	57	21	3 = 0 1085	0	2 = 0 4833	0	1 = 1 4705	12	

Table 2 Test set of the Iris data Notation same as in table 1 The correct classification is, according to Fisher,[22] class 1 for objects 26–50, class 2 for objects 76–100 and class 3 for objects 126–150

26	0	50	30	16	02	1 = 0 1250	0	2 = 1 4127	1	3 = 1 8152	1	
27	0	50	34	16	04	1 = 0 1151	0	2 = 1 5156	0	3 = 1 8290	1	
28	0	52	35	15	02	1 = 0 0661	0	2 = 1 7002	0	3 = 2 0428	1	– – – –
29	0	52	34	14	02	1 = 0 1340	0	2 = 1 7115	0	3 = 2 0605	1	
30	0	47	32	16	02	1 = 0 0804	0	2 = 1 4125	1	3 = 1 7583	1	
31	0	48	31	16	02	1 = 0 0050	0	2 = 1 3995	1	3 = 1 7678	1	
32	0	54	34	15	04	1 = 0 1901	0	2 = 1 6730	0	3 = 2 0188	1	
33	0	52	41	15	01	1 = 0 2419	0	2 = 1 9549	0	3 = 2 2334	1	
34	0	55	42	14	02	1 = 0 1181	0	2 = 2.0886	0	3 = 2 3543	12	
35	0	49	31	1.5	0.2	1 = 0 0840	0	2 = 1 4728	1	3 = 1 8429	1	– – – –
36	0	50	32	12	02	1 = 0 2077	0	2 = 1 6840	1	3 = 2 0261	1	
37	0	55	35	13	02	1 = 0 2602	0	2 = 1 8770	0	3 = 2 2353	1	
38	0	49	36	14	01	1 = 0 1258	0	2 = 1 7356	1	3 = 2 0441	1	
39	0	44	30	13	02	1 = 0 0322	0	2 = 1 4032	1	3 = 1 7230	1	
40	0	51	34	15	02	1 = 0 0579	0	2 = 1 6369	0	3 = 1 9840	1	
41	0	50	35	13	03	1 = 0 0614	0	2 = 1 7233	1	3 = 2 0132	1	
42	0	45	23	13	03	1 = 0 3456	0	2 = 1 1538	1	3 = 1 5717	1	
43	0	44	32	13	02	1 = 0 0853	0	2 = 1 4785	1	3 = 1 7719	1	
44	0	50	35	1.6	06	1 = 0 2569	0	2 = 1 5128	0	3 = 1 7748	1	
45	0	51	38	19	04	1 = 0 2893	0	2 = 1 5465	0	3 = 1 8445	1	
46	0	48	30	14	03	1 = 0 1438	0	2 = 1 4363	1	3 = 1 7870	1	– – – –
47	0	51	38	16	02	1 = 0 1657	0	2 = 1 7411	0	3 = 2 0430	1	
48	0	46	32	14	02	1 = 0 0409	0	2 = 1 4817	1	3 = 1 8014	1	
49	0	53	37	15	02	1 = 0 0626	0	2 = 1 8011	0	3 = 2.1253	1	
50	0	50	33	14	02	1 = 0 0882	0	2 = 1 6221	1	3 = 1 9655	1	
76	0	66	30	44	14	2 = 0 1914	0	1 = 0 7542	2	3 = 0.7887	0	
77	0	68	28	48	14	2 = 0 0378	0	3 = 0 6505	0	1 = 0 7430	2	
78	0	67	30	50	17	2 = 0 1729	0	3 = 0 4619	0	1 = 1 0228	2	– – 3 –
79	0	60	29	45	15	2 = 0 0868	0	3 = 0 4643	0	1 = 0 9845	2	
80	0	57	2.6	35	10	2 = 0 3310	0	1 = 0 5200	2	3 = 0 8913	0	
81	0	55	24	38	11	2 = 0 0362	0	3 = 0 6099	0	1 = 0 6525	2	
82	0	55	24	37	10	2 = 0 1141	0	1 = 0 5731	2	3 = 0 6896	0	
83	0	58	27	39	12	2 = 0 1527	0	1 = 0 7016	2	3 = 0 7073	0	
84	0	60	27	51	16	3 = 0 1887	0	2 = 0 4833	0	1 = 1 1415	2	3 3 3 3
85	0	54	30	45	15	2 = 0 2213	0	3 = 0 3632	0	1 = 1 2071	2	– 3 – –
86	0	60	34	45	16	2 = 0 0787	0	3 = 0 5539	0	1 = 1 1532	2	
87	0	67	31	47	15	2 = 0 0806	0	3 = 0 6813	0	1 = 0 8619	2	
88	0	63	23	44	13	2 = 0 1457	0	3 = 0 5768	0	1 = 0 6752	2	
89	0	56	30	41	13	2 = 0 1044	0	3 = 0 5998	0	1 = 0 9293	2	
90	0	55	25	40	13	2 = 0 0798	0	3 = 0 4733	0	1 = 0 8284	2	
91	0	55	26	44	12	2 = 0 2668	0	3 = 0 4249	0	1 = 0 8948	2	
92	0	61	30	46	14	2 = 0 0870	0	3 = 0 5321	0	1 = 0 9468	2	
93	0	58	26	40	12	2 = 0 0653	0	3 = 0 6434	0	1 = 0 7039	2	
94	0	50	23	33	10	2 = 0 1355	0	1 = 0 6148	2	3 = 0 6499	0	
95	0	56	2.7	42	13	2 = 0 0826	0	3 = 0 4820	0	1 = 0 8841	2	
96	0	57	30	42	12	2 = 0 1576	0	3 = 0 6460	0	1 = 0 8741	2	
97	0	57	29	42	13	2 = 0 0628	0	3 = 0 5674	0	1 = 0 8982	2	
98	0	62	29	43	13	2 = 0 1119	0	3 = 0 6945	0	1 = 0 7667	2	
99	0	51	25	30	11	2 = 0 3929	0	1 = 0 6380	2	3 = 0 8251	0	
100	0	57	28	41	13	2 = 0 0430	0	3 = 0 5758	0	1 = 0 8507	2	
126	0	72	32	60	18	3 = 0 2657	0	2 = 0 4819	1	1 = 1 2056	12	– – – 2
127	0	62	28	48	18	3 = 0 2883	0	2 = 0 3227	0	1 = 1 1448	2	– – 2 –
128	0	61	30	49	18	3 = 0 2436	0	2 = 0 2967	0	1 = 1 2349	2	– – 2 –
129	0	64	28	56	21	3 = 0 1008	0	2 = 0 7044	0	1 = 1 4312	2	
130	0	72	30	58	16	3 = 0 3608	0	2 = 0 4260	0	1 = 1 0018	12	– – – 2
131	0	74	28	61	19	3 = 0 1492	0	2 = 0 6337	1	1 = 1 1332	12	
132	0	79	38	64	20	2 = 0 3199	1	3 = 0 4032	0	1 = 1 3289	12	2 2 – 2
133	0	64	28	56	22	3 = 0 1536	0	2 = 0 7389	0	1 = 1 4883	2	
134	0	63	28	51	15	3 = 0 3254	0	2 = 0 3555	0	1 = 1 0152	2	– 2 – 2
135	0	61	26	56	14	3 = 0 3370	0	2 = 0 7605	0	1 = 1 1279	2	
136	0	77	30	61	23	3 = 0 3119	0	2 = 0 6230	1	1 = 1 3290	12	
137	0	63	34	56	24	3 = 0 1195	0	2 = 0 6001	0	1 = 1 7559	2	
138	0	64	31	55	18	3 = 0 2111	0	2 = 0 4720	0	1 = 1 3146	2	
139	0	60	30	48	18	3 = 0 2480	0	2 = 0 2795	0	1 = 1 2421	2	– – 2 –
140	0	69	31	54	21	3 = 0 2795	0	2 = 0 4114	0	1 = 1 3027	12	
141	0	67	31	56	24	3 = 0 1719	0	2 = 0 6467	0	1 = 1 5759	2	
142	0	69	31	51	23	3 = 0 4460	0	2 = 0 4916	0	1 = 1 3759	12	
143	0	58	27	51	19	3 = 0 0772	0	2 = 0 6021	0	1 = 1 3654	2	
144	0	68	32	59	23	3 = 0 0665	0	2 = 0 6515	0	1 = 1 5741	12	
145	0	67	33	57	25	3 = 0 1434	0	2 = 0 6455	0	1 = 1 6923	12	

Table 2—*continued*

146 0	6 7	3 0	5 2	2 3	3 = 0 3172	0	2 = 0 5422	0	1 = 1 4255	2
147 0	6 3	2 5	5 0	1 9	3 = 0 2218	0	2 = 0 5334	0	1 = 1 1690	2
148 0	6 5	3 0	5 2	2 0	3 = 0 2041	0	2 = 0 4096	0	1 = 1 2966	2
149 0	6 2	3 4	5 4	2 3	3 = 0 0549	0	2 = 0 5063	0	1 = 1 6845	2
150 0	5 9	3 0	5 1	1 8	3 = 0 1786	0	2 = 0 4373	0	1 = 1 3497	2

to belong to either of a number of given classes These classes are usually defined by means of objects with "known" classification These objects constitute the reference sets, sometimes also called the training sets (one set for each class) The number of classes in the example are three, the Iris is considered to belong to either of the species *Iris setosa* (1), *Iris versicolor* (2) or *Iris virginica* (3) In the present illustration, these Irises are divided into two parts, the first 25 individuals in each class are taken as reference sets and the latter 25 as test sets (assumed to be of unknown classification) See further Tables 1 and 2

The data are denoted by y_{ik} and together form the observation matrix **Y** with dimensions **M** × **N** (Fig 2) In the present article, it will be assumed that the matrix **Y** is complete, i e all the **M** variables have been measured for all the **N** objects This is no necessary assumption, however, the models work also when data are missing, see further Section 3

The observations made on a single object form an **M**-dimensional vector and can consequently be represented as a point in an **M**-dimensional space, here called the measurement space

2 1 1 Transformation of data Ideally, the variables should be weighted according to their relevance for the particular classification problem However, prior information regarding this relevance is seldom available Instead, it is customary to transform the variables to give them equal weight (equal variance), so called autoscaling This can be done on the basis of the reference sets or using all available data In the present example, the variance of the four variables is similar however, and no scaling of the data has been made

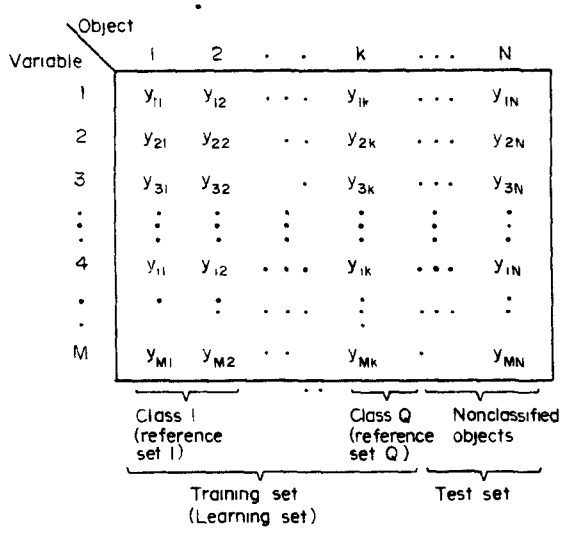

Fig 2 The available data in the pattern recognition problem form a matrix of dimensions M times N

If the distribution of the values of a variable is very skew, e g so that most measurements are rather small but a few are very large, it might be practical to correct for this by taking the logarithm or the square root of the observed values or to use other special transformations [23] In the present example, no such transformation has been made

2.2 Similarity models

The present treatment defines one separate model for each separate class Thus, let us consider one single class of **n** objects, which, by definition, are in some way similar On each object, the values of **M** variables have been measured giving the data matrix **Y** with the elements y_{ik} (see Section 2 1) If all the objects in the class are *identical*, the values of one variable **i** are the same for all the objects apart from small deviations ϵ_{ik} due to errors of measurement Hence, for this simple case, the data within one class can be described by the model

$$y_{ik} = \alpha_i + \epsilon_{ik} \tag{1}$$

However, equation (1) often is unrealistically simple The fundamental assumption that the objects within one class are so similar that they are virtually identical is rarely fulfilled in real situations If we instead assume that the objects differ slightly from each other, the second simplest model is obtained (see Appendix and Ref 11 for derivations)

$$y_{ik} = \alpha_i + \beta_i\theta_k + \epsilon_{ik} \tag{2}$$

Finally, a larger variation between the objects in the class, leads to the similarity model (see Appendix)

$$y_{ik} = \alpha_i + \sum_{a=1}^{A} \beta_{ia}\theta_{ak} + \epsilon_{ik} \tag{3}$$

It is seen that the three models (1–3) all are principal components models with the number of components being zero, one and **A** respectively

If now the objects come from a number of classes (with class index **q**), the data can consequently be described by a number of disjoint models

$$y_{ik}^{(q)} = \alpha_i^{(q)} + \sum_{a=1}^{A_q} \beta_{ia}^{(q)}\theta_{ak}^{(q)} + \epsilon_{ik}^{(q)}, \tag{4}$$

with A_q being zero, one, or larger corresponding to the single class models (1), (2) and (3) respectively Let us further adopt a limiting residual variance for each class, σ_q^2 The geometrical representation of the model (1) for a single class is then a hyper-sphere

with radius σ_q in measurement space (Fig. 3). This simple model is, in fact, the basis of many *PaRC* applications. the Euclidian distance between two points is a measure of the dissimilarity between the corresponding two objects. The geometrical representation of model (2) in measurement space is a cylinder with radius σ_q (Fig 4) and the model (3) is represented by higher volumes

The model (4) has been derived using only assumptions of continuity of the measurement variation between objects and variables. Thus, provided that these assumptions are fulfilled, the data within a single class can be described by model (3) *regardless of their structure*, if sufficiently many terms (A) are included. Hence, model (4) is, in turn, a complete description of the data in the Q classes provided only that the dimensions of the data-matrices of the reference sets are sufficiently large to allow estimation of the A_q product terms for each class.

These properties make the model (4) an ideal model for use in *PaRC* applications. In order to get a first crude working method, very little needs to be known about the data structure within each class and nothing needs to be known about the variation between the classes In order to reach near-optimal efficiency, however, one usually needs to verify such things as unimodality within classes and study the relation between sample size and classification performance, see discussions by Foley[18] and Fukunaga (Ref 13, Ch 5)

Model (4) gives, after the parameters are estimated from the reference sets, a parametrized structure of each class in terms of the values of the parameters α, β and θ These values can then, in addition to the ordinary use for classification purposes, be used for discussion and interpretation of the class structures, differences in structure between classes and so on

2.2.1 Number of terms (A) in the model The similarity models (4) are fitted separately to the reference set of each class Before doing this, one must in some way determine the dimensionality of the data in each class In the present case, this dimensionality is measured by the number of product terms, A_q in Equation (4) For a more general discussion of the problem, the reader is referred to Fukunaga (Ref 13, Ch 8 and 10) and Kanal (Ref 1, Section VI).

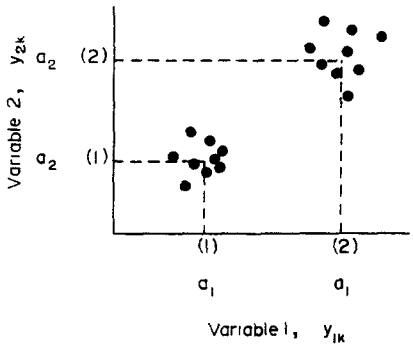

Fig 3 Two classes which are each described by equation (1) in a two-dimensional measurement space

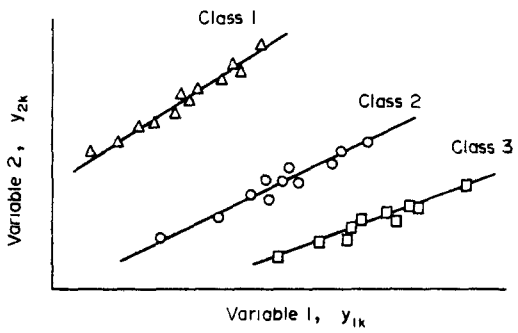

Fig 4 Three classes which are all described by equation (2) in a two-dimensional measurement space To be included in a class, an object should lie in a band with width $2\sigma_q$ around the line representing the class model

An approach that has worked well in practice with the present models is to use cross validation techniques[24,25] For each separate class this is done as follows

(a) Divide the objects in the class into T groups (index t) with T being *ca.* 5–10 Each group shall be as representative for the whole class as possible In the present example with 25 objects in each reference set (Table 1), 8 groups were used with the first containing objects 1, 9, 17 and 25, the second objects 2, 10 and 18, the third objects 3, 11 and 19 and so on until the eighth containing objects 8, 16 and 24

(b) Form a reduced data matrix $\mathbf{Y}^-$ by deleting the objects in group 1 from the reference set The number of objects in $\mathbf{Y}^-$ is n^-

(c) Fit the similarity model (3) to $\mathbf{Y}^-$ with A being consecutively 0, 1, 2, up to $M-2$ or $n^- - 2$ whichever is smaller (see Section 2.2.2, below)

(d) Fit the objects in the deleted group to the similarity models obtained in step (c) with $A = 0$, 1, and **fixed** α- and β-parameters (see Section 2.2.4 below) Calculate the deviations ϵ_{ik} for each A-value for the deleted objects and form Δ_A as the sum of the squares of these deviations

(e) Put back the deleted group into the data matrix **Y**.

(f) Form a new reduced data matrix $\mathbf{Y}^-$ by deleting the next group of objects, go back to step **c** and then **d**, **e** and **f**. If all groups already have been deleted once, instead go on to step **g**

(g) For each value of A, form the sum D_A by adding the corresponding values of Δ_A These D-values are a measure of how well the model (3) predicts the behaviour of the reference-set for each value of A By making F-tests on $(D_{A-1} - D_A)/n$ vs $D_A/[n(M - A - 1)]$ one can determine whether the last product term (number A) was significant or not This technique is exactly analogous to that using F-tests to determine the significance of the last term in polynomial regression[26] Therefore, it also rests on the same assumptions about independence between the observations

Thus, the cross-validation technique determines the number of product terms A_q for each class so that the prediction ability of model (4) is maximized with respect to the reference sets Figure 5 shows the results of this technique for the Iris data It is seen that for all three classes inclusion of the first component results in significantly better fit, i e the data contain more structure than can be described by the simple distance related model (1) For class three, the inclusion of a second term gives a better fit which is on the border of significance Hence the Iris-data are adequately described by three disjoint one component models ($A = 1$ in equation 4) A second component ($A = 2$) might give better fit for class three

2 2 2 *Estimation of the parameter values in model (4)* Before the models (4) can be used, e g for the classification of a new object, values of the parameters $\alpha_i^{(q)}$, $\beta_{ia}^{(q)}$ and σ_q^2 for $q = 1, 2, \ldots Q$ (Q = number of classes), $i = 1, 2, \ldots M$ (M = number of variables), $a = 1, 2, \ldots A_q$ (A_q = number of product terms in model 4 for class q) and $k = 1, 2, \ldots n_q$ (n_q = number of objects in qth reference set) must be determined from the data in the reference-sets

This corresponds to the estimation of the principal components of each data matrix of the reference sets after subtracting the averages $\alpha_i^{(q)}$ This, in turn corresponds to a diagonalization of the matrices $Z^{(q)}Z^{(q)+}$ where $Z^{(q)}$ denotes the matrix obtained form the data matrix of the qth reference-set after subtracting the average of each variable $\alpha_i^{(q)}$

A number of practical methods are available [27–29] In the present applications, I have preferred the NIPALS method[29] which is an iterative method calculating one eigenvalue with the corresponding eigenvector at a time consecutively one after another (for numerical details, see Refs 29, 30) This method is advantageous in connection with the cross-validation techniques (Section 2 2 1) since it can utilize the values from previous $\mathbf{Y}^-$ matrices as starting values in the latest calculation, thereby converging very rapidly Hence, using any diagonalization method, values of the parameters β_{ia} and θ_{ak} are obtained for each class The deviations $\epsilon_{ik}^{(q)}$ are then calculated by subtracting the appropriate number (A_q) of product terms from the z-values and the variances σ_q^2 are then estimated from these deviations as $s_0^{(q)2}$

$$s_0^{(q)2} = \sum_k^{n_q} \sum_i^M (\epsilon_{ik}^{(q)})^2 / [(n_q - A_q - 1)(M - A_q)] \quad (5)$$

The first summation is made over the number of objects in the reference set (n_q) and the second over the number of variables (M)

Thus, for each class the similarity model (4) is "calibrated" by means of the data in the reference sets The calibrated models can then be used to determine the classification of new objects and also for other things as discussed below Table 3 gives the resulting parameters for the Iris The one component model projected onto the plane of variables 1 and 3 is also shown in Fig 6

Due to the eigenvector properties of the coefficients, they are easy to interpret The $\alpha_i^{(q)}$-values are simply the mean value of the ith variable for the qth class The β-vectors are orthogonal to each other and so are the θ-vectors within each class and all have the mean values zero Hence, the β-variables express the variation of the corresponding variables around the class averages In order to make the interpretation of the parameter values easier, it is sometimes, especially when many components are needed to describe the data within a class, useful to rotate the β- and θ-matrices (dimensions M A and A M respectively) by multiplication by a unitary matrix This is not done in the present example, for details the reader is referred to any standard text on factor analysis [31]

In the Iris example, where all the variables are lengths or widths, the first θ vector expresses the size of the corresponding individuals and the first β-vector the "rate" at which the corresponding variable changes when the size is changed within the class

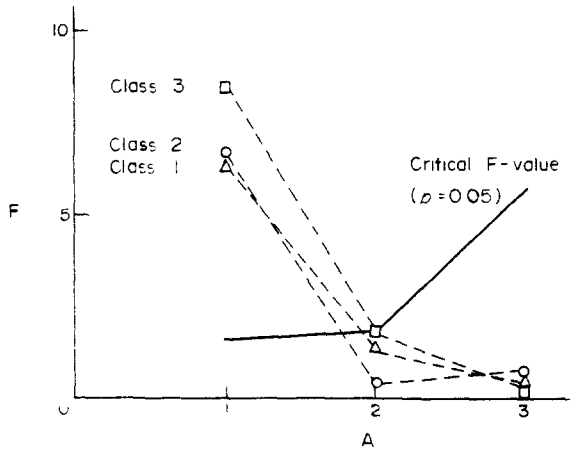

Fig 5 The results of a crossvalidatory determination of the number of product terms (A_q in equation 4) needed to adequately describe the three reference data matrices of the Iris-data The solid line shows the critical F-value ($p = 0.05$) for the corresponding number of degrees of freedom (n vs $n(M - A_q - 1)$)

Table 3(a) Parameters for the three Iris class models (see equation 4) Class 1 = *I setosa* class 2 = *I versicolor* class 3 = *I virginica* β-values normalised so that $\sum_{i=1}^{4} \beta_{ia}^2 = 1$

		$\alpha_i^{(q)}$	$\beta_{i1}^{(q)}$	$\beta_{i2}^{(q)}$
	$i = 1$	5 028	0 7365	0 3260
Class 1	2	3 480	0 6598	−0 4870
($q = 1$)	3	1 460	0 0969	0 8099
	4	0 248	0 1133	0 0242
	$i = 1$	6 012	0 7216	−0 5471
Class 2	2	2 776	0 3543	0 7480
($q = 2$)	3	4 312	0 5632	0 1083
	4	1 344	0 1912	0 3598
	$i = 1$	6 576	0 7301	−0 1642
Class 3	2	2 928	0 1948	0 8350
($q = 3$)	3	5 640	0 6451	−0 1548
	4	2 044	0 1137	0 5018

Table 3(b) Parameter values for components (θ_{ak}) for $a = 1,2$ for the objects in the three Iris reference sets Class numbering same as in Table 3(a) object numbering (k) same as in Table 1

k	$\theta_{1k}^{(1)}$	$\theta_{2k}^{(1)}$	k	$\theta_{1k}^{(2)}$	$\theta_{2k}^{(2)}$	k	$\theta_{1k}^{(3)}$	$\theta_{2k}^{(3)}$
1	0 0550	−0 0360	51	1 092	−0 1613	101	0 1550	0 5290
2	−0 4222	0 1423	52	0 5659	0 1813	102	−0 9757	−0 0517
3	−0 4473	−0 1013	53	1 117	−0 1237	103	0 5707	−0 0381
4	−0 5675	0 0768	54	−0 7222	−0 1255	104	−0 2605	−0 0943
5	0 0473	−0 1173	55	0 5527	−0 1617	105	0 0795	0 1261
6	0 5916	0 1148	56	−0 1192	0 1932	106	1 387	−0 2285
7	−0 3679	−0 1479	57	0 6610	0 3685	107	−2 081	−0 0784
8	−0 0750	0 0611	58	−1 571	0 0938	108	0 9211	−0 3669
9	−0 8565	0 0280	59	0 6220	−0 2136	109	0 0826	−0 5250
10	−0.3579	0 1722	60	−0 8342	0 3630	110	0 9350	0 6163
11	0 4176	0 0453	61	−1 528	−0 2384	111	−0 3558	0 3011
12	−0 2126	0 0769	62	−0 0347	0 2728	112	−0 4086	−0 1811
13	−0 5072	0 1073	63	−0 4542	−0 5818	113	0 0936	0 0731
14	−0 9045	−0 2987	64	0 3367	0 1068	114	−1 141	−0 1366
15	0 8810	−0 2133	65	−0 6628	0 2252	115	−0 8993	0 2828
16	1 123	−0 1929	66	0 6715	−0 1044	116	−0 2657	0 4371
17	0 5528	−0 2092	67	−0 0822	0 4695	117	−0 1595	−0.0282
18	0 0663	−0 0336	68	−0 3651	−0.0876	118	1 692	0 4578
19	0 7352	0 2588	69	0 0673	−0 4572	119	1 599	−0 5250
20	0 2739	−0.0987	70	−0.6738	−0 1134	120	−1 037	−0 6873
21	0 2390	0.3534	71	0 4315	0 5953	121	0 3573	0 2931
22	0 2193	−0 0476	72	−0 1121	−0 0798	122	−1 220	0 1458
23	−0 2861	−0 5717	73	0 4711	−0 2442	123	1 474	−0 4776
24	−0 0139	0 3116	74	0.2630	−0 0400	124	−0 7510	−0 1530
25	−0 1835	0 3199	75	0 3087	−0 1367	125	0 2081	0 3091
S D	0 522	0 213		0 717	0 282		0 974	0 353
Range	1 123	0 3534		1 117	0 5953		1 692	0 6163
	−	−		−	−		−	−
	−0 9045	−0 5717		−1 571	−0 5818		−2 081	−0 6873

One can see [Table 3(b), bottom] that the variation in size is almost twice as large in class three compared with class 1. In class three, the doubling in size doubles the length of both the sepal ($\beta_{11}^{(3)}$) and the petal ($\beta_{31}^{(3)}$) while in class one, the petal length and width is almost size independent ($\beta_{31}^{(1)}$ and $\beta_{41}^{(1)}$) Since this is not the place to make a detailed interpretation of Fisher's data, I will not discuss these results any further; the given examples are sufficient to illustrate how the parameters within one class give a quantitative picture of the "class structure"

2 2 3 The distribution of the values of θ within one reference-set The fitting of the model (4) to the data matrix of the qth reference set (class q) gives, for each object in the set, values of the parameters $\theta_{ak}^{(q)}$ If one wishes, these values can be used to determine a region for each $\theta_a^{(q)}$, in order for a non-classified object to be considered as a member of the class, a small residual variance (of the same order as that of the class, equation 5) should be obtained with parameters within the acceptable regions However, the classification of a new object becomes rather complicated in this way and since, in addition, this procedure is unnecessary in most applications, I recommend a simple control of the parameter values (c_a) after the fitting according to 2 2 4

In the fitting of the non-classified Iris (see below) to the three class models, a parameter value (c_a) falling outside the corresponding range plus or minus the standard deviation [bottom of Table 3(b)] has been flagged by a "misfit" indicator (see Table 2) When, for instance, this "misfit" criterion for object 34 has the value 12 for class 3 (see Table 2), this means that both c_1 and c_2 falls outside the interval

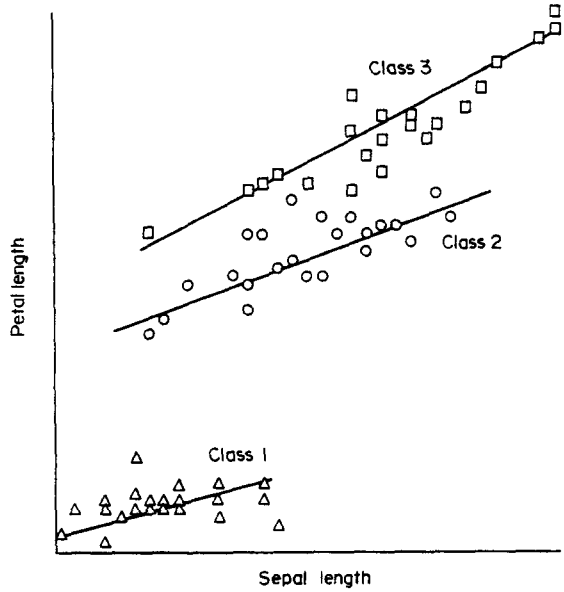

Fig 6 The data of the three Iris reference sets together with the resulting class models (solid lines) when $A_q = 1$ in equation (4), projected on the plane of the variables one (sepal length) and three (petal length)

(range ± S D for the corresponding parameters in class 3

2.2.4 Fitting of a non-classified object to the model (4) The similarity model, including values of the parameters is as shown above, completely defined by the data matrix of the reference set of the class When, later, one wishes to try how well a specified object (with index **p**) fits this class, one proceeds as follows

(a) Fit the observations of the object, say y_{ip}, to model (4) with the same number of product terms and the same values of the parameters $\alpha_i^{(q)}$ and $\beta_i^{(q)}$ as were obtained in the "calibration" of the model in Section 2.2.2 This fitting corresponds to a simple linear multiple regression with $z_i = y_{ip} - \alpha_i^{(q)}$ as dependent variable and $\beta_{ia}^{(q)}$, $a = 1, 2, \ldots A_q$ as independent variables

$$y_{ip} - \alpha_i^{(q)} = z_i = \sum_{a=1}^{A_q} c_a \beta_{ia}^{(q)} + \epsilon_i^{(q)} \qquad (6)$$

(b) The variance of the deviations (ϵ) then indicates how well object **p** fits class **q**

$$s_p^{(q)2} = \sum_{i=1}^{M} (\epsilon_i^{(q)})^2/(M - A_q) \qquad (7)$$

If $s_p^{(q)2}$ is "much" larger (F-test) than the standard deviation of the reference set (equation 5), then the object **p** does not belong to class **q**. It is seen that it might well happen that an object is classified as belonging to several classes, in such a case the observed data are not sufficient to uniquely determine the class of the object It might also happen that an object is classified as not belonging to any of the classes, it is an object of a new kind

In the Iris example, 25 individuals in each class were saved to constitute a "test"-population The residual variances (equation 7) for each of these 75 individuals are shown in Table 2 together with the resulting classification and a comparison with the corresponding results of a linear discriminant analysis[32] and a K- nearest neighbour analysis[33] It is seen that the present method compares very favourably with the other methods, 71 or 73 of the 75 objects in the test-population are correctly classified by the present method depending on whether one or two components were used in the analysis, vs 70 by the other two methods If all 150 objects are classified, the present method classifies 145 and 147 correctly with one and two components respectively, vs 145 and 141 for the other two methods

2.3 Similarity–dissimilarity between two classes

By fitting all objects in the reference-set **r** to the calibrated model for class **q** and vice versa, one can get a measure of the similarity between these two classes The variances according to equation (8) can be compared with the variances according to equation (5) to give a quantitative comparison

$$s^{(r,q)2} = \sum_{k=1}^{n_r} \sum_{i=1}^{M} (\epsilon_{ik}^{(q)})^2/[n_r(M - A_q)] \qquad (8)$$

The deviations (ϵ_{ik}) in equation (8) are calculated by fitting the objects in reference-set **r** to the similarity model for class **q** The variance $s^{(q,r)2}$ is obtained analogously by changing the index **r** to **q** and vice versa

The variances obtained from the Iris data are shown in Table 4 It is seen that both for $A = 1$ and $A = 2$, the three classes are well separated The closest distances in both cases are those between classes 2 and 3 (matrix elements 2,3 and 3,2) The inclusion of the second component makes these distances between classes 2 and 3 slightly larger

2.4 The relevance of a variable

If the number of variables (M) is three or larger one can obtain a measure of the explanatory power of each variable by comparing the size of the variances $s_{y,i}^2$ and $s_{\epsilon,i}^2$ below, i.e. for each variable the residual variance is compared with the variance of the original data y_{ik} If the data have been auto-scaled, the latter variance (equation 9) is the same for all variables **i**

$$s_{y,i}^2 = \sum_{q=1}^{Q} \sum_{k=1}^{n_q} (y_{ik}^{(q)} - \bar{\bar{y}}_i)^2 \Big/ \left[\left(\sum_q n_q\right) - 1\right], \qquad (9)$$

$$\bar{\bar{y}}_i = \sum_{q=1}^{Q} \sum_{k=1}^{n_q} y_{ik}^{(q)} \Big/ \sum_q n_q, \qquad (10)$$

$$s_{\epsilon,i}^2 = \sum_{q=1}^{Q} \sum_{k=1}^{n_q} (\epsilon_{ik}^{(q)})^2 \Big/ \sum_q (n_q - A_q - 1) \qquad (11)$$

The ratio between the residual and data variances for a variable **i** (equation 12) is smaller the larger explanatory power the variable **i** has Hence, $1 - U_i$ is a measure of the relevance of the ith variable, the closer to one, the larger relevance and the closer to zero, the smaller relevance

$$U_i = s_{\epsilon,i}^2/s_{y,i}^2 \qquad (12)$$

Table 5 shows the residual and data variances for each variable in the Iris example calculated on the reference sets It is seen that for one component ($A = 1$) variables one and three are more "relevant" than variables 2 and 4

Another way to measure the relevance of a variable is to study its discriminatory power One can, for a given variable **i**, compare (1) the variance of the residuals when all objects in the reference sets are fitted to all classes but "their own" with (2) the variance of the residuals when the same objects are fitted to

Table 4 "Distances" between the three Iris classes expressed as the S D of the objects in reference set **r** when these are fitted to the class model **q**

	One component ($A = 1$)			Two components ($A = 2$)		
	$q = 1$	$q = 2$	$q = 3$	$q = 1$	$q = 2$	$q = 3$
$r = 1$	0 192	1 38	1 65	0 173	1 67	1 99
2	1 88	0 275	0 535	0 853	0 273	0 609
3	2 71	0 578	0 307	1 40	0 650	0 259

Table 5 Measures of variable relevance for the Iris-data The first row ($1 - U_i$) indicates the importance of the variables within the classes (see equation 12) The second row, $s^2_{not-class}/s^2_{in-class}$, is a measure of how well the variables participate in the discrimination between the classes (see equation 13)

	$i = 1$	1	2	3	4
$1 - U_i$	$A = 1$	0 929	0 524	0 845	0 232
	2	0 963	0 914	0 896	0 596
s^2_{nc}/s^2_{ic}	$A = 1$	26 2	24 5	169 8	41 2
	2	63 8	56 7	55 2	84 5

"their own" classes The ratio between these two variances will give an indication of how much the corresponding variable discriminates between "correct" and "incorrect" class

$$s^2_{not\text{-}class,i}/s^2_{in\text{-}class,i} = \sum_{q=1}^{Q} \sum_{\substack{r=1 \\ q \neq r}}^{Q} \sum_{k=1}^{n_r} (\epsilon_{ikr}^{(q)})^2 \Big/ \left\{ (Q-1) \sum_{r=1}^{Q} \sum_{k=1}^{n_r} (\epsilon_{ikr}^{(r)})^2 \right\} \quad (13)$$

Here $\epsilon_{ikr}^{(q)}$ denotes the residuals after fitting an object with index **k** belonging to class **r** to the class model **q**. Since the summation in $s^2_{not-class}$ is made over $Q - 1$ classes for each of the objects, the sum is corrected by dividing by $(Q - 1)$ on the right side. The values of equation 13 for the Iris-data are given in Table 5 It is seen that for $A = 1$, variable no 3 is much superior in discriminatory power while for $A = 2$, the variables are more similar.

2 5 A control of the objects in the reference sets

Just in the same way as the relevance of the variables is studied by comparing the residual variances with the data variances for each variable, the relevance of the objects can be studied by comparing the residual variance of each object (equation 7) with the residual variance for the whole class (equation 5) (F-test) The smaller the residual variance of an object, the larger its relevance. The residual variances of each object in the reference sets of the Iris example are given above in Table 1 One can see that none of the objects in the reference sets have an abnormally high variance, there is no need to exclude any "outliers"

2 6 Alternatives

It should be noted that the measures of typicality of variables and objects are discussed above within the framework of class models which are PC models General entropy based methods for handling the same problems have been developed by Wong *et al*[34] who also give references to other methods The entropy methods are especially attractive in applications where the modelling approach is less useful, for instance when the data are discrete or qualitative

2 7 Summary of present procedure

To handle a pattern recognition problem by the present method, I recommend the following procedure

1 Check the distribution of each variable within each class by making histograms for the reference sets (separately) Very skew distributions should be corrected by, for instance, taking the logarithm of the observations in that variable

2 Auto-scale the data so that, over all classes, each variable gets the mean zero and variance one

3 By means of crossvalidation (Section 2 2 1, example Fig 5), estimate how many product terms (A_q) are needed to adequately describe the reference set of each class by equation (4) If the "optimum" A_q values differ by more than one between the classes, then use these "optimum" values in the procedures below If the A_q values all are within one as in the Iris-example, then use the same value for A_q for all classes, namely the largest value found for one class (two in the Iris example)

4 Fit separate principal components models with A_q terms (equation 4) to each of the reference sets This gives values for the parameters α_i, β_{ia} and $\theta_{ak}(a = 1, A_q)$ for each class (Section 2 2 2, example in Table 3)

5. Fit all objects in the reference sets by linear regressions to all class-models with the parameters α and β fixed to the values obtained in step 4. The residuals will give information about "distances" between classes (Section 2.3, example in Table 4), variable "relevance" (Section 2.4, example Table 5) and possible outliers among the objects in the reference sets (Section 2 5)

6. Fit all the objects in the test set (the unclassified objects) to all class models by linear regressions, again with the parameters α and β fixed as in step 5

7 The residuals for each object will give information about "closest class" for that object, whether this closest class is close enough for the object to be classified as belonging to that class and whether the closest class is significantly closer than the next closest class (Section 2 2 4, example in Table 2).

The regression coefficients c_a (equation 6) for each object and class can be checked if they fall within the "normal" interval for the class (Section 2.2 3), if they do not, this is an indication of misfit

This concludes the analysis Naturally, in actual cases the analysis is made iteratively with the results of one cycle guiding such matters as transformation, exclusion or inclusion of variables, deletion of objects from the reference sets and so on in the next cycle Also, in each analysis specific questions arise which might be answered by a pattern recognition analysis, but which are difficult to discuss in general terms in the present context. Hence, the scheme above is a "standard procedure" and deviations from the scheme and additional steps are expected and encouraged in actual applications

3 DISCUSSION

By representing the objects in separate classes by completely separate models, it is possible to get a very simple and still powerful method of pattern

recognition The fact that it is possible to adequately represent any data (provided that a few assumptions are fulfilled) by a principal components model with few terms, makes the present method rather generally applicable The assumptions underlying the method are few, (a) that the data observed on the objects can be thought to derive from a continuous function in two vector variables, and (b) that the number of variables and objects in the reference sets are sufficient to "carry" the principal components model

The latter assumption is, in related forms, the basis for all methods of data analysis It corresponds to the assumption that the data in the reference sets are sufficiently representative for the actual class structure, if this assumption is not fulfilled, any method of analysing the data will fail

The first assumption about the "continuity" of the data seems to be very reasonable in many applications in natural sciences, but will not be fulfilled in data of "yes–no" type The inclusion of such data into the present methods have not yet been tried, but might consequently create special problems

In order to get a near optimally efficient classification method, however, more of the information contained by the data should be utilized for the method design as shown by Fukunaga *et al*[14] and Foley *et al*[18]

The present method does not directly use the information that the objects in different reference sets are indeed different Hence, the separation between different classes is not exaggerated The disjoint character of the method is also to advantage when new classes are introduced into a problem which has previously been analysed The old class models need not be recalculated, one only needs to analyse the new classes with their data and the fit of these new data to the old models

The fact that the present method works directly on the matrix of the original data makes it computationally fast and little storage demanding No distance matrix with $Nx(N-1)/2$ elements needs to be calculated and stored

The models used in the present approach, i e equation (4), are very similar to the models used by Snee[35] to analyse shape This indicates that the disjoint principal components models might work also in more classical areas of pattern recognition, such as the recognition of hand written characters The variables used in such applications should then be such where the continuity assumption is fairly well fulfilled, for example the lengths of various cross-sections of the characters

The present method has been applied to oil-data by Kowalski *et al*[36] and found to work very well In addition, the present similarity models (equation 4 with $A = 1$) have also been used in an application of pattern cognition (cluster analysis), where gaschromatography column packings ($N = 226$) were grouped according to their empirical similarity (10 variables)[37]

Several problems remain to be solved before the PC method is generally applicable The most important is that of missing data, i e the data matrices for the reference sets and test set are not complete Estimation procedures have been developed for the one component model (equation 2) by Christoffersson[38] and might be possible to extend to the two component model ($A = 2$ in equation 4) Development of pattern recognition methods based on these procedures for estimation of parameters in disjoint class models is under way in our laboratory

Acknowledgements—Much of the present work was done while I spent a most enjoyable time at the Department of Statistics, University of Wisconsin Madison, and I am greatly indebted to all members of the department for their kind support and helpful comments To professor Herman Wold, I am very grateful for his enthusiastic and most valuable help in all phases of the project A referee has been most helpful with pertinent references and constructive criticism

The project has been supported by the Swedish Natural Science Research Council, the Institute of Applied Mathematics, Stockholm, and the Graduate School of University of Wisconsin

REFERENCES

1 L Kanal, Patterns in pattern recognition 1968–74, *IEEE Trans Inform Theory* **20**, 697 (1974)
2 T Cacoullos and G P H Styan, A Bibliography of discriminant analysis, *Discriminant Analysis and Applications* (T Cacoullos, ed) p 375 Academic Press, N Y (1973)
3 S Das Gupta, Theories and methods in classification A review, *Discriminant Analysis and Applications* (T Cacoullos, ed), p 77 Academic Press, N Y (1973)
4 T W Anderson, S Das Gupta and G P H Styan *A Bibliography of Multivariate Analysis* Oliver & Boyd, Edinburgh (1972)
5 M S Watanabe (ed) *Frontiers of Pattern Recognition* Academic Press, N Y (1972)
6 B R Kowalski Pattern recognition in chemical research, *Computers in Chemical and Biochemical Research* Vol 2 (C E Klopfenstein and C L Wilkins eds) Academic Press, N Y (1974)
7 P H A Sneath and R R Sokal, *Numerical Taxonomy* Freeman, San Francisco (1973)
8 J E Leffler and E Grunwald *Rates and Equilibria of Organic Reactions* Wiley, N Y (1963)
9 V A Palm *Osnovy Kolitchestvennoi Teorii Organicheskikh Reaktsii Izd Khimiya*, Leningrad 1967 (German translation, Akademie Verlag, Berlin DDR 1971)
10 N B Chapman and J Shorter (eds) *Advances in Linear Free Energy Relations* Plenum N Y (1972)
11 S Wold, A theoretical foundation of extrathermodynamic relationships (LFER), *Chem Scripta* **5**, 97 (1974)
12 G S Hammond, Information management and original thought in chemical education *J chem Educ* **51**, 558 (1974)
13 K Fukunaga, *Introduction to Statistical Pattern Recognition* Ch 8 Academic Press, N Y (1972)
14 K Fukunaga and W L G Koontz Application of the Karhunen–Loeve expansion to feature selection and ordering *IEEE Trans Comp* **19**, 311 (1970)
15 M S Watanabe and N Pakvasa, Subspace method in pattern recognition *Proc 1st Intl Joint Conf on Pattern Recognition* Washington D C October 1973 IEEE Catalog No 73 CHO 821-9C, p 25

16 M S Watanabe, P F Lambert, C A Kulikowski, J L Buxton and R Walker Evaluation and selection of variables in pattern recognition, *Computer and Information Sciences* Vol **2**, (J Tou, ed), p 91 Academic Press, N Y (1967)
17 W J Dixon and R I Jennrich, Computer graphical analysis and discrimination, *Discriminant Analysis and Applications* (T Cacoullos, ed), p 161 Academic Press, N Y (1973)
18 D H Foley and J W Sammon, Jr, An optimal set of discriminant vectors, *IEEE Trans Comp* **24**, 281 (1975)
19 S Wold, Some comments on possible types of analogy models in the study of relations between chemical structure and biological activity, *Proc Conf Chem Structure—Biol Activity Rel s* Prague, June 1973 Akadémiai Kiado, Hung Acad Sci Budapest (1976).
20 S Wold, Pattern cognition and recognition based on disjoint principal components models *Tech Rep 357*, Dept of Statistics, University of Wisconsin, Madison WI 53706 (1974)
21 S Wold, Pattern cognition and recognition based on disjoint principal components models *Proc 2nd Int Jnt Conf Pattern Recognition* Copenhagen August 1974, IEEE Catalog No 74 CHO 885–4 C, p. 43
22 R A Fisher, The use of multiple measurements in taxonomic problems, *Ann Eugenics* **7**, 179 (1936)
23 G E P Box and D R Cox, An analysis of transformations *J R statist Soc* **B26**, 211 (1964)
24 F Mosteller and D L Wallace, Inference in an authorship problem, *J Am statist Ass* **58**, 275 (1963)
25 M Stone, Crossvalidatory choice and assessment of statistical predictions *J R statist Soc* **B36**, 111 (1974)
26 N R Draper and H Smith, *Applied Regression Analysis* Wiley, N Y (1966)
27 J Greenstadt, The determination of the characteristic roots of a matrix by the Jacobi method, *Mathematical Methods for Digital Computers* (A Ralston and H S Wilf, eds), p 84 Wiley, N Y (1960)
28 B T Smith *et al Matrix eigensystem routines—EISPACK guide* Springer, Berlin (1974)
29 H Wold, Nonlinear estimation by iterative least squares procedures, *Festschrift for J Neyman* (F N David, ed), p 411 Wiley N Y (1966).
30 S Wold and M. Sjostrom, Statistical analysis of the Hammett equation, *Chem. Scripta* **2**, 49 (1972)
31 H H Harman, *Modern Factor Analysis*, 2nd Edn University of Chicago Press, Chicago, IL (1967)
32 N J Nilsson *Learning Machines* McGraw-Hill, N Y (1965)
33 E Fix and J L Hodges, Nonparametric discrimination Consistency properties U S Air Force School of Aviation Medicine, Rep No 4, Randolph Field, TX (1951) See further Ref 3
34 A K C Wong and T S Liu, Typicality, diversity and feature pattern of an ensemble, *IEEE Trans Comp* **24**, 158 (1975)
35 R D Snee and H P Andrews, Statistical design and analysis of shape studies, *Appl Stat* **20**, 250 (1971)
36 D L Duewer, B R Kowalski and T F Schatzki, Source identification of oil spills by pattern recognition analysis of natural elemental composition, *Anal Chem* **47**, 1573 (1975)
37 S Wold Analysis of similarities and dissimilarities between gaschromatographic liquid phases by means of pattern cognition, *J Chromatogr Sci* (1975 in press)
38 A Christoffersson, *The one component model with incomplete data*, *Thesis*, Dept of Statistics, University of Uppsala, Sweden, 1970 Summary in *Bull int Statist Inst*, **1**, 31 (1969)

APPENDIX

The following treatment is an attempt to show that equation (4) can, with sufficiently many product terms (A), adequately describe any data matrix obtained from measurements made on objects in a class of similar objects Let us introduce two sets of "microscopic" variables denoted by the vector variables **z** and **x** These microscopic variables should be distinguished from the measured, macroscopic, variables that can be directly observed Let us further assume that the observed variables (**y**) can be written as a function of the microscopic variables

$$\mathbf{y} = F(\mathbf{z},\mathbf{x}) \tag{A1}$$

Let us finally assume that the vector variables **z** and **x** have been chosen so that everything that changes between the measured variables is included in the microscopic variable **z** and everything that changes between the objects is included in the vector **x**. It should be noted that in general, one does not know which microscopic factors that do change between the macroscopic, measured, variables or between the different objects in a class Usually, however, the assumption that the observed data are functions of microscopic variables is rather well accepted, at least in the natural sciences In chemistry, these microscopic variables are electronic variables such as charge distribution and spin distribution in various orbitals, steric interactions, van der Waals forces, solute–solvent interactions and whatever is fashionable to "explain" the behaviour of chemical systems In biology, such variables as the conformation of proteins, microscopic structure of membranes, chemical equilibria in cells and sequences of nucleic acids are thought to govern the behaviour of macroscopic biological systems

If we now start to look at a class of *identical* objects, this identity means that all the objects have identical values of all the elements in their x-vectors Consequently, the observations are functions only of the vector variable **z**, which in turn, means that they can be described by equation (1) in Section 2 2 If now this identity is relaxed just a little so that we study a class of very similar objects, the variation in the elements in **x** is so small that they all vary linearly with respect to each other Another way to express this is to say that the variation in the elements in **x** is so small that, if this variation is described as a Taylor expansion in a common variable, all the expansions contain linear terms only Consequently, the vector **x** contains only one independent element and the observations can be modelled as

$$\mathbf{y} = F(\mathbf{z},\mathbf{x}) \tag{A2}$$

This function can now, in turn, be expanded around an arbitrary point $(\mathbf{z}_0,\mathbf{x}_0)$ Using standard notation, we have ($R(T)$ denotes a remainder containing only terms of degree T and higher)

$$\begin{aligned} F(\mathbf{z},\mathbf{y}) = F_{00} &+ \sum_i F'_i \Delta z_i + F'_x \Delta x + \sum_i \sum_{i\leqslant j} F''_{ij} \Delta z_i \Delta z_j \\ &+ \sum_i F''_{ix} \Delta z_i \Delta x + F''_{xx} \Delta x^2 + \sum_i \sum_{i\leqslant j} \sum_{i\leqslant j\leqslant k} F'''_{ijk} \Delta z_i \Delta z_j \Delta z_k \\ &+ \sum_i \sum_{i\leqslant j} F'''_{ijx} \Delta z_i \Delta z_j \Delta x + \sum_i F'''_{ixx} \Delta z_i \Delta x^2 \\ &+ F'''_{xxx} \Delta x^3 + \quad + F^{(N)}_{x\ldots x} \Delta x^N + R(N+1) \end{aligned} \tag{A3}$$

Rearranging, we get

$$\begin{aligned} F(\mathbf{z},\mathbf{x}) = f(\mathbf{z}) &+ \Delta x \left\{ F'_x + \sum_i F''_{ix} \Delta z_i + \sum_i \sum_{i\leqslant j} F'''_{ijx} \Delta z_i \Delta z_j + \quad \right\} \\ &+ \Delta x^2 \left\{ F''_{xx} + \sum_i F'''_{ixx} \Delta z_i + \sum_i \sum_{i\leqslant j} F''''_{ijxx} \Delta z_i \Delta z_j + \quad \right\} \\ &+ \quad + \Delta x^{N-1} \left\{ F^{(N-1)}_{x\ldots x} + \sum_i F^{(N)}_{ix\ldots x} \Delta z_i \right\} \\ &+ \Delta x^N F^{(N)}_{xx\ldots x} + R(N+1) \end{aligned} \tag{A4}$$

This can, finally, be rewritten in the following form

$$F(\mathbf{z}, x) = f(\mathbf{z}) + \Delta x\left(F'_x + \sum_i F''_{ix}\Delta z_i\right) + \Delta x^2 F''_{xx} + R(3)$$

$$= f(\mathbf{z}) + \Delta x\left(F_x + \sum_i F''_{ix}\Delta z_i\right)$$

$$+ \Delta x^2\left(F'_x + \sum_i F''_{ix}\Delta z_i\right) F''_{xx}/F'_x$$

$$\underbrace{- \Delta x^2\left(\sum_i F''_{ix}\Delta z_i\right) F''_{xx}/F'_x + R(3)}_{R(3)}$$

$$= f(\mathbf{z}) + h(\mathbf{z})\, g(x) + R(3) \qquad \text{(A5)}$$

If we now translate equation (A5) back to the model of the observations, we get, since only $\mathbf{z}$ changes with the variables i and only x changes with the objects k

$$y_{ik} = F(\mathbf{z}, x) = f(\mathbf{z}) + h(\mathbf{z})\, g(x) + R(3) = \alpha_i + \beta_i \theta_k + \epsilon_{ik} \qquad \text{(A6)}$$

Hence if the objects in the class are sufficiently similar they can be described by equation (A6) (equation 2 in Section 2.2) The deviations (ϵ_{ik}) are seen to contain both model errors in terms of the remainder $R(3)$ and errors of measurement If the latter are large in comparison with the former, the model (A6) will adequately describe the data

If now the similarity between the objects in the class is relaxed somewhat further, the variation in the elements in $\mathbf{x}$ between the objects will increase and the variation in the elements will not longer be linear with respect to each other However, at first, each variable can be expressed as one linear and one quadratic term in a common variable When the variation between the objects becomes larger, three terms including a cubic are needed and so on Hence, when the variation between the objects becomes larger the vector $\mathbf{x}$ will contain, in turn, two, three, four, independent elements Below, an inductive proof will be given that when the vector variable $\mathbf{x}$ contains P independent terms, the function F can be approximated by a function in only $\mathbf{z}$ plus a sum of P product terms of functions in $\mathbf{z}$ and functions in $\mathbf{x}$

$$F(\mathbf{z}, \mathbf{x}) = f(\mathbf{z}) + \sum_{a=1}^{P} h_a(\mathbf{z})\, g_a(\mathbf{x}) + R(3) \qquad \text{(A7)}$$

This has already been proved for $P = 1$ in equation (A5) above Let us then assume that equation (A7) is valid up to P and try to prove it for $P + 1$ terms If the additional term is separately denoted by u, we wish to prove that

$$F(\mathbf{z}, \mathbf{x}\ u) = \Phi_a(\mathbf{z}) + \sum_{a=1}^{P} \gamma_a(\mathbf{z})\ \gamma_a(\mathbf{x}) + \Phi(\mathbf{z})\ \Psi(\mathbf{x}\ u) + R(3) \qquad \text{(A8)}$$

The validity of equation (A8) is best seen if we first study the extra terms in the Taylor expansion of $F(\mathbf{z}\,\mathbf{x})$

$$F(\mathbf{z}, x) = F_{00} + \sum_i F'_i \Delta z_i + \sum_r F'_r \Delta x_r + \sum\sum_{i \leqslant j} F''_{ij}\Delta z_i \Delta z_j + \sum\sum_{i\ r} F''_{ir}\Delta z_i \Delta x_r + \sum\sum_{r \leqslant s} F''_{rs}\Delta x_r \Delta x_s + \quad + R(3) \qquad \text{(A9)}$$

These extra terms are

$$\sum_i F''_{iu}\Delta z_i \Delta u + F'_u \Delta u + \sum_r F''_{ru}\Delta x_r \Delta u + F''_{uu}\Delta u^2 \qquad \text{(A10)}$$

These terms can, in fact, be written as $\Phi(\mathbf{z})\Psi(\mathbf{x})$ plus a remainder with third and higher order terms in the following way

$$\Phi(\mathbf{z})\ \Psi(\mathbf{x}) = (1 + \sum F''_{iu}\Delta z_i / F'_u)$$

$$(\Delta u\ \ F_u + \sum \Delta x_r \Delta u\ \ F''_{ru} + \Delta u^2\ \ F''_{uu}) \qquad \text{(A11)}$$

This proves the validity of equation (A8) which, in turn, proves the validity of equation (A7) It is seen that when equation (A7) is translated back to a model of the observations, we obtain, in analogy with equation (A6), equation (3) in Section 2.2 Hence, it has been proved that, for a collection of similar objects, the measurements made on these objects can be described by equation (3)

The proofs are seen to rest on the assumption that the derivatives F'_x in equation (A5) and F'_u in equation (A11) are different from zero This means that the expansion (A7) is valid only around points which are not local extremum points in any of the elements contained in $\mathbf{x}$ If the expansion is made around such an extremum point, more terms have to be included in the expansions, which means that more than P terms are needed in equation (A7) and equation (3) This number of terms will always be finite, however, which is easily seen by following analogous arguments as in the proofs above

About the Author—SVANTE WOLD was born in Stockholm, Sweden on 14 March 1941 He received a B S degree from University of Uppsala in 1963 in Mathematics Chemistry and Theoretical Physics and the Ph D degree in Chemistry from Umeå University in 1971

Dr Wold has been, since 1972, Assistant Professor in Organic Chemistry at Umeå University where he works with research in Chemometrics, i e the art of extracting chemically useful information from real data In 1973–1974, Dr Wold was on leave to the Dept of Statistics, University of Wisconsin, Madison, where he was Statistician in Residence at the Statistics Consulting Laboratory

Dr Wold is, together with Dr Bruce Kowalski, University of Washington, Seattle, a founder of the Chemometrics Society He is also member of the Pattern Recognition Society, Biometric Society and the Swedish Chemical Society