massimizzare $a^T F' a$ con il vincolo $|a| = 1$; questo equivale a dire che a deve essere l'autovettore di F' corrispondente al suo massimo autovalore. Sia U la matrice che ha per colonna gli autovettori ordinati in modo tale che sulla prima colonna si abbia l'autovettore corrispondente all'autovalore massimo, e così via[1]. Le funzioni discriminanti sono quindi:

$$X'U = XSU.$$

In R questo tipo di analisi si può effettuare facendo uso della funzione *lda* disponibile all'interno della libreria $MASS$, la quale in realtà opera con una logica leggermente diversa preferendo alla ricerca degli autovettori la decomposizione ai valori singolari delle matrici in questione.

Esempio

Riprendendo l'esempio delle anfore cretesi, si supponga che sia nota la loro datazione. Le prime 5 anfore risalgono ad un periodo più antico, le ultime 6 sono le più recenti e le rimanenti 4 hanno un'età intermedia. Si cerca di ricavare una classificazione analoga utilizzando solo i parametri di dimensione $x1, \ldots, x4$. Per prima cosa si crea il fattore *grp* che tiene traccia del gruppo di appartenenza reale delle anfore:

```
> grp <- factor( c(1,1,1,1,1,2,2,2,2,3,3,3,3,3,3) )
```

A questo punto l'analisi discriminante si esegue con la semplice chiamata:

```
> library(MASS)
> discr <- lda(grp ~ ., data=X)
```

Dove X è, come in precendenza, il data frame che contiene i valori delle dimensioni, rilevate sulle 15 anfore. L'output della funzione è abbastanza ricco:

```
> discr
Call:
lda(grp ~ ., data = X)

Prior probabilities of groups:
        1         2         3
0.3333333 0.2666667 0.4000000

Group means:
        x1       x2    x3     x4
1 22.46000 19.20000 29.66  9.900
2 24.60000 21.05000 32.85 10.925
3 24.28333 20.96667 33.10 10.550

Coefficients of linear discriminants:
          LD1        LD2
x1 -1.8266939 -3.0112168
x2  2.2243913  2.7660701
x3 -0.6399994  0.3771298
x4 -0.7181312 -0.2750134

Proportion of trace:
   LD1    LD2
0.8942 0.1058
```

[1] In generale è possibile costruire al massimo $\min(p, r-1)$ funzioni discriminanti dato che questo è il numero massimo di autovalori non nulli che ammette la matrice F'.

Dapprima vengono valutate le probabilità a priori di far parte di un dato gruppo. In mancanza di informazioni specificate dallo sperimentatore (tramite l'opzione *prior* della funzione *lda*) esse sono esattamente le proporzioni di soggetti nei vari gruppi di *grp*. A questa informazione segue una tabella in cui vengono riepilogate le medie dei predittori all'interno dei tre gruppi definiti dal fattore *grp*. Questi vettori di coordinate definiscono i *centroidi* dei gruppi. Infine vi è la tabella dei coefficienti delle funzioni lineari che meglio separano i soggetti rispetto alle classi specificate dal fattore cronologico in esame. Dato che i gruppi sono tre è possibile costruire due funzioni di questo genere (indicate da R con le sigle LD1 e LD2). Secondo i dati dell'analisi, la funzione lineare dei predittori che meglio separa i dati, identificata da LD1, è:

$$LD1 = -1.83 * x1 + 2.22 * x2 - 0.64 * x3 - 0.72 * x4.$$

Tale funzione separa le tre popolazioni di anfore molto meglio di quanto non farebbe la funzione LD2, come si evince dall'ultima riga dell'output. Per ogni funzione discriminante il valore riportato è infatti il rapporto fra l'autovalore corrispondente della matrice F' e la somma degli autovalori stessi. Questa quantità rappresenta proprio la proporzione di varianza fra gruppi interpretata dalle funzioni lineari trovate.

Per verificare i calcoli dell'algoritmo è possibile effettuare manualmente la procedura:

```
> S <- cov(X)          # matrice di covarianza totale
> M <-  discr$means    # coordinate dei centroidi
> G <- NULL; for(i in 1:3) G <- cbind(G, as.numeric(grp == levels(grp)[i]))
> E <- t(as.matrix(anfore) - G %*% M) %*% (as.matrix(anfore) - G %*% M) / 12
> F <- (14 * S - 12 * E)/2

> eigen( solve(E) %*% F )
$values
[1]  8.892942e+00  1.052540e+00 -5.553715e-14 -1.207411e-14

$vectors
           [,1]       [,2]       [,3]        [,4]
[1,] -0.6019148 -0.7316967 -0.5933035 -0.68597423
[2,]  0.7329603  0.6721284  0.7546568  0.06526138
[3,] -0.2108865  0.0916389 -0.1208490  0.19837373
[4,] -0.2366318 -0.0668256  0.2527443  0.69701375
```

Si hanno le componenti degli autovettori (non scalate) e i corrispondenti autovalori. Per questo problema si possono avere un massimo di due funzioni discriminanti. In effetti si nota che, entro la precisione algoritmica, gli ultimi due autovalori sono nulli; nei calcoli seguenti si possono quindi trascurare i relativi autovettori. La proporzione di varianza fra gruppi spiegata dalla prima funzione discriminante è:

```
> 8.892942e+00/(8.892942e+00 + 1.052540e+00)
[1] 0.894169
```

risultato coincidente con quanto visto in precedenza. Per quanto riguarda la matrice degli autovettori si procede alla scalatura in modo da rendere approssimativamente circolare la dispersione dei punti nei gruppi. Detta U la matrice degli autovettori questo risultato si ottiene dividendo ogni autovettore u_k per la radice quadrata della quantità $u_k^T E u_k$ (la varianza entro gruppi dell'autovettore u_k). In notazione matriciale la matrice normalizzata C è:

$$C = U(U^T EU)^{-1/2}$$

dove la matrice di scalatura $U^T EU$ è diagonale. Nel caso dell'esempio in questione si ha:

```
> U <- eigen( solve(E) %*% F )$vectors[,1:2]
```

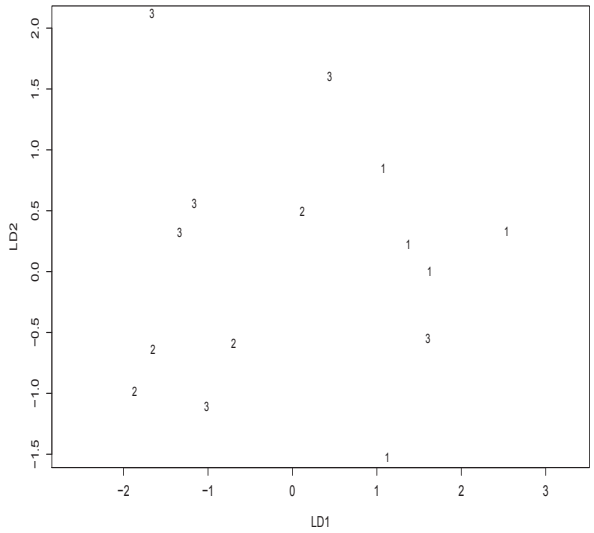

Figura 9.1: Grafico dei gruppi, corrispondenti alle diverse datazioni delle 15 anfore (dalle più antiche individuate dal numero 1, alle più recenti individuate dal numero 3), sul piano individuato dalle prime due funzioni lineari discriminanti. Si nota una buona separazione fra le tre popolazioni.

```
> scale <- sqrt(diag(diag((t(U) %*% E %*% U))))    # matrice di scalatura
> C <- U %*% solve(scale)

> C
           [,1]       [,2]
[1,] -1.8266939 -3.0112168
[2,]  2.2243913  2.7660701
[3,] -0.6399994  0.3771298
[4,] -0.7181312 -0.2750134
```

Si verifica che questo è quanto riportato dall'algoritmo *lda* (eventualmente a meno del segno).

Per avere un'idea della bontà delle funzioni discriminanti individuate è possibile graficare i valori che assumono LD1 e LD2 sugli individui e segnare ogni punto con il corrispondente valore del fattore *grp*. Se, in questa particolare visualizzazione, le popolazioni risultano ben separate si può concludere che la tecnica ha raggiunto il suo scopo (si noti per inciso che questo avviene quando la varianza all'interno dei gruppi è piccola rispetto a quella fra gruppi). In R è possibile realizzare questo grafico con la chiamata:

```
> plot(discr)
```

che produce l'output di Fig. 9.1, da cui si può concludere che i tre gruppi di anfore sono sufficientemente ben separate dalle due funzioni LD1 e LD2. Nel calcolare i valori che le funzioni assumono sui dati campionari, al posto dei vettori $x1, \ldots, x4$ vengono usati i vettori traslati $x1 - \bar{x1}, \ldots, x4 - \bar{x4}$.

9.1.1 Allocazione dei soggetti nelle classi

Per assegnare un soggetto sperimentale a una classe è necessario introdurre un algoritmo di allocazione. Il più semplice di essi è quello di calcolare il valore delle funzioni discriminanti sul soggetto in questione, fare lo stesso sui centroidi delle classi e assegnare il dato alla classe il cui centroide è più vicino. Nel far questo si trascurano le probabilità (a priori) di far parte delle diverse classi.