

EXERCISE 11.4 Repeat the 8-point example (Example 11.5) using the complete linkage and the Ward algorithm. Explain the difference to single linkage.

EXERCISE 11.6 Repeat the bank notes example (Example 11.6) with another random sample of 20 notes.

EXERCISE 11.7 Repeat the bank notes example (Example 11.6) with another clustering algorithm.

EXERCISE 11.9 Analyze the U.S. companies example (Table B.5) using the Ward algorithm and the L_2 -norm.

EXERCISE 11.10 Analyze the U.S. crime data set (Table B.10) with the Ward algorithm and the L_2 -norm on standardized variables (use only the crime variables).

EXERCISE 12.2 Apply the rule from Theorem 12.2 (b) for $p = 1$ and compare the result with that of Example 12.3.

EXERCISE 12.3 Calculate the ML discrimination rule based on observations of a one-dimensional variable with an exponential distribution.

EXERCISE 12.5 Apply the Bayes rule to the car data (Table B.3) in order to discriminate between Japanese, European and U.S. cars, i.e., $J = 3$. Consider only the “miles per gallon” variable and take the relative frequencies as prior probabilities.

EXERCISE 12.9 Recalculate Example 12.3 with the prior probability $\pi_1 = \frac{1}{3}$ and $C(2|1) = 2C(1|2)$.

EXERCISE 12.10 Explain the effect of changing π_1 or $C(1|2)$ on the relative location of the region $R_j, j = 1, 2$.

EXERCISE 12.12 Suppose that $x \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and

$\Pi_1 : X \sim \text{Bi}(10, 0.2)$ with the prior probability $\pi_1 = 0.5$;

$\Pi_2 : X \sim \text{Bi}(10, 0.3)$ with the prior probability $\pi_2 = 0.3$;

$\Pi_3 : X \sim \text{Bi}(10, 0.5)$ with the prior probability $\pi_3 = 0.2$.

Determine the sets R_1, R_2 and R_3 . (Use the Bayes discriminant rule.)

EXAMPLE 11.5 Here we describe the single linkage algorithm for the eight data points displayed in Figure 11.1. The distance matrix (L_2 -norms) is

$$D = \begin{pmatrix} 0 & 10 & 53 & 73 & 50 & 98 & 41 & 65 \\ & 0 & 25 & 41 & 20 & 80 & 37 & 65 \\ & & 0 & 2 & 1 & 25 & 18 & 34 \\ & & & 0 & 5 & 17 & 20 & 32 \\ & & & & 0 & 36 & 25 & 45 \\ & & & & & 0 & 13 & 9 \\ & & & & & & 0 & 4 \\ & & & & & & & 0 \end{pmatrix}$$

and the dendrogram is shown in Figure 11.2.

If we decide to cut the tree at the level 10, three clusters are defined: $\{1, 2\}$, $\{3, 4, 5\}$ and $\{6, 7, 8\}$.

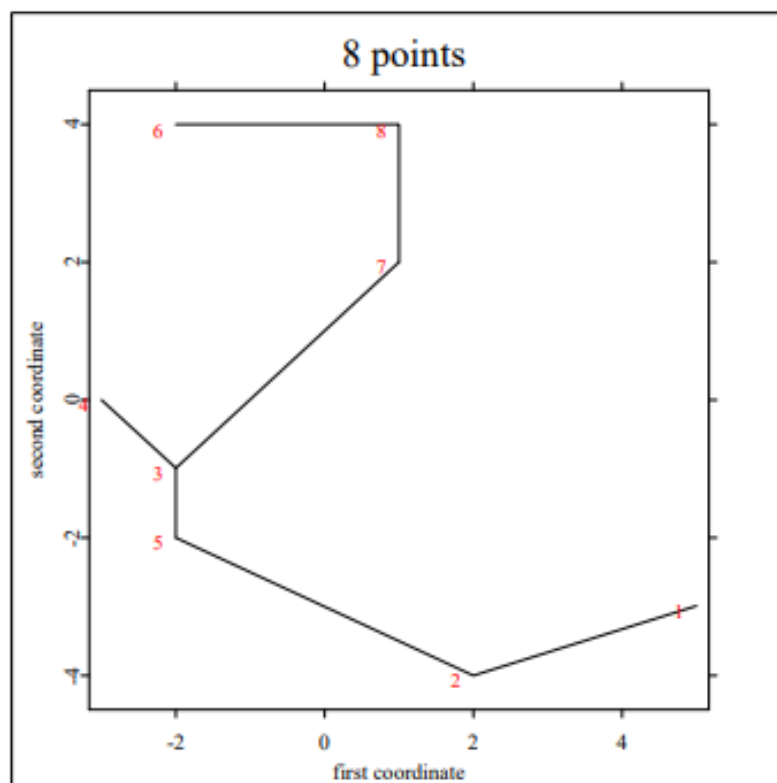


Figure 11.1. The 8-point example. [MVAclus8p.xpl](#)

The single linkage algorithm defines the distance between two groups as the smallest value of the individual distances. Table 11.4 shows that in this case

$$d(R, P + Q) = \min\{d(R, P), d(R, Q)\}. \quad (11.11)$$

This algorithm is also called the *Nearest Neighbor* algorithm. As a consequence of its construction, single linkage tends to build large groups. Groups that differ but are not well separated may thus be classified into one group as long as they have two approximate points. The *complete linkage* algorithm tries to correct this kind of grouping by considering the largest (individual) distances. Indeed, the complete linkage distance can be written as

$$d(R, P + Q) = \max\{d(R, P), d(R, Q)\}. \quad (11.12)$$

It is also called the *Farthest Neighbor* algorithm. This algorithm will cluster groups where all the points are proximate, since it compares the largest distances. The *average linkage* algorithm (weighted or unweighted) proposes a compromise between the two preceding

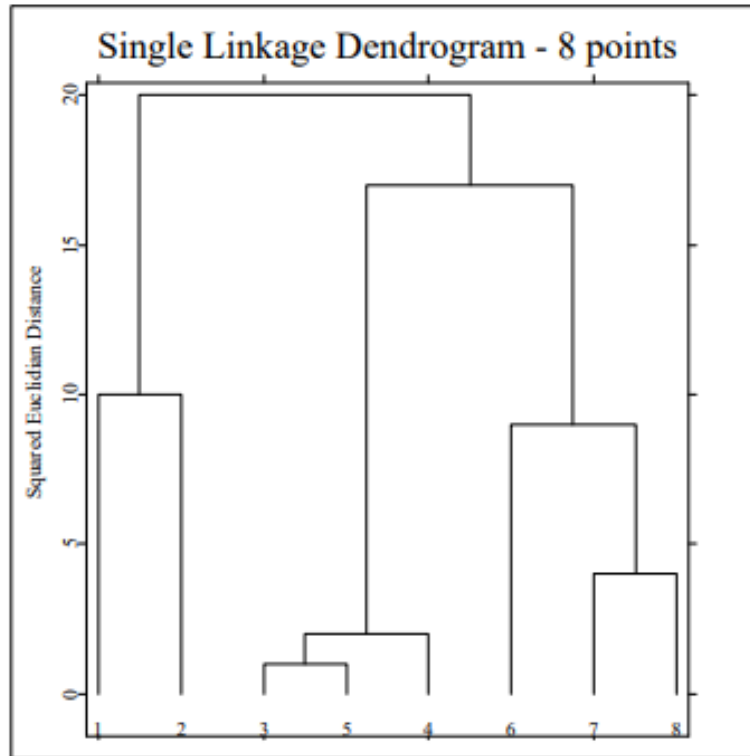


Figure 11.2. The dendrogram for the 8-point example, Single linkage algorithm. [MVAclus8p.xpl](#)

algorithms, in that it computes an average distance:

$$d(R, P + Q) = \frac{n_P}{n_P + n_Q}d(R, P) + \frac{n_Q}{n_P + n_Q}d(R, Q). \quad (11.13)$$

The *centroid* algorithm is quite similar to the average linkage algorithm and uses the natural geometrical distance between R and the weighted center of gravity of P and Q (see Figure 11.3):

$$d(R, P + Q) = \frac{n_P}{n_P + n_Q}d(R, P) + \frac{n_Q}{n_P + n_Q}d(R, Q) - \frac{n_P n_Q}{(n_P + n_Q)^2}d(P, Q). \quad (11.14)$$

The *Ward clustering* algorithm computes the distance between groups according to the formula in Table 11.4. The main difference between this algorithm and the linkage procedures is in the unification procedure. The Ward algorithm does not put together groups with smallest distance. Instead, it joins groups that do not increase a given measure of heterogeneity

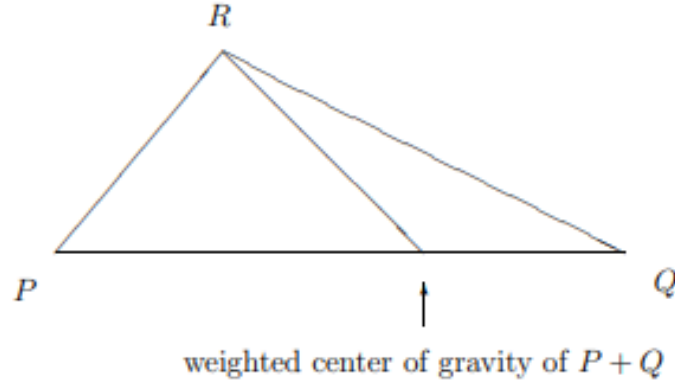


Figure 11.3. The centroid algorithm.

“too much”. The aim of the Ward procedure is to unify groups such that the variation inside these groups does not increase too drastically: the resulting groups are as homogeneous as possible.

The heterogeneity of group R is measured by the inertia inside the group. This inertia is defined as follows:

$$I_R = \frac{1}{n_R} \sum_{i=1}^{n_R} d^2(x_i, \bar{x}_R) \quad (11.15)$$

where \bar{x}_R is the center of gravity (mean) over the groups. I_R clearly provides a scalar measure of the dispersion of the group around its center of gravity. If the usual Euclidean distance is used, then I_R represents the sum of the variances of the p components of x_i inside group R .

When two objects or groups P and Q are joined, the new group $P + Q$ has a larger inertia I_{P+Q} . It can be shown that the corresponding increase of inertia is given by

$$\Delta(P, Q) = \frac{n_P n_Q}{n_P + n_Q} d^2(P, Q). \quad (11.16)$$

In this case, the Ward algorithm is defined as an algorithm that “joins the groups that give the smallest increase in $\Delta(P, Q)$ ”. It is easy to prove that when P and Q are joined, the new criterion values are given by (11.9) along with the values of δ_i given in Table 11.4, when the centroid formula is used to modify $d^2(R, P + Q)$. So, the Ward algorithm is related to the centroid algorithm, but with an “inertial” distance Δ rather than the “geometric” distance d^2 .

As pointed out in Section 11.2, all the algorithms above can be adjusted by the choice of the metric \mathcal{A} defining the geometric distance d^2 . If the results of a clustering algorithm are illustrated as graphical representations of individuals in spaces of low dimension (using principal components (normalized or not) or using a correspondence analysis for contingency tables), it is important to be coherent in the choice of the metric used.

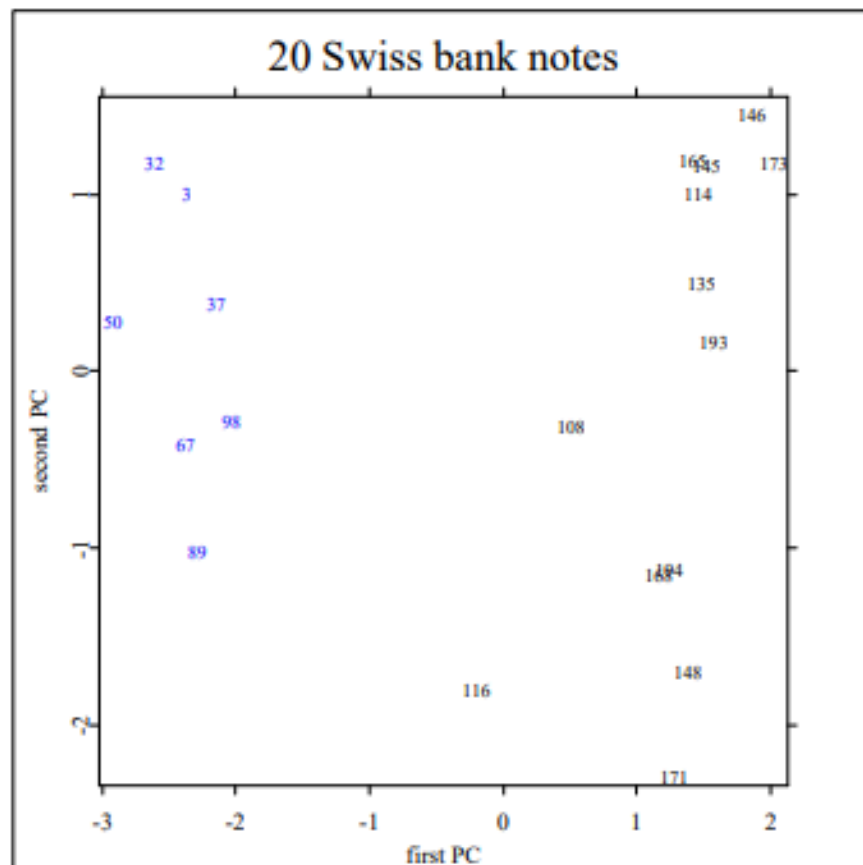


Figure 11.4. PCA for 20 randomly chosen bank notes.
[Q MVAclusbank.xpl](#)

EXAMPLE 11.6 *As an example we randomly select 20 observations from the bank notes data and apply the Ward technique using Euclidean distances. Figure 11.4 shows the first two PCs of these data, Figure 11.5 displays the dendrogram.*

B.3 Car Data

The car data set (Chambers, Cleveland, Kleiner and Tukey, 1983) consists of 13 variables measured for 74 car types. The abbreviations in Table B.3 are as follows:

X_1 :	P	Price,
X_2 :	M	Mileage (in miles per gallon),
X_3 :	R78	Repair record 1978 (rated on a 5-point scale; 5 best, 1 worst),
X_4 :	R77	Repair record 1977 (scale as before),
X_5 :	H	Headroom (in inches),
X_6 :	R	Rear seat clearance (distance from front seat back to rear seat, in inches),
X_7 :	Tr	Trunk space (in cubic feet),
X_8 :	W	Weight (in pound),
X_9 :	L	Length (in inches),
X_{10} :	T	Turning diameter (clearance required to make a U-turn, in feet),
X_{11} :	D	Displacement (in cubic inches),
X_{12} :	G	Gear ratio for high gear,
X_{13} :	C	Company headquarter (1 for U.S., 2 for Japan, 3 for Europe).

Model	P	M	R78	R77	H	R	Tr	W	L	T	D	G	C
AMC-Concord	4099	22	3	2	2.5	27.5	11	2930	186	40	121	3.58	1
AMC-Pacer	4749	17	3	1	3.0	25.5	11	3350	173	40	258	2.53	1
AMC-Spirit	3799	22	–	–	3.0	18.5	12	2640	168	35	121	3.08	1
Audi-5000	9690	17	5	2	3.0	27.0	15	2830	189	37	131	3.20	1
Audi-Fox	6295	23	3	3	2.5	28.0	11	2070	174	36	97	3.70	3
BMW-320i	9735	25	4	4	2.5	26.0	12	2650	177	34	121	3.64	3
Buick-Century	4816	20	3	3	4.5	29.0	16	3250	196	40	196	2.93	1
Buick-Electra	7827	15	4	4	4.0	31.5	20	4080	222	43	350	2.41	1
Buick-Le-Sabre	5788	18	3	4	4.0	30.5	21	3670	218	43	231	2.73	1
Buick-Opel	4453	26	–	–	3.0	24.0	10	2230	170	34	304	2.87	1
Buick-Regal	5189	20	3	3	2.0	28.5	16	3280	200	42	196	2.93	1
Buick-Riviera	10372	16	3	4	3.5	30.0	17	3880	207	43	231	2.93	1
Buick-Skylark	4082	19	3	3	3.5	27.0	13	3400	200	42	231	3.08	1
Cad.-Deville	11385	14	3	3	4.0	31.5	20	4330	221	44	425	2.28	1
Cad.-Eldorado	14500	14	2	2	3.5	30.0	16	3900	204	43	350	2.19	1
Cad.-Seville	15906	21	3	3	3.0	30.0	13	4290	204	45	350	2.24	1
Chev.-Chevette	3299	29	3	3	2.5	26.0	9	2110	163	34	231	2.93	1
Chev.-Impala	5705	16	4	4	4.0	29.5	20	3690	212	43	250	2.56	1
Chev.-Malibu	4504	22	3	3	3.5	28.5	17	3180	193	41	200	2.73	1
Chev.-Monte-Carlo	5104	22	2	3	2.0	28.5	16	3220	200	41	200	2.73	1
Chev.-Monza	3667	24	2	2	2.0	25.0	7	2750	179	40	151	2.73	1
Chev.-Nova	3955	19	3	3	3.5	27.0	13	3430	197	43	250	2.56	1
Datsun-200-SX	6229	23	4	3	1.5	21.0	6	2370	170	35	119	3.89	2
Datsun-210	4589	35	5	5	2.0	23.5	8	2020	165	32	85	3.70	2
Datsun-510	5079	24	4	4	2.5	22.0	8	2280	170	34	119	3.54	2
Datsun-810	8129	21	4	4	2.5	27.0	8	2750	184	38	146	3.55	2
Dodge-Colt	3984	30	5	4	2.0	24.0	8	2120	163	35	98	3.54	2
Dodge-Diplomat	5010	18	2	2	4.0	29.0	17	3600	206	46	318	2.47	1
Dodge-Magnum-XE	5886	16	2	2	3.5	26.0	16	3870	216	48	318	2.71	1
Dodge-St.-Regis	6342	17	2	2	4.5	28.0	21	3740	220	46	225	2.94	1
Fiat-Strada	4296	21	3	1	2.5	26.5	16	2130	161	36	105	3.37	3
Ford-Fiesta	4389	28	4	–	1.5	26.0	9	1800	147	33	98	3.15	1

Ford-Mustang	4187	21	3	3	2.0	23.0	10	2650	179	42	140	3.08	1
Honda-Accord	5799	25	5	5	3.0	25.5	10	2240	172	36	107	3.05	2
Honda-Civic	4499	28	4	4	2.5	23.5	5	1760	149	34	91	3.30	2
Linc.-Continental	11497	12	3	4	3.5	30.5	22	4840	233	51	400	2.47	1
Linc.-Cont-Mark-V	13594	12	3	4	2.5	28.5	18	4720	230	48	400	2.47	1
Linc.-Versailles	13466	14	3	3	3.5	27.0	15	3830	201	41	302	2.47	1
Mazda-GLC	3995	30	4	4	3.5	25.5	11	1980	154	33	86	3.73	1
Merc.-Bobcat	3829	22	4	3	3.0	25.5	9	2580	169	39	140	2.73	1
Merc.-Cougar	5379	14	4	3	3.5	29.5	16	4060	221	48	302	2.75	1
Merc.-Cougar-XR-7	6303	14	4	4	3.0	25.0	16	4130	217	45	302	2.75	1
Merc.-Marquis	6165	15	3	2	3.5	30.5	23	3720	212	44	302	2.26	1
Merc.-Monarch	4516	18	3	-	3.0	27.0	15	3370	198	41	250	2.43	1
Merc.-Zephyr	3291	20	3	3	3.5	29.0	17	2830	195	43	140	3.08	1
Olds-98	8814	21	4	4	4.0	31.5	20	4060	220	43	350	2.41	1
Olds-Cutlass	4733	19	3	3	4.5	28.0	16	3300	198	42	231	2.93	1
Olds-Cutl-Supr	5172	19	3	4	2.0	28.0	16	3310	198	42	231	2.93	1
Olds-Delta-88	5890	18	4	4	4.0	29.0	20	3690	218	42	231	2.73	1
Olds-Omega	4181	19	3	3	4.5	27.0	14	3370	200	43	231	3.08	1
Olds-Starfire	4195	24	1	1	2.0	25.5	10	2720	180	40	151	2.73	1
Olds-Tornado	10371	16	3	3	3.5	30.0	17	4030	206	43	350	2.41	1
Peugeot-604-SL	12990	14	-	-	3.5	30.5	14	3420	192	38	163	3.58	3
Plym.-Arrow	4647	28	3	3	2.0	21.5	11	2360	170	37	156	3.05	1
Plym.-Champ	4425	34	5	4	2.5	23.0	11	1800	157	37	86	2.97	1
Plym.-Horizon	4482	25	3	-	4.0	25.0	17	2200	165	36	105	3.37	1
Plym.-Sapporo	6486	26	-	-	1.5	22.0	8	2520	182	38	119	3.54	1
Plym.-Volare	4060	18	2	2	5.0	31.0	16	3330	201	44	225	3.23	1
Pont.-Catalina	5798	18	4	4	4.0	29.0	20	3700	214	42	231	2.73	1
Pont.-Firebird	4934	18	1	2	1.5	23.5	7	3470	198	42	231	3.08	1
Pont.-Grand-Prix	5222	19	3	3	2.0	28.5	16	3210	201	45	231	2.93	1
Pont.-Le-Mans	4723	19	3	3	3.5	28.0	17	3200	199	40	231	2.93	1
Pont.-Phoenix	4424	19	-	-	3.5	27.0	13	3420	203	43	231	3.08	1
Pont.-Sunbird	4172	24	2	2	2.0	25.0	7	2690	179	41	151	2.73	1
Renault-Le-Car	3895	26	3	3	3.0	23.0	10	1830	142	34	79	3.72	3
Subaru	3798	35	5	4	2.5	25.5	11	2050	164	36	97	3.81	2
Toyota-Cecila	5899	18	5	5	2.5	22.0	14	2410	174	36	134	3.06	2
Toyota-Corolla	3748	31	5	5	3.0	24.5	9	2200	165	35	97	3.21	2
Toyota-Corona	5719	18	5	5	2.0	23.0	11	2670	175	36	134	3.05	2
VW-Rabbit	4697	25	4	3	3.0	25.5	15	1930	155	35	89	3.78	3
VW-Rabbit-Diesel	5397	41	5	4	3.0	25.5	15	2040	155	35	90	3.78	3
VW-Scirocco	6850	25	4	3	2.0	23.5	16	1990	156	36	97	3.78	3
VW-Dasher	7140	23	4	3	2.5	37.5	12	2160	172	36	97	3.74	3
Volvo-260	11995	17	5	3	2.5	29.5	14	3170	193	37	163	2.98	3

B.5 U.S. Companies Data

The data set consists of measurements for 79 U.S. companies. The abbreviations in Table B.5 are as follows:

- X_1 : A Assets (USD),
- X_2 : S Sales (USD),
- X_3 : MV Market Value (USD),
- X_4 : P Profits (USD),
- X_5 : CF Cash Flow (USD),
- X_6 : E Employees.

Company	A	S	MV	P	CF	E	Sector
Bell Atlantic	19788	9084	10636	1092.9	2576.8	79.4	Communication
Continental Telecom	5074	2557	1892	239.9	578.3	21.9	Communication
American Electric Power	13621	4848	4572	485.0	898.9	23.4	Energy
Brooklyn Union Gas	1117	1038	478	59.7	91.7	3.8	Energy
Centra Illinois Publ. Serv.	1633	701	679	74.3	135.9	2.8	Energy
Cleveland Electric Illum.	5651	1254	2002	310.7	407.9	6.2	Energy
Columbia Gas System	5835	4053	1601	-93.8	173.8	10.8	Energy
Florida Progress	3494	1653	1442	160.9	320.3	6.4	Energy
Idaho Power	1654	451	779	84.8	130.4	1.6	Energy
Kansas Power & Light	1679	1354	687	93.8	154.6	4.6	Energy
Mesa Petroleum	1257	355	181	167.5	304.0	0.6	Energy
Montana Power	1743	597	717	121.6	172.4	3.5	Energy
Peoples Energy	1440	1617	639	81.7	126.4	3.5	Energy
Phillips Petroleum	14045	15636	2754	418.0	1462.0	27.3	Energy
Publ. Serv. Co of New Mexico	3010	749	1120	146.3	209.2	3.4	Energy
San Diego Gas & Electric	3086	1739	1507	202.7	335.2	4.9	Energy
Valero Energy	1995	2662	341	34.7	100.7	2.3	Energy
American Savings Bank FSB	3614	367	90	14.1	24.6	1.1	Finance
Bank South	2788	271	304	23.5	28.9	2.1	Finance
H & R Block	327	542	959	54.1	72.5	2.8	Finance
California First Bank	5401	550	376	25.6	37.5	4.1	Finance
Cigna	44736	16197	4653	-732.5	-651.9	48.5	Finance
Dreyfus	401	176	1084	55.6	57.0	0.7	Finance
First American	4789	453	367	40.2	51.4	3.0	Finance
First Empire State	2548	264	181	22.2	26.2	2.1	Finance
First Tennessee National	5249	527	346	37.8	56.2	4.1	Finance
Marine Corp	3720	356	211	26.6	34.8	2.4	Finance
Mellon Bank	33406	3222	1413	201.7	246.7	15.8	Finance
National City	12505	1302	702	108.4	131.4	9.0	Finance
Norstar Bancorp	8998	882	988	93.0	119.0	7.4	Finance
Norwest	21419	2516	930	107.6	164.7	15.6	Finance
Southeast Banking	11052	1097	606	64.9	97.6	7.0	Finance
Sovran Financial	9672	1037	829	92.6	118.2	8.2	Finance
United Financial Group	4989	518	53	-3.1	-0.3	0.8	Finance
Apple Computer	1022	1754	1370	72.0	119.5	4.8	HiTech
Digital Equipment	6914	7029	7957	400.6	754.7	87.3	HiTech
Eg & G	430	1155	1045	55.7	70.8	22.5	HiTech
General Electric	26432	28285	33172	2336.0	3562.0	304.0	HiTech
Hewlett-Packard	5769	6571	9462	482.0	792.0	83.0	HiTech
IBM	52634	50056	95697	6555.0	9874.0	400.2	HiTech
NCR	3940	4317	3940	315.2	566.3	62.0	HiTech
Telex	478	672	866	67.1	101.6	5.4	HiTech
Armstrong World Industries	1093	1679	1070	100.9	164.5	20.8	Manufacturing
CBI Industries	1128	1516	430	-47.0	26.7	13.2	Manufacturing
Fruehauf	1804	2564	483	70.5	164.9	26.6	Manufacturing
Halliburton	4662	4781	2988	28.7	371.5	66.2	Manufacturing
LTV	6307	8199	598	-771.5	-524.3	57.5	Manufacturing
Owens-Corning Fiberglas	2366	3305	1117	131.2	256.5	25.2	Manufacturing
PPG Industries	4084	4346	3023	302.7	521.7	37.5	Manufacturing
Textron	10348	5721	1915	223.6	322.5	49.5	Manufacturing
Turner	752	2149	101	11.1	15.2	2.6	Manufacturing
United Technologies	10528	14992	5377	312.7	710.7	184.8	Manufacturing
Commun. Psychiatric Centers	278	205	853	44.8	50.5	3.8	Medical
Hospital Corp of America	6259	4152	3090	283.7	524.5	62.0	Medical

AH Robins	707	706	275	61.4	77.8	6.1	Medical
Shared Medical Systems	252	312	883	41.7	60.6	3.3	Medical
Air Products	2687	1870	1890	145.7	352.2	18.2	Other
Allied Signal	13271	9115	8190	-279.0	83.0	143.8	Other
Bally Manufacturing	1529	1295	444	25.6	137.0	19.4	Other
Crown Cork & Seal	866	1487	944	71.7	115.4	12.6	Other
Ex-Cell-O	799	1140	633	57.6	89.2	15.4	Other
Liz Claiborne	223	557	1040	60.6	63.7	1.9	Other
Warner Communications	2286	2235	2306	195.3	219.0	8.0	Other
Dayton-Hudson	4418	8793	4459	283.6	456.5	128.0	Retail
Dillard Department Stores	862	1601	1093	66.9	106.8	16.0	Retail
Giant Food	623	2247	797	57.0	93.8	18.6	Retail
Great A & P Tea	1608	6615	829	56.1	134.0	65.0	Retail
Kroger	4178	17124	2091	180.8	390.4	164.6	Retail
May Department Stores	3442	5080	2673	235.4	361.5	77.3	Retail
Stop & Shop Cos	1112	3689	542	30.3	96.9	43.5	Retail
Supermarkets General	1104	5123	910	63.7	133.3	48.5	Retail
Wickes Cos	2957	2806	457	40.6	93.5	50.0	Retail
FW Woolworth	2535	5958	1921	177.0	288.0	118.1	Retail
AMR	6425	6131	2448	345.8	682.5	49.5	Transportation
IU International	999	1878	393	-173.5	-108.1	23.3	Transportation
PanAm	2448	3484	1036	48.8	257.1	25.4	Transportation
Republic Airlines	1286	1734	361	69.2	145.7	14.3	Transportation
TWA	2769	3725	663	-208.4	12.4	29.1	Transportation
Western AirLines	952	1307	309	35.4	92.8	10.3	Transportation

B.10 U.S. Crime Data

This is a data set consisting of 50 measurements of 7 variables. It states for one year (1985) the reported number of crimes in the 50 states of the U.S. classified according to 7 categories (X_3 – X_9).

- X_1 : land area (land)
- X_2 : population 1985 (popu 1985)
- X_3 : murder (murd)
- X_4 : rape
- X_5 : robbery (robb)
- X_6 : assault (assa)
- X_7 : burglary (burg)
- X_8 : larceny (larc)
- X_9 : autothieft (auto)
- X_{10} : US states region number (reg)
- X_{11} : US states division number (div)

<i>division numbers</i>		<i>region numbers</i>	
New England	1	Northeast	1
Mid Atlantic	2	Midwest	2
E N Central	3	South	3
W N Central	4	West	4
S Atlantic	5		
E S Central	6		
W S Central	7		
Mountain	8		
Pacific	9		

abb. of state	land area	popu 1985	murd	rape	robb	assa	burg	larc	auto	reg	div
ME	33265	1164	1.5	7.0	12.6	62	562	1055	146	1	1
NH	9279	998	2.0	6	12.1	36	566	929	172	1	1
VT	9614	535	1.3	10.3	7.6	55	731	969	124	1	1
MA	8284	5822	3.5	12.0	99.5	88	1134	1531	878	1	1
RI	1212	968	3.2	3.6	78.3	120	1019	2186	859	1	1
CT	5018	3174	3.5	9.1	70.4	87	1084	1751	484	1	1
NY	49108	17783	7.9	15.5	443.3	209	1414	2025	682	1	2
NJ	7787	7562	5.7	12.9	169.4	90	1041	1689	557	1	2
PA	45308	11853	5.3	11.3	106.0	90	594	11	340	1	2
OH	41330	10744	6.6	16.0	145.9	116	854	1944	493	2	3
IN	36185	5499	4.8	17.9	107.5	95	860	1791	429	2	3
IL	56345	11535	9.6	20.4	251.1	187	765	2028	518	2	3
MI	58527	9088	9.4	27.1	346.6	193	1571	2897	464	2	3
WI	56153	4775	2.0	6.7	33.1	44	539	1860	218	2	3
MN	84402	4193	2.0	9.7	89.1	51	802	1902	346	2	4
IA	56275	2884	1.9	6.2	28.6	48	507	1743	175	2	4
MO	69697	5029	10.7	27.4	2.8	167	1187	2074	538	2	4
ND	70703	685	0.5	6.2	6.5	21	286	1295	91	2	4
SD	77116	708	3.8	11.1	17.1	60	471	1396	94	2	4
NE	77355	1606	3.0	9.3	57.3	115	505	1572	292	2	4
KS	82277	2450	4.8	14.5	75.1	108	882	2302	257	2	4
DE	2044	622	7.7	18.6	105.5	196	1056	2320	559	3	5

MD	10460	4392	9.2	23.9	338.6	253	1051	2417	548	3	5
VA	40767	5706	8.4	15.4	92.0	143	806	1980	297	3	5
WV	24231	1936	6.2	6.7	27.3	84	389	774	92	3	5
NC	52669	6255	11.8	12.9	53.0	293	766	1338	169	3	5
SC	31113	3347	14.6	18.1	60.1	193	1025	1509	256	3	5
GA	58910	5976	15.3	10.1	95.8	177	9	1869	309	3	5
FL	58664	11366	12.7	22.2	186.1	277	1562	2861	397	3	5
KY	40409	3726	11.1	13.7	72.8	123	704	1212	346	3	6
TN	42144	4762	8.8	15.5	82.0	169	807	1025	289	3	6
AL	51705	4021	11.7	18.5	50.3	215	763	1125	223	3	6
MS	47689	2613	11.5	8.9	19.0	140	351	694	78	3	6
AR	53187	2359	10.1	17.1	45.6	150	885	1211	109	3	7
LA	47751	4481	11.7	23.1	140.8	238	890	1628	385	3	7
OK	69956	3301	5.9	15.6	54.9	127	841	1661	280	3	7
TX	266807	16370	11.6	21.0	134.1	195	1151	2183	394	3	7
MT	147046	826	3.2	10.5	22.3	75	594	1956	222	4	8
ID	83564	15	4.6	12.3	20.5	86	674	2214	144	4	8
WY	97809	509	5.7	12.3	22.0	73	646	2049	165	4	8
CO	104091	3231	6.2	36.0	129.1	185	1381	2992	588	4	8
NM	121593	1450	9.4	21.7	66.1	196	1142	2408	392	4	8
AZ	1140	3187	9.5	27.0	120.2	214	1493	3550	501	4	8
UT	84899	1645	3.4	10.9	53.1	70	915	2833	316	4	8
NV	110561	936	8.8	19.6	188.4	182	1661	3044	661	4	8
WA	68138	4409	3.5	18.0	93.5	106	1441	2853	362	4	9
OR	97073	2687	4.6	18.0	102.5	132	1273	2825	333	4	9
CA	158706	26365	6.9	35.1	206.9	226	1753	3422	689	4	9
AK	5914	521	12.2	26.1	71.8	168	790	2183	551	4	9
HI	6471	1054	3.6	11.8	63.3	43	1456	3106	581	4	9

EXAMPLE 12.3 Consider two normal populations

$$\begin{aligned}\Pi_1 &: N(\mu_1, \sigma_1^2), \\ \Pi_2 &: N(\mu_2, \sigma_2^2).\end{aligned}$$

Then

$$L_i(x) = (2\pi\sigma_i^2)^{-1/2} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu_i}{\sigma_i} \right)^2 \right\}.$$

Hence x is allocated to Π_1 ($x \in R_1$) if $L_1(x) \geq L_2(x)$. Note that $L_1(x) \geq L_2(x)$ is equivalent to

$$\frac{\sigma_2}{\sigma_1} \exp \left\{ -\frac{1}{2} \left[\left(\frac{x - \mu_1}{\sigma_1} \right)^2 - \left(\frac{x - \mu_2}{\sigma_2} \right)^2 \right] \right\} \geq 1$$

or

$$x^2 \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) - 2x \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right) \leq 2 \log \frac{\sigma_2}{\sigma_1}. \quad (12.5)$$

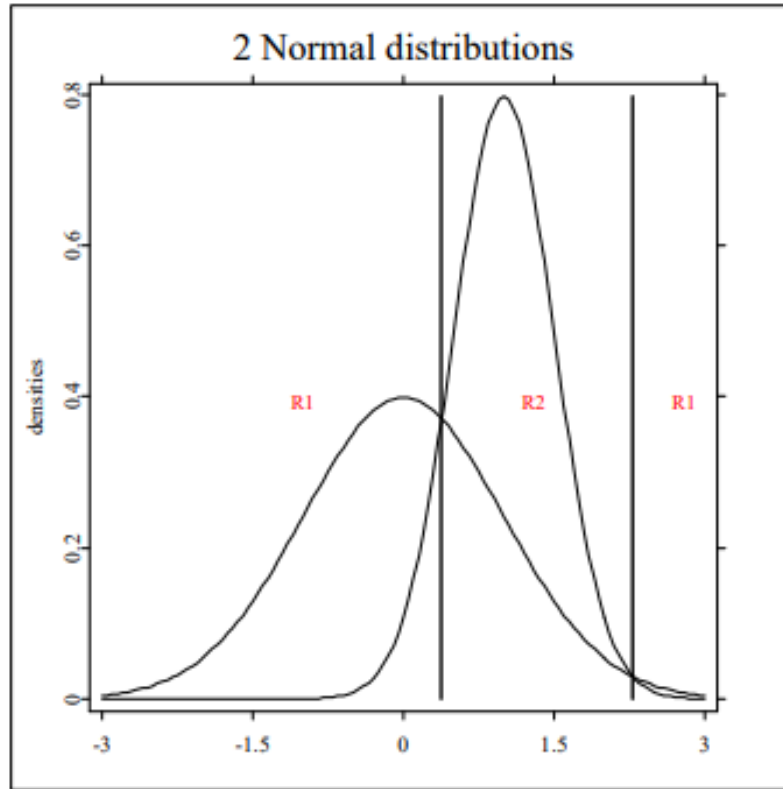



Figure 12.1. Maximum likelihood rule for normal distributions.
 `MVAdisnorm.xpl`

Suppose that $\mu_1 = 0$, $\sigma_1 = 1$ and $\mu_2 = 1$, $\sigma_2 = \frac{1}{2}$. Formula (12.5) leads to

$$\begin{aligned} R_1 &= \left\{ x : x \leq \frac{1}{3} \left(4 - \sqrt{4 + 6 \log(2)} \right) \text{ or } x \geq \frac{1}{3} \left(4 + \sqrt{4 + 6 \log(2)} \right) \right\}, \\ R_2 &= \mathbb{R} \setminus R_1. \end{aligned}$$

This situation is shown in Figure 12.1.

The situation simplifies in the case of equal variances $\sigma_1 = \sigma_2$. The discriminant rule (12.5) is then (for $\mu_1 < \mu_2$)

$$\begin{aligned} x &\rightarrow \Pi_1, & \text{if } x \in R_1 &= \{x : x \leq \tfrac{1}{2}(\mu_1 + \mu_2)\}, \\ x &\rightarrow \Pi_2, & \text{if } x \in R_2 &= \{x : x > \tfrac{1}{2}(\mu_1 + \mu_2)\}. \end{aligned} \tag{12.6}$$

Theorem 12.2 shows that the ML discriminant rule for multinormal observations is intimately connected with the Mahalanobis distance. The discriminant rule is based on linear combinations and belongs to the family of Linear Discriminant Analysis (LDA) methods.

THEOREM 12.2 Suppose $\Pi_i = N_p(\mu_i, \Sigma)$.

(a) The ML rule allocates x to Π_j , where $j \in \{1, \dots, J\}$ is the value minimizing the square Mahalanobis distance between x and μ_i :

$$\delta^2(x, \mu_i) = (x - \mu_i)^\top \Sigma^{-1} (x - \mu_i), \quad i = 1, \dots, J.$$

(b) In the case of $J = 2$,

$$x \in R_1 \iff \alpha^\top (x - \mu) \geq 0,$$

where $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\mu = \frac{1}{2}(\mu_1 + \mu_2)$.

Proof:

Part (a) of the Theorem follows directly from comparison of the likelihoods.

For $J = 2$, part (a) says that x is allocated to Π_1 if

$$(x - \mu_1)^\top \Sigma^{-1} (x - \mu_1) \leq (x - \mu_2)^\top \Sigma^{-1} (x - \mu_2)$$

Rearranging terms leads to

$$-2\mu_1^\top \Sigma^{-1} x + 2\mu_2^\top \Sigma^{-1} x + \mu_1^\top \Sigma^{-1} \mu_1 - \mu_2^\top \Sigma^{-1} \mu_2 \leq 0,$$

which is equivalent to

$$2(\mu_2 - \mu_1)^\top \Sigma^{-1} x + (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 + \mu_2) \leq 0,$$

$$(\mu_1 - \mu_2)^\top \Sigma^{-1} \left\{ x - \frac{1}{2}(\mu_1 + \mu_2) \right\} \geq 0,$$

$$\alpha^\top (x - \mu) \geq 0.$$

□