UNIVERSITÀ DEGLI STUDI DI TRENTO

DIPARTIMENTO DI MATEMATICA

Corso di Laurea in Matematica

Distribuzioni mistura e algoritmo EM nella lotta alla contraffazione: il caso dei mosti concentrati rettificati

Supervisore:

Prof. Pier Luigi Novi Inverardi

Candidato:

Giuseppe Nicola Liso

Firma

Pier Luigi Novi Inverardi

ANNO ACCADEMICO 2023/2024

Indice

Capitolo 1

Introduzione e flessibilità dei modelli mistura

L'importanza dei modelli mistura finiti nell'analisi statistica dei dati è facilmente riscontrabile grazie all'enorme quantità di articoli pratici e teorici in cui questi appaiono. Il motivo principale dell'enorme importanza delle misture è dovuta al fatto che queste ultime sono spesso utilizzate per fornire buone rappresentazioni di modelli di distribuzioni che trattano dati di fenomeni casuali.
Le distribuzioni mistura trovano applicazioni in numerosi campi, grazie proprio alla loro flessibilità, versatilità e propensione all'adattamento ai diversi modelli della realtà. Molto spesso, infatti, queste sono state utilizzate per risolvere problemi relativi all'agricoltura, astronomia, biologia, medicina e molti altri campi che riguardano la fisica e le scienze sociali. In queste applicazioni, i modelli mistura finiti sorreggono una grande varietà di tecniche che riguardano le maggiori aree della statistica, come l'analisi delle classi latenti o la così detta *cluster analysis*, ma giocano un ruolo importante soprattutto nell'inferenza e nell'analisi dei modelli statistici in cui, considerando un insieme di dati, ci si accorge che non può essere studiato tramite una singola distribuzione.
Il problema principale risolto efficacemente dalle distribuzioni mistura emerge quando i dati non sono disponibili per ciascuna componente separatamente, ma solo per l'intero modello mistura. Questa situazione è comune nella realtà, poiché spesso esistono più popolazioni o sottopopolazioni con caratteristiche diverse e non è possibile studiare o osservare alcune variabili sottostanti che suddividono le osservazioni in gruppi distinti e, di conseguenza, è necessario analizzare la distribuzione congiunta. In altre parole, le distribuzioni mistura riescono a catturare l'eterogeneità nei dati di un modello generico, cosa che una singola distribuzione non potrebbe fare. Questo avviene perché, quando i dati provengono da diverse sottopopolazioni o gruppi, ognuno con caratteristiche distintive, come ad esempio una diversa media (centro), varianza, o entrambe, una singola distribuzione non

è in grado di rappresentare adeguatamente la complessità della struttura dei dati. Consideriamo il caso di un insieme di dati proveniente da un fenomeno in cui coesistono più gruppi con distribuzioni normali. Se ciascun gruppo ha una media (centro) diversa, una varianza diversa, o una combinazione di entrambe, il tentativo di descrivere l'intero insieme dei dati utilizzando una singola distribuzione normale risulterebbe in una rappresentazione distorta e inefficace. Una distribuzione singola, infatti, tenderebbe a mediare le differenze tra i gruppi, fornendo una stima unica della media e della varianza che non riflette accuratamente la realtà dei dati. Di conseguenza, la stima della distribuzione risulterebbe approssimativa e non in grado di cogliere le peculiarità dei singoli gruppi. Al contrario, un modello mistura è capace di rappresentare ciascun gruppo con la sua propria distribuzione, permettendo di modellare la variabilità tra gruppi in modo più preciso. Questo approccio consente di catturare l'eterogeneità intrinseca dei dati, in quanto combina le diverse distribuzioni che riflettono le caratteristiche dei gruppi sottostanti, risultando in una rappresentazione più fedele della distribuzione congiunta dei dati osservati. Di conseguenza, le distribuzioni mistura sono strumenti potenti per l'analisi di dati complessi, in cui l'eterogeneità e la presenza di sottogruppi sono elementi chiave da considerare. Questo obiettivo è frequentemente raggiunto identificando vari gruppi all'interno del modello, ciascuno con caratteristiche diverse, e studiando complessivamente questi gruppi quando, come detto in precedenza, i dati non sono classificati.
In queste situazioni, l'analisi e la ricerca si focalizza soprattutto sullo stimare le *proporzioni della mistura* e sulla stima dei parametri della distribuzione. I modelli mistura, infatti, rappresentano i meccanismi generatori dei dati osservati, dove ciascun dato può essere pensato come appartenente a una delle diverse componenti della mistura, corrispondenti a popolazioni distinte. Queste componenti riflettono la variabilità tra sottopopolazioni, spesso modellate con distribuzioni normali, ma adattabili ad altre distribuzioni per meglio rappresentare le caratteristiche specifiche dei dati.
Un ulteriore aspetto significativo delle distribuzioni mistura finite riguarda il fatto che queste possono essere utilizzate in due maniere differenti, definite *applicazioni dirette* e *applicazioni indirette*.

Le *applicazioni dirette* comprendono le situazioni in cui si crede nell'esistenza di c *categorie* sottostanti, dette anche *sorgenti*, in modo che l'unità statistica ω su cui è stata effettuata l'osservazione derivi da una di queste categorie. In questo contesto, la popolazione Ω è considerata composta da un certo numero di gruppi o categorie $\Omega_1, \Omega_2, \ldots, \Omega_c$, ciascuno con la propria distribuzione $f_j(\cdot)$ che descrive la distribuzione di X assumendo che l'unità ω provenga dalla categoria j. In questa particolare forma di applicazioni delle misture, $f_i(\cdot)$ riassume la distribuzione

di probabilità di ω dato che l'osservazione deriva dalla categoria i, e π_i indica la probabilità che l'osservazione provenga da questa sorgente. Questa applicazione è utile in contesti dove esiste una vera eterogeneità nei dati e il compito principale è identificare e analizzare le diverse sorgenti.
Al fine di chiarire questo primo caso, è possibile prendere in considerazione un dataset studiato in Campbell and Mahon (1974) riguardante granchi appartenti alla specie *Leptograpsus*. Ogni osservazione è basata su 5 variabili: ampiezza della bocca (FL), ampiezza dela parte posteriore (RW), lunghezza lungo la linea media (CL), massima ampiezza (CW) e profondità (BD) della carapace. I dati sono stati raggrupati per sesso. Infatti, osservando la fig. 1.1 e la fig. 1.2, è possibile notare la presenza di due gruppi distinti.

Le *applicazioni indirette*, invece, si riferiscono alle situazioni in cui le distribuzioni mistura sono utilizzate come *strumento matematico* per ottenere una forma dell'analisi flessibile e facilmente trattabile. In questo caso, non si presume necessariamente l'esistenza di categorie sottostanti reali, ma le misture sono impiegate per modellare e analizzare i dati in maniera efficace.
Per esempio, una mistura di due normali, in cui una componente presenta una varianza *inflated*, ovvero una delle distribuzioni normali ha una varianza significativamente più grande dell'altra, può essere utilizzata per descrivere densità con code pesanti. Le "code pesanti" si riferiscono a una caratteristica di alcune distribuzioni in cui c'è una probabilità relativamente alta di osservare valori estremi (molto al di fuori della media). Le distribuzioni normali standard hanno code "leggere", il che significa che i valori estremi sono rari. Tuttavia, in molti fenomeni reali, i valori estremi sono più comuni di quanto non sarebbe previsto da una normale distribuzione gaussiana. In una mistura di due normali, se una delle componenti ha una varianza molto grande, questa componente contribuirà a una maggiore dispersione dei dati, producendo valori molto distanti dalla media. Di conseguenza, la distribuzione complessiva (la mistura) avrà una maggiore probabilità di valori estremi rispetto a una singola normale, generando code pesanti. La componente con la varianza maggiore genera valori che si distribuiscono molto più ampiamente attorno alla media, il che aumenta la probabilità di osservare valori estremi. Quando queste due componenti sono mescolate insieme, la distribuzione risultante non seguirà una normale standard ma avrà code pesanti, riflettendo l'effetto della componente con la varianza inflated. Dunque, se usiamo una mistura di normali, una componente con una varianza elevata può catturare i valori estremi (outliers) o la dispersione più ampia, mentre l'altra componente può rappresentare i dati più vicini alla media. La combinazione di queste componenti produce una distribuzione complessiva che meglio riflette la realtà osservata, in particolare per fenomeni che generano code pesanti. Questo tipo di distribuzione può essere utile

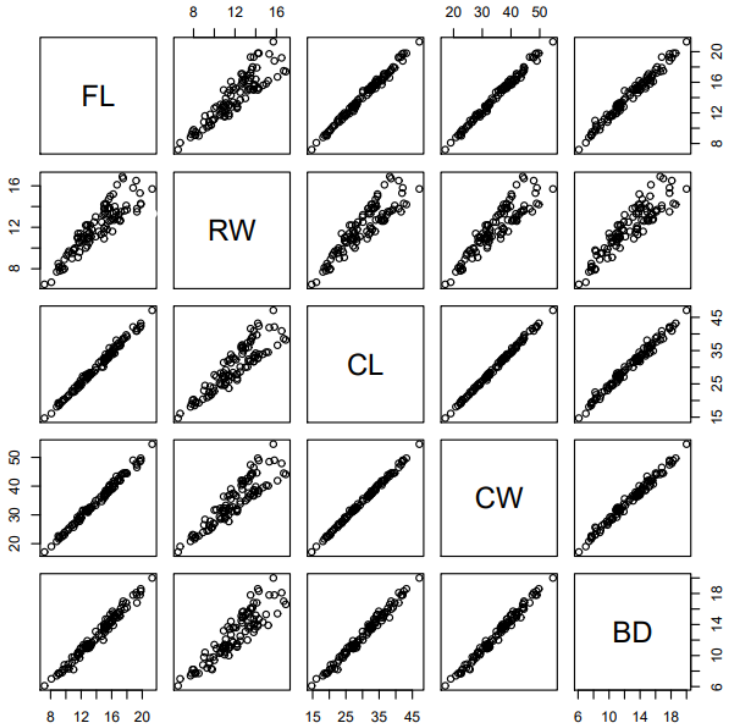


Figura 1.1: Matrice di diagrammi a dispersione del crab dataset

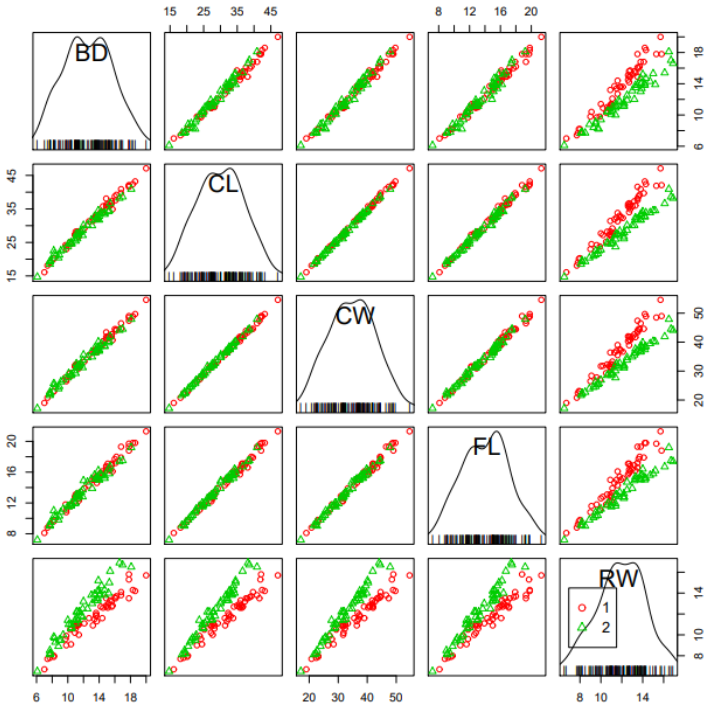


Figura 1.2: Matrice di diagrammi a dispersione del crab dataset con grafico della funzione densità per ciascuna variabile e classificazione dati

in molti contesti dove i fenomeni osservati non seguono una distribuzione normale standard, ma presentano una frequenza di valori estremi più alta del previsto, come nei mercati finanziari (dove i crolli o i picchi di prezzo sono più comuni di quanto una normale potrebbe prevedere) o nei modelli di rischio.
La dicotomia tra queste due applicazioni è essenziale perché influisce sia sul problema di inferenza che sulle tecniche statistiche utilizzate per risolverlo. Nelle applicazioni dirette, l'enfasi è sulla stima delle componenti della mistura e delle probabilità associate, mentre nelle applicazioni indirette l'obiettivo è trovare un modello che meglio si adatta ai dati complessivi. Questa distinzione aiuta a chiarire gli obiettivi di ricerca e le strategie analitiche appropriate per diversi scenari applicativi.

Capitolo 2

Definizioni e teoremi

2.1 Definizione di una mistura di distribuzioni

Nel seguente capitolo sarà fornita una definizione formale delle distribuzioni mistura e saranno trattati alcuni dei principali problemi legati allo studio delle stesse.

Definizione. Sia $f(\mathbf{x};\boldsymbol{\theta})$ una funzione di densità di probabilità d-dimensionale che dipende da un vettore di parametri m-dimensionale $\boldsymbol{\theta}$, e sia $H(\boldsymbol{\theta})$ una funzione di ripartizione m-dimensionale. Allora

$$F(\mathbf{x}) = \int f(\mathbf{x};\boldsymbol{\theta}) dH(\boldsymbol{\theta}) \tag{2.1}$$

è chiamata *densità della mistura*. Inoltre, in altre parole, si può affermare che H denota la misura di probabilità sullo spazio dei parametri.

Osservando questa definizione, è chiaro che $F(\mathbf{x})$ può essere interpretata come una densità marginale di una densità *(d + m)*-variata. Una diretta conseguenza di questa interpretazione è che qualsiasi funzione di densità $F(\mathbf{x})$ può essere vista come una densità di mistura semplicemente immaginando variabili aggiuntive che sono state integrate.
La definizione appena data è del tutto generale e può essere applicata in qualsiasi situazione, ma la maggior parte delle applicazioni delle distribuzioni mistura si occupano di un sottoinsieme di questa definizione generale. Questo sottoinsieme comprende il caso speciale in cui H è discreto e assegna una probabilità positiva solo ad un numero finito di punti $\{\boldsymbol{\theta}_i; i = 1, \ldots, c\}$ che appartengono tutti allo stesso spazio $\boldsymbol{\Theta}$.
Quando siamo in questo caso è possibile sostituire l'integrale in (2.1) con una somma finita per ottenere la *mistura finita*

$$F(\mathbf{x}|\boldsymbol{\Psi}) = \sum_{i=1}^{c} H_i(\boldsymbol{\theta}_i) f_i(\mathbf{x}; \boldsymbol{\theta}_i). \tag{2.2}$$

Gli elementi che costituiscono il vettore dei parametri nella formula (2.2) si possono suddividere in tre tipologie. La prima comprende c, che indica il numero di componenti della mistura finita. La seconda comprende le cosiddette proporzioni (o pesi) della mistura $H_i(\boldsymbol{\theta}_i)$, che saranno denotati con π_i, aventi le seguenti proprietà:

$$\pi_i > 0, \quad i = 1, \ldots, c; \quad \sum_{i=1}^{c} \pi_i = 1.$$

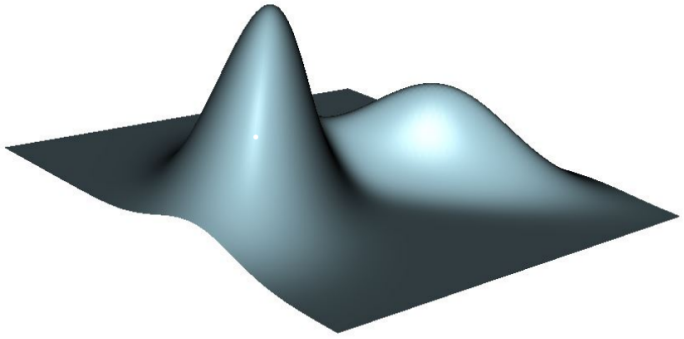


Figura 2.1: Esempio di mistura di 2 densità Normali bivariate

Segue subito che $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$ può essere pensato come un vettore che definisce una distribuzione di probabilità su $\boldsymbol{\Theta}$, con $\pi_i = \mathbb{P}(\boldsymbol{\theta} = \boldsymbol{\theta}_i)$, $i = 1, \ldots, c$. L'ultima categoria comprende i vettori delle componenti dei parametri $\boldsymbol{\theta}_i$, che a priori si suppongono distinti. In molte applicazioni le componenti f_i della densità appartengono alla stessa famiglia, ma, potenzialmente, potrebbero anche essere differenti.
Per comodità, il simbolo $\boldsymbol{\theta}$ indicherà la collezione dei parametri distinti che appaiono nelle componenti della densità e il simbolo $\boldsymbol{\Psi}$ indicherà la collezione di parametri distinti che appaiono nell'intero modello della mistura, cioè $\boldsymbol{\Psi} = (\boldsymbol{\xi}^t, \pi_1, \ldots, \pi_c)^t$, dove $\boldsymbol{\xi}_i$ contiene gli elementi di $\boldsymbol{\theta}_i$.

CHIARIMENTO SULLA NOTAZIONE : negli esempi e nella trattazione che seguirà, verrà adottata la notazione, solitamente più utilizzata, seguente in luogo della (2.2): $f(x|\boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i f_i(x|\boldsymbol{\theta}_i)$.

Esempio di una mistura con due componenti Un esempio significativo e spesso utilizzato di modello di una mistura con due componenti Normali omoschedastiche (dunque $c = 2$) ha la forma:

$$f(x|\boldsymbol{\Psi}) = \pi\phi(x|\mu_1, \sigma) + (1 - \pi)\phi(x|\mu_2, \sigma),$$

in cui $\phi(x|\mu_i, \sigma)$, i = 1,2, denota la densità normale univariata di media μ_i e varianza σ^2. In questo caso si ha che $\pi_1 = \pi$, $\pi_2 = 1 - \pi$, $\boldsymbol{\theta}_1 = (\mu_1, \sigma)$, $\boldsymbol{\theta}_2 = (\mu_2, \sigma)$, $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma)$ e $\boldsymbol{\Psi} = (\pi, \mu_1, \mu_2, \sigma)$.

2.2 Problemi statistici legati alle distribuzioni mistura

Nella seguente sezione, saranno trattati e spiegati brevemente alcuni dei problemi statistici più importanti legati alle distribuzioni mistura.

2.2.1 Modalità di campionamento

In qualsiasi indagine o analisi statistica, il punto di partenza riguarda sempre la modalità di campionamento, cioè, detto in altre parole, le forme attraverso cui i dati sono ottenuti. Nella maggior parte delle applicazioni si parte da una realizzazione $\boldsymbol{X_1} = \boldsymbol{x_1}, \ldots, \boldsymbol{X_n} = \boldsymbol{x_n}$ del cosiddetto *campione casuale*, in cui la distribuzione di ogni $\boldsymbol{X_i}$ è descritta tramite una densità parametrica di mistura finita della forma (2.2). Questo mette in evidenza, infatti, il fatto che operativamente si ha a disposizione un'informazione limitata, dal momento che non è possibile osservare l'intera popolazione, cioè dello spazio campionario, bensì solo qualche realizzazione della stessa, realizzazione che può essere relativa ad una sola variabile o ad una pluralità di variabili.
La grande maggioranza dei metodi statistici parte considerando la *funzione di verosimiglianza*

$$L_0(\boldsymbol{\Psi}) = \prod_{i=1}^{n} f_i(\boldsymbol{x_i}|\boldsymbol{\Psi}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{c} \pi_j f_i(\boldsymbol{x_i}|\boldsymbol{\theta}_j) \right]$$

In alcune applicazioni dirette delle distribuzioni mistura, oltre ad esserci il campione casuale descritto poco sopra, potrebbe essere considerato anche un campione casuale di osservazioni che si sa essere derivato da singole categorie sottostanti. Ad esempio, nello studiare alcune popolazioni ittiche, si potrebbero avere campioni delle lunghezze dei pesci derivanti da pesci di età *nota*, in aggiunta ad un campione di lunghezze delle popolazione della mistura. Questi dati aggiuntivi si possono denotare come

$$\{x_{jh} : j = 1, \ldots, c; h = 1, \ldots, n_j\}$$

dove j indica la categoria (ad esempio, l'età dei pesci), e h è l'indice delle osservazioni in quella categoria. Inoltre, almeno uno degli n_j è diverso da zero. Dunque, avendo una panoramica completa dei dati riguardanti il fenomeno, è possibile riscrivere la funzione di verosimiglianza, tenendo conto sia delle osservazioni classificate che quelle non classificate, nel seguente modo

$$L_1(\mathbf{\Psi}) = L_0(\mathbf{\Psi}) \prod_{j=1}^{c} \prod_{h=1}^{n_j} f_j(x_{jh}|\boldsymbol{\theta}_j).$$

in cui j indica la categoria e h indica l'osservazione in quella categoria.
Inoltre, se si suppone che le osservazioni classificate per categoria siano indipendenti, con pesi $\pi_1, \ldots, \pi_c$ per le singole categorie, questo aggiunge ulteriori informazioni e la funzione di verosimiglianza appropriata risulta

$$L_2(\mathbf{\Psi}) = L_1(\mathbf{\Psi}) \prod_{j=1}^{c} \pi_j^{n_j}.$$

Nella pratica, è fondamentale decidere quale delle tre funzioni di verosimiglianza sia appropriata per uno specifico caso. In particolare, se sono disponibili informazioni riguardanti le osservazioni classificate per categoria, è importante utilizzare $L_1(\mathbf{\Psi})$ piuttosto che $L_0(\mathbf{\Psi})$, dal momento che le informazioni aggiuntive considerate nella prima espressione potrebbero essere sostanziali. Ad esempio, si immagini di aver raccolto dei dati e di avere due tipi di informazioni:

1. **Dati non classificati**: Si ha un campione di lunghezze di pesci, ma non si sa nulla sull'età di questi pesci.

2. **Dati classificati**: Oltre a queste lunghezze generiche, si hanno anche dati specifici su pesci di cui si conosce l'età.

Se si hanno solo dati non classificati (le lunghezze dei pesci senza sapere l'età), si utilizza la funzione $L_0(\mathbf{\Psi})$. Questa funzione permette di stimare i parametri del modello basandosi solo su queste osservazioni generiche. Se, invece, oltre ai dati non classificati, si conosce l'età di alcuni pesci, è preferibile utilizzare $L_1(\mathbf{\Psi})$, perché questa funzione di verosimiglianza tiene conto sia delle osservazioni non classificate che delle osservazioni classificate per età. Questo significa che si stanno sfruttando al massimo tutte le informazioni disponibili, permettendo di fare stime più precise.
D'altro canto, le informazioni riguardanti le osservazioni categoriche sono spesso ottenute selezionando $n_1, \ldots, n_c$, e in questo caso $L_2(\mathbf{\Psi})$ non si può applicare.

Infatti, come detto poco prima, la funzione $L_2(\mathbf{\Psi})$ include un termine che considera i pesi delle categorie $(\pi_j^{n_j})$, il che presuppone che n_j rifletti un processo di campionamento naturale. Tuttavia, se gli n_j vengono scelti intenzionalmente, questo termine non riflette più correttamente la probabilità dei dati osservati. Ad esempio, si supponga di avere tre categorie (ad esempio, età dei pesci: giovani, adulti, anziani). Se si decide di raccogliere esattamente 50 osservazioni per ciascuna categoria, si sta influenzando la composizione del campione. In altre parole, i valori n_1, n_2, n_3 non sono il risultato naturale del processo di campionamento, ma piuttosto una scelta intenzionale. Nella trattazione che segue, utilizzando la notazione di Hosmer, saranno denotate con $M0, M1$ ed $M2$ le tre strutture dati che danno origine a L_0, L_1 ed L_2 rispettivamente.

2.2.2 Numero di componenti

In alcuni contesti, l'incertezza sul numero delle componenti fa sorgere numerosi problemi statistici correlati alla *cluster analysis gerarchica*, ovvero quella tecnica di analisi dei dati che raggruppa un insieme di oggetti in sottoinsiemi, chiamati *cluster*, in modo che gli oggetti all'interno di ciascun cluster siano simili tra loro e diversi dagli oggetti in altri cluster, solitamente con assunzioni forti sulla struttura parametrica. Ad esempio, solitamente si assume che tutte le densità facciano parte della famiglia della distribuzione normale (univariata o multivariata in base al caso in questione).
In altri casi, una grande quantità di dati è disponibile direttamente dalla mistura, il che permette di stimare correttamente la forma di $f(x)$ in (2.2). Fornendo assunzioni parametriche sulle componenti sottostanti, il problema di diventa un problema di *curve fitting*, ovvero trovare la migliore rappresentazione di una serie di dati attraverso una funzione matematica, in modo che la funzione si adatti il più possibile ai dati osservati.
A volte, si è interessati a trovare la mistura con il minor numero di componenti in modo tale che questa si adatti in maniera soddisfacente ai dati. Molto spesso, infatti, ci si chiede se nel modello considerato si deve assumere l'esistenza di due componenti sottostanti o soltanto di una componente.
Effettivamente, quando si è in presenza di un modello di misture in cui si assume che le componenti della densità siano in forma parametrica, un'ipotetica mistura avente meno componenti può essere interpretata come imposizione di un'*ipotesi nulla* sul modello originario. Il problema di confrontare i due modelli mistura tramite un *test d'ipotesi*, a questo punto, sembra essere riconducibile ad un problema di *test tra ipotesi annidate*.
Tuttavia, risulta subito evidente che i modelli classici di test non sono così facili da applicare in questo contesto. Si consideri, ad esempio, il problema, apparentemente semplice, di testare le seguenti ipotesi:

$$H_0: \quad p(x) = \phi(x|\mu, \sigma)$$
$$H_1: \quad p(x) = \pi\phi(x|\mu_1, \sigma_1) + (1-\pi)\phi(x|\mu_2, \sigma_2)$$

in cui $x \in \mathbb{R}$ e i parametri sotto H_0 e H_1 sono sconosciuti. In altre parole, ci si sta chiedendo se vi è una sola componente normale oppure due nel modello considerato.
Pensando alle procedure tradizionali con cui si affrontano questo tipo di problemi, è naturale pensare al *rapporto di verosimiglianza generalizzato*, soprattutto se si dispone di un grande campione, facendo riferimento ad un importantissimo risultato fornito dal Teorema di Wilks. Sorge, però, subito un problema, dal momento che non vi è un modo univoco per ottenere H_0 da H_1. Un esempio sarebbe quello di imporre

$$\pi = 0 \quad (1 \text{ vincolo})$$

oppure

$$\mu_1 = \mu_2; \quad \sigma_1 = \sigma_2 \quad (2 \text{ vincoli}).$$

Il confronto tra modelli mistura tramite test di ipotesi rivela una notevole complessità, che richiede approcci più sofisticati per gestire le ipotesi annidate. I modelli mistura spesso implicano un gran numero di parametri e strutture complesse, il che rende difficile applicare i tradizionali test statistici. In particolare, l'interpretazione delle ipotesi annidate richiede un'analisi approfondita delle configurazioni parametriche, tenendo conto dei vincoli specifici imposti dalle ipotesi alternative. Quando si confrontano modelli mistura, è essenziale esplorare come diverse configurazioni parametriche influenzano i risultati. Per esempio, si potrebbe dover considerare se una ipotesi alternativa implica vincoli particolari sui parametri, come l'uguaglianza delle medie o delle varianze tra le componenti di una mistura. Questi vincoli possono complicare l'applicazione dei test di ipotesi tradizionali, che spesso non sono progettati per gestire la complessità intrinseca dei modelli mistura. Inoltre, i modelli mistura sono sensibili alla scelta dei vincoli sui parametri e alla definizione delle ipotesi. L'approccio tradizionale potrebbe non essere adeguato per catturare la variabilità e la struttura dei dati in modo accurato, richiedendo quindi metodi statistici più avanzati e flessibili. Questi metodi devono essere in grado di valutare i modelli mistura con precisione e di accettare o rifiutare le ipotesi in modo robusto, tenendo conto delle complessità strutturali e parametriche. In conclusione, la gestione e la valutazione dei modelli mistura richiedono strumenti statistici che superino le limitazioni dei test tradizionali. È necessario adottare metodi che possano affrontare le sfide uniche poste dai modelli complessi e dalle ipotesi annidate, garantendo così una valutazione accurata e significativa dei modelli e delle ipotesi associate.

2.2.3 Parametri sconosciuti

Assumendo un valore noto di c, è molto frequente, in una formulazione parametrica, la necessità di dover fare inferenza sui parametri sconosciuti del modello della mistura. Potrebbero sorgere diversi casi differenti.
In alcune applicazioni dirette, è possibile sviluppare studi approfonditi e dettagliati delle singole componenti della distribuzione separatamente dal problema della mistura. Nel contesto di distribuzioni mistura della forma

$$f(x|\boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i f_i(x|\boldsymbol{\theta}_i)$$

se le componenti $f_1(\cdot|\boldsymbol{\theta}_1), \ldots, f_c(\cdot|\boldsymbol{\theta}_c)$ fossero conosciute, allora i problemi di inferenza sarebbero stati legati soltanto ai pesi $\pi_1, \ldots, \pi_c$ della mistura.
In altre applicazioni, invece, potrebbero sollevarsi dubbi e incertezze circa i parametri $\boldsymbol{\theta}_i$, considerando invece i pesi π_i noti. In questo caso, dunque, si avrebbe un problema con pesi π_i noti, ma parametri $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c$ incogniti.
È ovvio che potrebbero presentarsi situazioni in cui sia i pesi π_i che i parametri $\boldsymbol{\theta}_i$ sono incogniti e, infatti, questo è il caso più frequente nelle diverse applicazioni delle misture.
Una volta che si è deciso l'approccio parametrico (ovviamente, esistono approcci non parametrici nel caso in cui le densità delle componenti siano sconosciute) esiste una grande quantità di metodi importanti volti allo studio e alla risoluzione dei problemi di stima dei parametri in un modello di misture finite.
In particolare, i metodi più utilizzati nelle applicazioni, a seconda della loro semplicità e adattamento in un caso specifico, sono il *metodo dei momenti, metodo di massima verosimiglianza, minimo* χ^2*, minimi quadrati* e *metodi Bayesiani.*

2.3 Identificabilità di una mistura

L'ipotesi di identificabilità, ovvero, in altre parole, l'univoca identificazione della mistura, è il cuore della gran parte della teoria e pratica della statistica, dal momento che soltanto in presenza di questa si ha stimabilità. Infatti, nel caso in cui non fosse identificabile, la mistura potrebbe essere compatibile con molte distribuzioni e, una volta stimati i parametri, non si saprebbe a quale distribuzione questi facciano riferimento. Inoltre, il problema dell'identificabilità è alquanto serio nel momento in cui lo scopo ultimo dello studio è la conoscenza delle singole componenti, ad esempio, nel caso in cui si vogliano classificare osservazioni future in una delle classi di cui sappiamo essere composta la distribuzione.

2.3.1 Definizione di identificabilità

Un'importante proprietà di una famiglia parametrica di distribuzioni $\mathcal{F} = \{f(\cdot;\theta)\}_{\theta\in\Theta}$ indicizzata da un parametro $\theta \in \Theta\mathcal{F} = \{f(\cdot;\theta)\}_{\theta\in\Theta}$ su un certo spazio campionario $\mathcal{X}$, è costituita dalla identificabilità. Diremo che la famiglia $\mathcal{F}$ è *identificabile* se e solo se, considerati due qualunque parametri $\theta_1, \theta_2 \in \Theta$, con $\theta_1 \neq \theta_2$, allora le due distribuzioni indicizzate da θ_1 e θ_2 sono diverse; in particolare risulta $f(\mathbf{x};\theta_1) \neq f(\mathbf{x};\theta_2)$ tranne al più per un insieme di punti $\mathbf{x} \in \mathcal{X}$ di misura nulla. In generale, si adotta la seguente definizione di identificabilità per distribuzioni mistura. Siano

$$f(\mathbf{x}|\psi) = \sum_{i=1}^{c} \alpha_i f_i(\mathbf{x}|\theta_i) \quad \text{e} \quad f(\mathbf{x}|\psi^*) = \sum_{i=1}^{c^*} \alpha_i^* f_i(\mathbf{x}|\theta_i^*)$$

due qualunque membri di una famiglia parametrica di densità mistura. Una mistura finita si dice *identificabile* rispetto a $\psi \in \Psi$ se risulta

$$f(\mathbf{x}|\psi) = f(\mathbf{x}|\psi^*)$$

se e solo se $c = c^*$ e se possiamo permutare gli indici delle componenti in modo tale che si abbia

$$\alpha_i = \alpha_i^* \quad \text{e} \quad f_i(\mathbf{x}|\theta_i) = f_i(\mathbf{x}|\theta_i^*) \quad i = 1, \ldots, c.$$

In particolare, se una mistura è identificabile, ci sono tante equazioni indipendenti quanti sono i parametri da stimare e dunque, avendo un sistema deterministico, è possibile ottenere le stime cercate.

2.3.2 Esempi in cui appare il problema dell'identificabilità

Esempio. Sia **B** la classe a cui appartiene la distribuzione binomiale $Bi(2, \theta)$, con θ che indica la probabilità di successo, dunque $0 < \theta < 1$ e 2 il numero di prove. Consideriamo un'arbitraria mistura di due componenti di **B**. Allora

$$f(0) = \pi(1-\theta_1)^2 + (1-\pi)(1-\theta_2)^2$$

e

$$f(1) = 2\pi\theta_1(1-\theta_1) + 2(1-\pi)\theta_2(1-\theta_2),$$

in cui π denota il peso della mistura. Ci sarebbe una terza equazione, per $p(2)$, ma è ovviamente una combinazione lineare delle due precedenti e, dunque, non è indipendente da esse. Quindi, date le probabilità della mistura $p(0), p(1)$, la loro rappresentazione in termini di $(\pi, \theta_1, \theta_2)$ non è chiaramente unica. Al fine di risolvere questa ambiguità per ottenere una soluzione unica, è necessario introdurre ulteriori vincoli sul modello. Una possibilità, ad esempio, sarebbe quella di imporre delle restrizioni sui valori che possono assumere θ_1 e θ_2.

Esempio. Si consideri, per esempio, una distribuzione di Bernoulli composta da una mistura di due componenti di Bernoulli con funzioni di probabilità

	$P(x=0)$	$P(x=1)$
Componente 1	$f_1(0;\theta_1)$	$f_1(1;\theta_1)$
Componente 2	$f_2(0;\theta_2)$	$f_2(1;\theta_2)$

Allora:

$$f(x) = \pi_1 f_1(x;\theta_1) + \pi_2 f_2(x;\theta_2)$$

Sfortunatamente, è possibile determinare un'unica equazione

$$f(0) = \pi_1 f_1(0;\theta_1) + \pi_2 f_2(0;\theta_2)$$

dall'equazione generale dal momento che $f(1) = 1 - f(0)$. Nel modello, però, ci sono tre parametri indipendenti da stimare, ovvero π_1, $f_1(0;\theta_1)$, $f_2(0;\theta_2)$, dunque non è possibile determinare univocamente i parametri. Si noti, inoltre, che $\pi_2 = 1 - \pi_1$, $f_1(1;\theta_1) = 1 - f_1(0;\theta_1)$ e $f_2(1;\theta_1) = 1 - f_2(0;\theta_1)$.

Esempio. In ultima battuta, si supponga di avere una serie di osservazioni provenienti da una mistura di due distribuzioni uniformi univariate e che lo scopo dello studio sia quello di analizzare la mistura nelle sue diverse componenti. Supponiamo che un'analisi produca la mistura

$$f(x) = \frac{1}{3}U(-1,1) + \frac{2}{3}U(-2,2)$$

in cui $U(a, b)$ indica una distribuzione uniforme di range (a, b). Tuttavia, dal momento che sono stati forniti solo alcuni campioni della mistura e non per ogni componente separatamente, non c'è alcun modo di affermare con sicurezza che la precedente decomposizione sia quella corretta, oppure che lo sia un'altra, ad esempio

$$f(x) = \frac{1}{2}U(-2, 1) + \frac{1}{2}U(-1, 2).$$

2.3.3 Teorema di Yakowitz e Spragins

Il problema dell'identificabilità di una mistura di distribuzioni, studiato inizialmente da Titterington, è stato risolto per le distribuzioni continue da Yakowitz e Spragins (1968) che hanno fornito condizioni necessarie e sufficienti utili a mostrare quali distribuzioni producono misture finite identificabili.

Teorema (Yakowitz e Spragins)**.** Una condizione necessaria e sufficiente affinchè la classe di tutte le misture finite dell'insieme

$$\{G(\mathbf{x}; \boldsymbol{\theta}); \mathbf{x} \in \mathbb{R}^d, \boldsymbol{\theta} \in \mathbb{R}^m\}$$

sia identificabile è che questo insieme sia linearmente indipendente rispetto al campo dei numeri reali $\mathbb{R}$.

Dimostrazione. (*Condizione necessaria*) Si inizi assumendo per assurdo che gli elementi $G(\mathbf{x}; \boldsymbol{\theta}_i)$ non siano linearmente indipendenti, dunque, per definizione, esiste una qualche funzione lineare

$$\sum_{i=1}^{c} p_i G(\mathbf{x}; \boldsymbol{\theta}_i) = 0$$

con $G(\mathbf{x}; \boldsymbol{\theta}_i) \neq G(\mathbf{x}; \boldsymbol{\theta}_j)$ se $i \neq j$. Senza perdita di generalità si può assumere che gli elementi p_i siano ordinati in modo tale che $p_i < 0$ se e solo se $i \leq c'$, con c' generico.
Allora

$$\sum_{i=1}^{c'} |p_i| G(\mathbf{x}; \boldsymbol{\theta}_i) = \sum_{i=c'+1}^{c} |p_i| G(\mathbf{x}; \boldsymbol{\theta}_i)$$

Poiché le funzioni $G(\mathbf{x}; \boldsymbol{\theta}_i)$ sono funzioni di ripartizione, per definizione si ha che $G(\infty; \boldsymbol{\theta}_i) = \lim_{\mathbf{x} \to +\infty} G(\mathbf{x}; \boldsymbol{\theta}_i) = 1$ e dunque

$$\sum_{i=1}^{c'} |p_i| = \sum_{i=c'+1}^{c} |p_i| = b > 0.$$

A questo punto, definendo $a_i := \frac{|p_i|}{b}$, si ottiene

$$\sum_{i=1}^{c'} a_i G(\mathbf{x}; \boldsymbol{\theta}_i) = \sum_{i=c'+1}^{c} a_i G(\mathbf{x}; \boldsymbol{\theta}_i)$$

che sono entrambe misture finite e, tuttavia, distinte. Dunque, non sono identificabili. Perciò, avendo ottenuto un assurdo, la condizione di necessità è verificata.

(Condizione sufficiente) Si supponga che la classe delle distribuzioni mistura finite definita precedentemente sia non identificabile. Grazie alla definizione, questo significa che esisterebbero alcune misture tali che

$$\sum_{i=1}^{c} p_i G(\mathbf{x}; \boldsymbol{\theta}_i) = \sum_{j=1}^{c'} p_j G(\mathbf{x}; \boldsymbol{\theta}_j)$$

in cui non tutti i p_i sono uguali ad alcuni p_j e/o non tutti i $\boldsymbol{\theta}_i$ sono uguali ad alcuni $\boldsymbol{\theta}_j$. L'equivalenza scritta nella riga precedente potrebbe essere riscritta, definendo $c'' := c + c'$, nel seguente modo

$$\sum_{i=1}^{c} p_i G(\mathbf{x}; \boldsymbol{\theta}_i) - \sum_{j=1}^{c'} p_j G(\mathbf{x}; \boldsymbol{\theta}_j) = \sum_{j=1}^{c''} p_j G(\mathbf{x}; \boldsymbol{\theta}_j) = 0$$

con nessun $G(\mathbf{x}; \boldsymbol{\theta}_i)$ uguale a nessun $G(\mathbf{x}; \boldsymbol{\theta}_j)$ e nessun $p_i = 0$. Questo, per definizione, vuol dire che i $G(\mathbf{x}; \boldsymbol{\theta}_i)$ non sono linearmente indipendenti. Esattamente come per la condizione necessaria, anche la condizione sufficiente è dimostrata. □

Capitolo 3

Metodi di stima

Uno dei problemi principali legati alle distribuzioni mistura riguarda la stima dei parametri delle misture stesse.
La maggior parte del capitolo riguarderà il metodo di massima verosimiglianza, per poi giungere all'algoritmo EM. Inoltre, nel seguente capitolo, sarà dato un cenno al metodo dei momenti, sia per ragioni storiche, in quanto è stato il primo metodo proposto da Pearson nel 1894, ma anche perchè il metodo dei momenti può essere utilizzato come meccanismo per inizializzare il processo di stima che verrà completato dall'algoritmo EM.
È bene specificare, però, che i metodi di stima dei parametri di una mistura di distribuzioni non si esauriscono ai pochi citati sopra, bensì esistono altri metodi del tutto validi per affrontare il problema, quali metodi qualitativi grafici, metodi Bayesiani, stima della minima distanza basata sulle funzioni di distribuzioni e decomposizioni numeriche delle misture. Nel seguito vengono presentati un teorema e una definizione che saranno citati successivamente.

Definizione. Sia X una variabile casuale avenete funzione di ripartizione $F_X(x;\theta)$ e sia $(X_1,\dots,X_n)$ un campione casuale da quest'ultima distribuzione. Sia $T_n = T_n(X_1,\dots,X_n)$ uno stimatore di θ. Diremo T_n stimatore *consistente* di θ se

$$T_n \xrightarrow{\mathrm{P}} \theta.$$

Teorema (Legge (debole) dei Grandi Numeri)**.** Sia $(X_1,\dots,X_n)$ una successione di variabili indipendenti e identicamente distribuite di media μ e supponiamo che le X_i ammettano momento secondo. Definita $Y_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ si ha

$$\lim_{n\to+\infty} P(|\overline{X}_n - \mu| \geq \epsilon) = 0, \text{per ogni } \epsilon > 0$$

In altre parole, $\overline{X}_n \xrightarrow{\mathrm{P}} \mu$.

3.1 Cenno al metodo dei momenti

Innanzitutto, è utile ricordare la definizione di momento campionario.

Definizione (Momento campionario di ordine s)**.** Sia $(X_1, \ldots, X_n)$ un campione casuale proveniente da una distribuzione $F_X(x;\theta)$ con $\theta \in \Theta \subseteq \mathbb{R}^k$ e sia $s \in \mathbb{N}$. Si definisce *momento campionario (non centrato) di ordine s* la quantità

$$m'_s = \frac{1}{n}\sum_{i=1}^{n} X_i^s$$

Ponendo i primi k momenti della popolazione $\mu'_1, \ldots, \mu'_k$ uguali ai corrispondenti momenti campionari si ricaverà un sistema di k equazioni

$$\mu'_s = m'_s$$

per $s = 1, \ldots, k$ e, poichè i momenti sono funzioni di $\theta = (\theta_1, \ldots, \theta_k)$, si avrà un sistema di k equazioni in k incognite la cui soluzione fornisce una stima di θ detta *stima con il metodo dei momenti.*

Si supponga di avere un insieme di dati di n osservazioni indipendenti che derivano da una popolazione il cui modello di probabilità dipende da r parametri sconosciuti, che saranno indicati con $\boldsymbol{\Psi}$.
Si supponga che $\boldsymbol{\mu(\Psi)}$ denoti un vettore di r momenti funzionalmente indipendenti e che $\mathbf{m}$ denoti il corrispondente insieme di momenti del campione. Lo *stimatore del metodo dei momenti* corrisponde a quel $\hat{\boldsymbol{\Psi}}$ che soddisfa

$$\boldsymbol{\mu}(\hat{\boldsymbol{\Psi}}) = \mathbf{m}. \tag{3.1}$$

In generale, però, sorgono una serie di potenziali problemi con gli stimatori dei momenti. Infatti:

(a) Una soluzione esplicita dell'equazione (3.1) potrebbe non essere facile da trovare o, a volte, nemmeno possibile.

(b) La soluzione dell'equazione (3.1) potrebbe non essere unica e potrebbe non trovarsi in una regione di $\mathbb{R}^r$.

(c) Nonostante la consistenza di $\boldsymbol{\mu}(\hat{\boldsymbol{\Psi}})$ e, conseguentemente, in alcuni casi, la consistenza di $\hat{\boldsymbol{\Psi}}$ dovuta dalla *Legge dei Grandi Numeri*, $\hat{\boldsymbol{\Psi}}$ potrebbe non essere asintoticamente efficiente.

(d) Il calcolo esatto della $\text{cov}(\hat{\boldsymbol{\Psi}})$ non è generalmente possibile. Tuttavia, un'espansione di Taylor potrebbe essere utilizzata al fine di dimostrare che, approsimativamente e per grandi campioni,

$$\boldsymbol{\mu}(\boldsymbol{\Psi_0}) + \mathbf{D}(\boldsymbol{\Psi_0})(\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi_0}) = \mathbf{m} \tag{3.2}$$

dove $\mathbf{D}$ indica la matrice quadrata delle derivate degli elementi in $\boldsymbol{\mu}$ e $\boldsymbol{\Psi_0}$ indica il valore esatto di $\boldsymbol{\Psi}$. Quindi, approssimativamente, si ha che

$$\begin{aligned} \text{cov}(\hat{\boldsymbol{\Psi}}) &= \mathbf{D}(\boldsymbol{\Psi}_0)^{-1}\text{cov}_{\boldsymbol{\Psi}_0}(\mathbf{m})[\mathbf{D}^t(\boldsymbol{\Psi}_0)]^{-1} \\ &\approx \mathbf{D}(\hat{\boldsymbol{\Psi}})^{-1}\text{cov}_{\hat{\boldsymbol{\Psi}}}(\mathbf{m})[\mathbf{D}^t(\hat{\boldsymbol{\Psi}})]^{-1}. \end{aligned} \tag{3.3}$$

Esempio (Mistura di due distribuzioni note)**.** Si conderi la mistura

$$f(x) = \pi f_1(x) + (1-\pi)f_2(x), \quad 0 < \pi < 1.$$

Sia $t(X)$ tale che $\mathbb{E}[t(X)]$ esiste per ogni componente f_1 e f_2 della densità e si denotino con μ_{1t} e μ_{2t} il valore atteso di $t(X)$ rispetto alle due componenti. Analogamente, indicando il momento campionario della mistura con m_t, si ottiene lo stimatore esplicito del metodo dei momenti per π, basato su t dalla formula

$$m_t = \hat{\pi}\mu_{1t} + (1-\hat{\pi})\mu_{2t},$$

che implica

$$\hat{\pi} = (m_t - \mu_{2t})/(\mu_{1t} - \mu_{2t}).$$

Si osservi, inoltre, che $\hat{\pi}$ è uno stimatore non distorto di π e la sua varianza è data da

$$\begin{aligned} \text{Var}(\hat{\pi}) &= \text{Var}(m_t)/(\mu_{1t} - \mu_{2t})^2 \\ &= \frac{1}{n}\text{Var}(t)/(\mu_{1t} - \mu_{2t})^2 \end{aligned}$$

in cui $Var(t)$ può essere scritta in termini dei primi due momenti di t corrisponenti alle due componenti della densità.

3.2 Metodo di massima verosimiglianza

In questa sezione sarà introdotto il metodo di massima verosimiglianza per la stima dei parametri di distribuzioni statistiche, supponendo di lavorare sempre con problemi regolari. Nella sezione 2.2.1, è stata già citata la funzione di verosimiglianza a proposito di problemi statistici legati alle distribuzioni mistura. Questa funzione, infatti, soprattutto per modelli parametrici semplici, fornisce un approccio al problema della stima dei parametri molto popolare per diverse ragioni, tra le quali è utile ricordare l'esistenza e il supporto della teoria asintotica.
Infatti, il metodo di massima verosimiglianza possiede proprietà statistiche molto utili. Ad esempio, sotto determinate condizioni, lo stimatore ottenuto dal metodo di massima verosimiglianza è consistente, ovvero converge in probabilità ai valori reali dei parametri stimati, e, inoltre, è asintoticamente normalmente distribuito, il che conferisce allo stimatore enormi proprietà utili a fare inferenza sullo stesso.
Dunque, considerato un campione casuale di n osservazioni indipendenti derivanti dalla mistura, la funzione di verosimiglianza è data dalla quantità

$$L_0(\mathbf{\Psi}) = \prod_{i=1}^{n} f(x_i|\mathbf{\Psi}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{c} \pi_j f_i(x_i|\boldsymbol{\theta}_j) \right] \tag{3.4}$$

in cui $\mathbf{\Psi}$ indica il vettore di parametri che si vuole stimare.
In quest'ottica, si potrebbe affermare che la funzione di verosimiglianza misura la probabilità che che diversi $\mathbf{\Psi}$ abbiano dato origine al campione osservato $\mathbf{X}$.
Adesso, guardando la funzione di verosimiglianza come funzione di $\mathbf{\Psi}$ piuttosto che funzione delle $\boldsymbol{x_i}$, è immediato voler cercare un valore $\mathbf{\Psi}_0$ particolare tale che la massimizzi. Solitamente, però, per semplicità di calcoli, si cerca di massimizzare la *funzione di log-verosimiglianza* data da $\mathscr{L}_0(\mathbf{\Psi}) = log(L_0(\mathbf{\Psi}))$.
Per molti problemi relativi alla stima dei parametri, dal momento che si suppone di avere un probema regolare, è possibile affrontare il problema della massimizzazione della funzione di verosimiglianza nel modo più classico possibile, ovvero differenziando L_0 rispetto alle componenti di $\mathbf{\Psi}$ e ponendo le derivate uguali a zero per ottenere le così dette *equazioni normali*

$$\frac{\partial \mathscr{L}}{\partial \Psi_i} = 0.$$

Queste equazioni sono risolte nelle variabili Ψ_i e le derivate del secondo ordine sono esaminate per verificare che la soluzione ottenuta sia effettivamente un massimo e non un punto stazionario qualunque.
Sfortunatamente, però, per le distribuzioni mistura le cose sono un po' più complicate. I problemi principali che si riscontrano con le distribuzioni mistura sono i seguenti. Innanzitutto, per le distribuzioni mistura, le equazioni normali non sono

sempre risolvibili esplicitamente nei parametri Ψ_i, ovvero non sempre esiste una soluzione analitica alle equazioni normali poichè potrebbero essere equazioni non linerari nei parametri, e dunque vengono applicate tecniche iterative di analisi numerica per la risoluzione delle equazioni, come le tecniche di *hill-climbing*, le quali sono utilizzate per massimizzare o minimizzare una funzione obiettivo, muovendosi iterativamente in direzione dei valori che migliorano l'obiettivo, che nel seguente caso è la massimizzazione della funzione di verosimiglianza.
In seconda battuta, un ulteriore problema potrebbe essere legato al fatto che, nel caso di distribuzioni mistura, la funzione di verosimiglianza, come anche le derivate parziali $\frac{\partial \mathscr{L}}{\partial \Psi_i}$, non sono limitate.
Per maggiore comprensione dell'argomento, in seguito vengono forniti due esempi, il primo riguardante una mistura di due densità note e il secondo riguardante una mistura di due densità normali univariate.

Esempio 3.2.1 (Mistura di due densità note)**.** Si consideri

$$\begin{aligned}\mathscr{L}_0(\mathbf{\Psi}) = \mathscr{L}_0(\pi) &= \sum_{i=1}^{n} \log[\pi f_1(x_i) + (1-\pi) f_2(x_i)] \\ &= \sum_{i=1}^{n} \log[\pi(f_{i1} - f_{i2}) + f_{i2}]\end{aligned}$$

dove $f_{ij} = f_j(x_i) \quad j = 1, 2; i = 1, \ldots, n.$
Se si considera $p_i = \pi f_1(x_i) + (1-\pi) f_2(x_i)$ allora l'equazione di verosimiglianza diventa

$$0 = \frac{\partial \mathscr{L}_0}{\partial \pi} = \sum_{i=1}^{n} \frac{f_{i1} - f_{i2}}{p_i}. \tag{3.5}$$

Ci sono due aspetti preoccupanti circa la soluzione dell'equazione (3.5). Il primo riguarda il fatto che la (3.5) è equivalente ad un'equazione polinomiale di grado fino a $(n-1)$ in π. Ad ogni modo, vi è almeno una soluzione reale per l'equazione (3.5) dal momento che la funzione $\mathscr{L}$ è concava. Infatti:

$$\frac{\partial^2 \mathscr{L}_0}{\partial \pi^2} = -\sum_{i=1}^{n} \left[\frac{f_{i1} - f_{i2}}{p_i} \right]^2 < 0.$$

Il secondo problema, invece, si presenta dal momento che la soluzione $\hat{\pi}$ di (3.5) potrebbe non soddisfare la richiesta $0 \leq \hat{\pi} \leq 1$, dunque lo stimatore di massima verosimiglianza di π è

(a) $\hat{\pi}$ se $0 \leq \hat{\pi} \leq 1$

(b) 0 se $\left. \frac{\partial \mathscr{L}_0}{\partial \pi} \right|_{\pi=0} < 0$

(c) 1 se $\left.\frac{\partial \mathscr{L}_0}{\partial \pi}\right|_{\pi=0} > 0$

Sebbene l'esistenza di una radice reale dell'equazione (3.5) sia molto utile, trovare una soluzione esplicita non è sempre possibile e, infatti, nella maggior parte dei casi si ricorre a metodi numerici, come il *Metodo di Newton-Rapshon*. Un ulteriore procedimento iterativo consiste nel sostituire f_{i2} in funzione di p_i e f_{i1} nella (3.5) e riarrangiare l'equazione per ottenere

$$\pi = \frac{1}{n}\sum_{i=1}^{n} \pi f_{i1}/p_i = \frac{1}{n}\sum_{i=1}^{n} w_{i1}(\pi)$$

dove $w_{i1}(\pi)$ è chiaramente tra 0 e 1. Questo conduce subito alla seguente procedura iterativa, data da

$$\pi^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{i1}(\pi^{(m)}) \quad m = 0, 1, \ldots. \tag{3.6}$$

Questa procedura sarà ripresa nella sezione successiva.

Esempio 3.2.2 (Mistura di due densità normali univariate)**.** Nel seguente caso, la funzione di (log)-verosimiglianza è data da

$$\begin{aligned}\mathscr{L}_0(\mathbf{\Psi}) &= \mathscr{L}_0(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \\ &= \sum_{i=1}^{n} \log[\pi\phi(x_i|\mu_1, \sigma_1) + (1-\pi)\phi(x_i|\mu_2, \sigma_2)]\end{aligned}$$

A questo punto, non è difficile notare che l'insieme delle equazioni di verosimiglianza non può essere risolto esplicitamente. Nonostante questo, ci sono problemi ancora più gravi. Sebbene questo sia il modello di misture più comune nelle applicazioni, la superficie di verosimiglianza risultante è piena di singolarità. Infatti, si supponga, ad esempio, che $\mu_1 = x_1$. Dunque, quando $\sigma_1 \to 0$, si ha che $\mathscr{L}_0(\cdot) \to \infty$. In conclusione, quindi, si avrà una moltitudine di massimi globali 'inutili'. Esistono, tuttavia, strategie per rendere la massima verosimiglianza ancora praticabile, a patto di tenere le varianze lontane dallo zero. Spesso, nella pratica, questo accade automaticamente richiedendo $\sigma_1 = \sigma_2$ oppure se si è in presenza di insiemi supplementari con dati categorici (come le strutture dati M1 ed M2 viste nella sezione 2.2.1) contententi almeno due osservazioni distinte per le due sottopopolazioni.

È interessante notare come l'iterazione definita dalla formula (3.6) è molto simile alla versione dell'Esempio 3.2.1 in cui i dati sono completamente classificati. Infatti, quest'ultima consiste nel definire $w_{i1}(\pi^{(m)})$ come zeri o uni.

In generale, i dati completi, ovvero classificati (dunque $\boldsymbol{Z}$ è noto), possono essere rappresentati tramite

$$\{y_i, i = 1, \ldots, n\} = \{(x_i, \boldsymbol{z}_i); i = 1, \ldots, n\},$$

in cui ogni $\boldsymbol{z}_i = (z_{ij}, j = 1, \ldots, c)$ è un vettore indicatore di lunghezza c con 1 nella posizione corrispondente alla propria categoria e 0 nelle restanti. La verosimiglianza corrispondente a $(y_1, \ldots, y_n)$ può essere scritta nella forma

$$g(y_1, \ldots, y_n|\boldsymbol{\Psi}) = \prod_{i=1}^{n} \prod_{j=1}^{c} \pi^{z_{ij}} f_j(x_i|\boldsymbol{\theta_j})^{z_{ij}} \tag{3.7}$$

con logaritmo

$$l_0(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \boldsymbol{z_i^T V(\pi)} + \sum_{i=1}^{n} \boldsymbol{z_i^T U_i(\theta)} \tag{3.8}$$

in cui $\boldsymbol{V(\pi)}$ ha la j-ma componente $\log(\pi_j)$ e $\boldsymbol{U_i(\theta)}$ ha la j-ma componente $\log f_j(x_i|\boldsymbol{\theta_j})$.

La funzione di verosimiglianza $L_0(\boldsymbol{\Psi})$ nella (3.4) è ottenuta sommando la funzione di log-verosimiglianza per ciascuna osservazione y_i, come indicato nell'equazione (3.7), su tutti i possibili vettori $\boldsymbol{z_i}$. Questo processo corrisponde a considerare tutti i possibili modi in cui i dati potrebbero essere assegnati alle diverse componenti della mistura, tenendo conto delle incertezze associate ai dati mancanti rappresentati dai vettori indicatori $\boldsymbol{z_i}$. Questa interpretazione suggerisce che i modelli mistura trattano i dati come incompleti, con le categorie non osservate che possono contribuire alla generazione dei dati. Di conseguenza, quando si utilizzano modelli mistura, è necessario tener conto di tutte le possibili configurazioni dei dati mancanti e considerare la somma su tutte queste configurazioni per ottenere la verosimiglianza complessiva del modello. Inoltre, questa interpretazione evidenzia che quando ci sono dati mancanti, la stima dei parametri del modello mediante massima verosimiglianza può diventare più complessa, poiché bisogna tener conto delle molteplici configurazioni dei dati mancanti durante il processo di stima. Viene in aiuto, tuttavia, una classe di algoritmi iterativi denominati EM,che sarà descritta nella sezione successiva, la cui idea è quella di "riempire" i dati mancanti stimando le probabilità di appartenenza (quindi completando virtualmente $\boldsymbol{Z}$) e poi usare queste stime per aggiornare i parametri del modello. Per comprendere al meglio l'utilizzo del vettore $\boldsymbol{Z}$ sopra citato viene fornito il seguente esempio.

Esempio 3.2.3. Si immagini di avere un insieme di dati che proviene da una popolazione che è in realtà una *mistura* di due sotto-popolazioni. Supponiamo che

ciascuna sotto-popolazione segua una distribuzione normale, ma con parametri (media e varianza) diversi.

Le due distribuzioni normali hanno:

- **Distribuzione 1**: media $\mu_1 = 0$, varianza $\sigma_1^2 = 1$
- **Distribuzione 2**: media $\mu_2 = 5$, varianza $\sigma_2^2 = 2$

Supponiamo che la popolazione complessiva sia composta per il 40% dalla Distribuzione 1 e per il 60% dalla Distribuzione 2. Quindi i pesi della mistura sono $\pi_1 = 0.4$ e $\pi_2 = 0.6$. Consideriamo di osservare un campione di dati $x_1, x_2, \ldots, x_n$ proveniente da questa mistura. In pratica, ogni osservazione x_i è stata generata o dalla Distribuzione 1 o dalla Distribuzione 2, ma non sappiamo da quale. Qui entra in gioco la variabile $\boldsymbol{Z}_i$. Per ogni osservazione x_i, definiamo un vettore indicatore $\boldsymbol{Z}_i$ che ci dice da quale componente della mistura proviene x_i. Per questo esempio con due componenti:

- $\boldsymbol{Z}_i = (1, 0)$ se l'osservazione x_i proviene dalla Distribuzione 1.
- $\boldsymbol{Z}_i = (0, 1)$ se l'osservazione x_i proviene dalla Distribuzione 2.

Si immagini di avere i seguenti dati osservati:

$$\{x_1 = -0.2, x_2 = 0.1, x_3 = 4.8, x_4 = 5.2, x_5 = 0.3\}$$

Se si suppone di sapere invece da quale distribuzione proviene ogni dato, allora:

- $x_1 = -0.2$ viene dalla Distribuzione 1 $\rightarrow \boldsymbol{Z}_1 = (1, 0)$
- $x_2 = 0.1$ viene dalla Distribuzione 1 $\rightarrow \boldsymbol{Z}_2 = (1, 0)$
- $x_3 = 4.8$ viene dalla Distribuzione 2 $\rightarrow \boldsymbol{Z}_3 = (0, 1)$
- $x_4 = 5.2$ viene dalla Distribuzione 2 $\rightarrow \boldsymbol{Z}_4 = (0, 1)$
- $x_5 = 0.3$ viene dalla Distribuzione 1 $\rightarrow \boldsymbol{Z}_5 = (1, 0)$

In concluzione, se $\boldsymbol{Z}$ fosse noto, ovvero non si è in presenza di dati incompleti, stimare i parametri delle distribuzioni sarebbe semplice, perché si potrebbero separare i dati nelle due distribuzioni e stimare i parametri di ciascuna distribuzione separatamente.

3.3 Algoritmo EM

Come già anticipato nel paragrafo precedente, l'algoritmo *Expectation-Maximization* (EM) è un algoritmo iterativo molto utilizzato al fine di calcolare le stime di massima verosimiglianza nel caso in cui ci si trovi davanti a dati incompleti, cioè dati non classificati, dal momento il problema della stima dei parametri di una mistura è affiancato dalla stima dell'appartenenza delle osservazioni ai vari gruppi. Molto spesso questo algoritmo è applicato nel momento in cui la funzione di verosimiglianza assume forme troppo complicate e, di conseguenza, i metodi numerici, come quello di Newton-Raphson, per la calcolare le stime diventano troppo complicati ed onerosi.
L'algoritmo EM, infatti, deve il suo successo e il suo grande impiego alla sua semplicità di implementazione e programmazione, ma anche alla sua generalità. Quest'ultimo, infatti, può essere applicato sia quando si hanno dati incompleti, ma anche in situazioni più generali in cui l'incompletezza dei dati non è evidente, come nei modelli log-lineari.
Le prime fonti riguardanti l'algortimo EM risalgono ai primi anni del Novecento, anche se la maggior parte del merito si deve a Dempster, Laird e Rubin che, nel 1977, hanno pubblicato un articolo in cui hanno formalizzato l'algoritmo e fornito alcuni esempi.
Lo scopo dell'algortimo EM è quello di formalizzare la procedura più naturale per affrontare il problema dei dati mancanti, ovvero:

- Sostituire i dati mancanti con i dati stimati
- Stimare i parametri
- Stimare nuovamente i dati mancanti assumendo che le stime dei parametri trovare nel punto precedente siano corrette
- Stimare nuovamente i parametri sulla base dei dati nel punto precedente e ripetere la procedura fino alla convergenza.

Come suggerisce il nome, l'algortimo è composto da due passi: un passo E (Expectation step) e un passo M (Maximization step).
Il passo M è il passo più semplice dei due, infatti consiste nel calcolo delle stime di massima verosimiglianza dei parametri sui dati completi, ovvero come se non ci fossero dati mancanti, dunque in questo passo si utilizzano gli stessi metodi numerici o computazionali utilizzati per i dati completi.
Nel passo E, invece, si calcolano i valori attesi condizionati dei dati mancanti rispetto ai dati osservati e alle corrette stime dei parametri di interesse e, dunque, si sostituiscono i valori mancanti con quelli attesi.

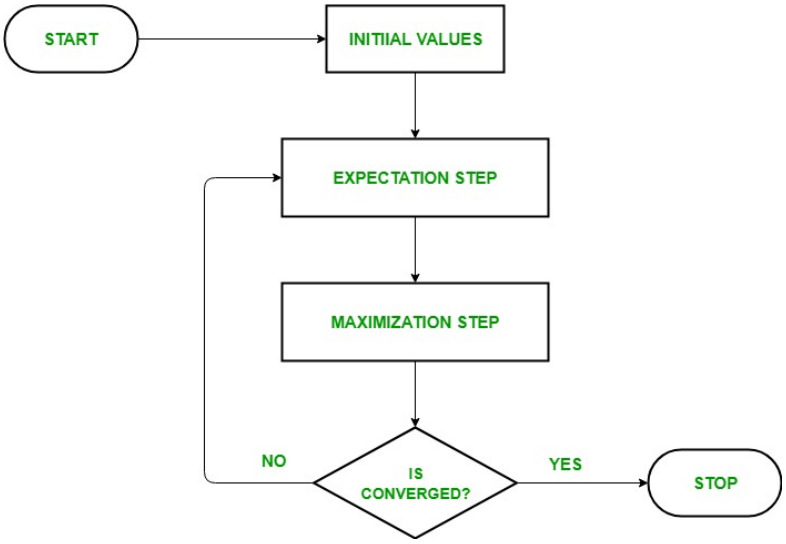


Figura 3.1: Diagramma di flusso dell'algortimo EM

Sebbene l'algoritmo possa essere applicato ad una vasta gamma di modelli, uno dei casi più significativi è quello in cui i dati completi provengono da una famiglia esponenziale. In questo caso, infatti, l'algoritmo risulta ancora più semplice, poichè nel passo E si deve calcolare il valore atteso condizionato delle statistiche sufficienti per i dati completi e, inoltre, il passo M è molto semplice a livello computazionale. Prima di formalizzare l'algoritmo, è utile osservare l'esempio seguente, nonchè lo stesso esempio utilizzato da Dempster, Laird e Rubin nel 1977 per presentare l'algoritmo EM.

3.3.1 Esempio di Dempster, Laird e Rubin

Consideriamo $Y = (y_1, y_2, y_3, y_4)$ la determinazione di una variabile aleatoria multinomiale, a cui è associata la probabilità

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4) = \left(\frac{1}{2} - \frac{\theta}{2}, \frac{\theta}{4}, \frac{\theta}{4}, \frac{1}{2}\right)$$

Si vuole trovare la stima di massima verosimiglianza del parametro θ. Il vettore dei dati osservati $Y_{obs} = (38, 34, 125)$ corrisponde all'osservazione della variabile casuale Y in cui

$$\begin{aligned} y_1 &= 38 \\ y_2 &= 34 \\ y_3 + y_4 &= 125. \end{aligned}$$

Dunque, in questo caso, si può indicare $Y_{obs} = (y_1, y_2, y_3 + y_4)$.
Si supponga che il vettore $Y_{obs} = (38, 34, 125)$ dei dati osservati derivi da una variabile casuale multinomiale con probabilità

$$\pi = (\pi_1, \pi_2, \pi_3) = \left(\frac{1}{2} - \frac{\theta}{2}, \frac{\theta}{4}, \frac{\theta}{4} + \frac{1}{2}\right)$$

Quindi il dato mancante riguarda la parte di y_3+y_4 corrispondente a y_3 (oppure y_4). Nel caso in cui il vettore Y sarebbe stato proprio il vettore di dati osservati, cioè se non ci fosse stata perdita di dati, la stima di massima verosimiglianza per il parametro θ si sarebbe trovata cercando il massimo della funzione di verosimiglianza per i dati completi data da

$$L(\theta|Y) = \frac{(y_1 + y_2 + y_3 + y_4)!}{y_1!y_2!y_3!y_4!} \cdot \pi_1^{y_1} \cdot \pi_2^{y_2} \cdot \pi_3^{y_3} \cdot \pi_4^{y_4}$$

da cui la funzione di log-verosimiglianza in questo caso è data da

$$\mathscr{L}(\theta|Y) \sim y_1 ln(1-\theta) + y_2 ln(\theta) + y_3 ln(\theta)$$

Risolvendo l'equazione di verosimiglianza rispetto alla variabile θ

$$\frac{d\mathscr{L}(\theta|Y)}{d\theta} = 0$$

si ricava facilmente lo stimatore cercato nel caso dei dati completi:

$$\hat{\theta} = \frac{y_2 + y_3}{y_1 + y_2 + y_3}.$$

Si osservi, però, che la log-verosimiglianza è lineare in Y, dunque il valore atteso della stessa rispetto a Y, condizionato a θ e Y_{obs}

$$E[Y_1 \ln(1-\theta) + y_2 \ln(\theta) + Y_3 \ln(\theta) \,|\, Y_{obs}, \theta^{(m)}]$$

comporta calcolare il valore atteso di Y condizionato a θ e Y_{obs}. Se, come in questo caso, ci si trova dinanzi a dati mancanti, è necessario sostituire questi ultimi con delle stime, dunque

$$\begin{aligned}
E[Y_1|\theta, Y_{obs}] &= 38 \\
E[Y_2|\theta, Y_{obs}] &= 34 \\
E[Y_3|\theta, Y_{obs}] &= 125(\theta/4)(1/2 + \theta/4) \\
E[Y_4|\theta, Y_{obs}] &= 125(1/2)(1/2 + \theta/4).
\end{aligned}$$

Dunque, alla m-esima iterazione, in cui si indica con $\theta^{(m)}$ la stima corrente del parametro θ, il passo E consiste nel calcolare

$$y_3^{(m)} = 125(\theta^{(m)}/4)(1/2 + \theta^{(m)}/4) \tag{3.9}$$

In seguito, nel passo M non si dovrà far altro che calcolare il massimo della funzione di log-verosimiglianza per dati completi come visto in precedenza, sostituendo semplicemente y_3 con $y_3^{(m)}$, da cui

$$\theta^{(m+1)} = \frac{34 + y_3^{(m)}}{72 + y_3^{(m)}} \tag{3.10}$$

Iterando le procedure definite nella (3.9) e (3.10) si trova l'algoritmo EM relativo al seguente problema. In seguito vi è una tabella con le iterazioni dell'algoritmo, in cui si mostra anche la convergenza, partendo da un valore iniziale per il parametro θ pari a $\theta^{(0)} = 0.5$.

m	$\theta^{(m)}$	$\theta^{(m)} - \hat{\theta}$	$(\theta^{(m+1)} - \hat{\theta})/(\theta^{(m)} - \hat{\theta})$
0	0.500000000	0.126821498	0.1465
1	0.608247423	0.018574075	0.1346
2	0.624321051	0.002500447	0.1330
3	0.626488879	0.000332619	0.1328
4	0.626777323	0.000044176	0.1328
5	0.626815632	0.000005866	0.1328
6	0.626820719	0.000000779	-
7	0.626821395	0.000000104	-
8	0.626821484	0.000000014	-

Il problema di stima appena visto può, inoltre, essere risolto anche utilizzando l'algoritmo di Newton-Rapshon, il quale, partendo sempre da $\theta^{(0)} = 0.5$, fornisce le prime due stime di θ pari a $\theta^{(1)} = 0.63636363$ e $\theta^{(2)} = 0.62696867$. Dunque, facendo un paragone tra questi valori e quelli riportati nella tabella, è evidente che, partendo dal medesimo dato iniziale, la convergenza del metodo di Newton-Rapshon è molto più veloce di quella dell'algoritmo EM, dal momento che il primo converge in circa due iterazione, mentre il secondo in cinque. Questo, purtoppo, mostra uno dei punti critici dell'algoritmo EM, ovvero la sua lentenza nella convergenza.

3.3.2 Descrizione dell'algoritmo EM

Si supponga di voler trovare $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$ per massimizzare la funzione di verosimiglianza $L(\boldsymbol{\Psi}) = f(\boldsymbol{x}|\boldsymbol{\Psi})$, in cui $\boldsymbol{x}$ indica un insieme di dati incompleti. Si denoti, invece, con $\boldsymbol{y}$ la versione 'completa' dei dati $\boldsymbol{x}$ e sia $\mathscr{Y}(\boldsymbol{x})$ l'insieme dei possibili valori di

$\boldsymbol{y}$ (nella mistura dell'esempio 3.2.1, $\mathscr{Y}(\boldsymbol{x})$ contiene c^n punti, corrispondenti alle c^n scelte per $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$). Si denoti la funzione di verosimiglianza per $\boldsymbol{y}$ con

$$g(\boldsymbol{y}|\boldsymbol{\Psi}).$$

L'algoritmo EM genera, da un'approssimazione iniziale $\boldsymbol{\Psi}^{(0)}$ una sequenza $\{\boldsymbol{\Psi}^{(m)}\}$ di stime. Ogni iterazione consiste nei due seguenti step:

$$\begin{aligned} \text{E step :} \quad & \text{Valutare } E[log(g(\boldsymbol{y}|\boldsymbol{\Psi}))|\boldsymbol{x}, \boldsymbol{\Psi}^{(m)}] = Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(m)}) \\ \text{M step :} \quad & \text{Trovare } \boldsymbol{\Psi} = \boldsymbol{\Psi}^{(m+1)} \text{ per massimizzare } Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(m)}). \end{aligned}$$

Molto spesso lo step M è il più semplice da calcolare. Inoltre, grazie ad una semplice dimostrazione basata sulla disuguaglianza di Jensen, si prova che

$$L(\boldsymbol{\Psi}^{(m+1)}) \geq L(\boldsymbol{\Psi}^{(m)}) \quad m = 0, 1, \ldots,$$

assicurando il fatto che le verosimiglianze di interesse siano monotone crescenti. L'uguaglianza, solitamente, indica che ci si trova in un punto stazionario della funzione di verosimiglianza. La versione dell'algoritmo EM per i problemi riguardanti le misture finite è definita come segue: ricordando che, in questo caso, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ indicano le quantità mancanti, si ottiene dalla (3.8)

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \boldsymbol{w}^T \boldsymbol{V}(\boldsymbol{\pi}) + \sum_{i=1}^{n} \boldsymbol{w}^T \boldsymbol{U_i}(\boldsymbol{\theta}),$$

dove

$$\boldsymbol{w} = \boldsymbol{w_i}(\boldsymbol{\Psi^{(m)}}) = E[\boldsymbol{z_i}|x_i, \boldsymbol{\Psi^{(m)}}]$$

Dunque,

$$\boldsymbol{w_{ij}}(\boldsymbol{\Psi^{(m)}}) = [\boldsymbol{w_i}(\boldsymbol{\Psi^{(m)}})]_j = \pi_j^{(m)} f_j(x_i|\boldsymbol{\theta_j^{(m)}})/p(x_i|\boldsymbol{\Psi^{(m)}}), \quad \text{per ogni } i, j.$$

Questi 'pesi' sono, dunque, le probabilità di appartenenza ad una categoria per l'osservazione i-esima, condizionata da x_i e, considerato ciò, il parametro è $\boldsymbol{\Psi^{(m)}}$. Solitamente, i parametri $\boldsymbol{\pi}$ e $\boldsymbol{\theta}$ sono stimati separatamente, con la conseguenza che lo **step M** per il parametro $\boldsymbol{\pi}$ è dato da

$$\pi_j^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}(\boldsymbol{\Psi^{(m)}}), \quad j = 1, \ldots, c.$$

Questo, appunto, spiega come nell'Esempio 3.2.1 l'iterazione (3.6) sia proprio l'algoritmo EM.

La forma dello **step M** per il parametro $\boldsymbol{\theta}$ è, per la maggior parte, specifica per ogni problema, sebbene corrisponda in generale alla massimizzazione della seconda sommatoria nel termine $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(m)})$. Dunque, lo scopo dell'algoritmo EM è produrre simultaneamente un aggiornamento dei parametri sia nella componente di descrizione del modello, ovvero $\boldsymbol{\theta}$, sia nella parte di descrizione della mistura, ovvero $\boldsymbol{\pi}$.

Esempio (continuazione Esempio 3.2.2)**.** Applicando direttamente l'algoritmo EM, si ottiene, ponendo $n_j^{(m)} = \sum_i w_{ij}(\boldsymbol{\Psi^{(m)}})$ per $j = 1, 2$

$$\begin{aligned}
\pi_j^{(m+1)} &= \frac{n_j^{(m)}}{n} \\
\mu_j^{(m+1)} &= \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} w_{ij}^{(m)} x_i \\
\sigma_j^{2(m+1)} &= \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} w_{ij}^{(m)} (x_i - \mu_j^{(m+1)})^2,
\end{aligned}$$

dove $w_{ij}^{(m)} = w_{ij}(\boldsymbol{\Psi^{(m)}})$. L'estensione di questa procedura per misture di c componenti è diretta.

3.3.3 Pregi e difetti dell'algoritmo EM

Dopo aver formalizzato l'algoritmo EM, è bene notare che quest'ultimo possiede diversi pregi e difetti. Alcuni degli aspetti positivi dell'algoritmo sono:

- Non serve calcolare e invertire matrici;
- È molto semplice da costruire dal momento che il passo E e il passo M si basano su calcoli fatti sui dati completi;
- È un algoritmo facile da implementare in un calcolatore;
- Trasforma il problema di massimizzare la funzione di verosimiglianza nel caso in cui ci siano dati mancanti in un problema statistico. Infatti, attraverso il passo E si completano i dati e tramite il passo M si calcola la stima di massima verosimiglianza usando i dati completi;
- Come visto, la log-verosimiglianza aumenta ad ogni iterazione e, spesso, converge ad un massimo locale.

Tra i difetti, invece, si annoverano:

- L'algoritmo risulta efficiente quando il passo E si può calcolare direttamente;
- La velocità di convergenza può risultare molto lenta, soprattutto se si è in presenza di molti dati mancanti;
- Non sempre l'algoritmo converge ad un massimo globale.

Capitolo 4

Applicazione dell'Algoritmo EM e delle distribuzioni mistura

Dopo aver presentato nel dettaglio i modelli mistura e l'algoritmo EM, utile a ricavare delle stime per i parametri del modello, è interessante studiare come l'algoritmo possa essere utile in diverse applicazioni pratiche. Nel seguente capitolo sarà analizzato uno studio nell'ambito della chimica enologica, dove è fondamentale identificare e quantificare la presenza di adulteranti nei mosti di vino. L'aggiunta di zuccheri estranei, ad esempio, può alterare significativamente le proprietà isotopiche e chimiche del mosto, rendendo possibile l'individuazione dell'adulterazione attraverso metodi statistici sofisticati. Il documento di riferimento, "Sugar Adulterations Control in Concentrated Rectified Grape Musts by Finite Mixture Distribution Analysis of the myo- and scyllo-Inositol Content and the D/H Methyl Ratio of Fermentative Ethanol" (Monetti et al., 1996), rappresenta una pionieristica integrazione delle metodologie statistiche avanzate con le tecniche di analisi chimica e isotopica, dimostrando come l'algoritmo EM possa essere utilizzato per migliorare la rilevazione delle adulterazioni nei mosti concentrati rettificati d'uva (CRM).

4.1 Analisi del documento

4.1.1 Contesto del CRM e delle adulterazioni

Il mosto concentrato rettificato (CRM) è un prodotto derivato dall'uva, ottenuto mediante un processo di concentrazione e rettificazione del mosto fresco. Il processo prevede l'evaporazione dell'acqua dal mosto d'uva fresco, aumentando così la concentrazione degli zuccheri naturali presenti. Successivamente, il mosto concentrato subisce una rettificazione, un processo che rimuove selettivamente alcune

componenti non zuccherine, come acidi e composti fenolici, per ottenere un prodotto con un alto contenuto di zuccheri e un sapore neutro.
Il CRM è apprezzato per il suo elevato contenuto zuccherino e la sua purezza, caratteristiche che lo rendono ideale per la produzione di vari prodotti alimentari e bevande. A causa della sua purezza, il CRM è spesso soggetto a pratiche di adulterazione, dove zuccheri esogeni come il saccarosio da barbabietola o canna da zucchero vengono aggiunti per aumentare il volume del prodotto. Tuttavia, proprio queste caratteristiche lo rendono vulnerabile a frodi alimentari, in cui produttori senza scrupoli aggiungono zuccheri di origine diversa dall'uva per ridurre i costi.
Lo studio raccoglie misurazioni su diversi campioni di mosto concentrato rettificato d'uva (CRM). Le principali variabili misurate sono:

1. **Contenuto di myo-inositolo e scyllo-inositolo:** Questi due polialcoli sono indicatori della genuinità del mosto. Il myo-inositolo è naturalmente presente nell'uva, mentre lo scyllo-inositolo, pur essendo un isomero, può fornire ulteriori informazioni sulla purezza del mosto.

2. D/H_I **(Rapporto isotopico D/H del metile, prima posizione):** Questo rapporto rappresenta il rapporto tra il deuterio (D, un isotopo dell'idrogeno con un protone e un neutrone) e l'idrogeno (H) nella posizione specifica di un atomo di carbonio all'interno del gruppo metile (-CH3) dell'etanolo. Il rapporto D/H_I si riferisce specificamente alla misurazione del deuterio nella prima posizione del metile.

3. D/H_{II} **(Rapporto isotopico D/H del metile, seconda posizione):** Analogamente, il rapporto D/H_{II} rappresenta il rapporto tra deuterio e idrogeno nella seconda posizione del metile dell'etanolo. Queste misurazioni isotopiche sono sensibili alle origini dei componenti zuccherini e possono variare a seconda della fonte del carbonio che viene fermentato per produrre l'etanolo.

Nella procedura utilizzata per rilevare le adulterazioni dello zucchero, i valori di un campione sospetto vengono confrontati con quelli di un campione di riferimento, ottenuto nelle stesse condizioni del campione in esame, se disponibile. In alternativa, si confrontano con una banca dati di campioni non adulterati, tenendo conto della variabilità naturale. Ad esempio, per i vini italiani, a partire dal 1987, è stata istituita una banca dati con circa 500 campioni all'anno, divisi in due macroregioni corrispondenti alle aree settentrionali e meridionali della penisola, e, dal 1991, i dati sono stati validati da un apposito comitato della UE. In questa banca dati sono presenti i range di riferimento di tutte le variabili precedentemente presentate, in modo tale che sia semplice confrontare i valori delle stesse per

classificare al meglio i campioni.

4.1.2 Limiti dei metodi tradizionali

Come detto in precedenza, tra le possibili scelte, gli zuccheri esogeni più interessanti per effettuare adulterazioni sono quelli di barbabietola e di canna. I metodi tradizionali di rilevazione delle adulterazioni si basano su analisi isotopiche (rapporti D/H e $^{13}C/^{12}C$) e sulla misurazione di specifici polialcoli come il myo-inositolo. Tuttavia, questi metodi possono avere limitazioni, specialmente quando le adulterazioni rientrano nei range di variabilità naturale dei parametri isotopici. Infatti, con un'aggiunta oculata, un CRM adulterato proveniente da alcune regioni del sud Italia potrebbe essere venduto come genuino delle aree del centro-nord, poiché i suoi parametri rientrano nella gamma della variabilità naturale e sono quindi accettabili.

4.1.3 Nuovo approccio

Lo studio propone un metodo innovativo che integra l'analisi isotopica, che misura i rapporti tra isotopi stabili dell'idrogeno e del carbonio presenti nel CRM, che possono variare a seconda dell'origine degli zuccheri, con la misurazione dei polialcoli myo- e scyllo-inositolo. Tuttavia, le sofisticate tecniche di adulterazione possono fare in modo che i parametri misurati rientrino nei range naturali, rendendo difficile la distinzione tra CRM genuino e adulterato.
La novità del metodo proposto risiede nella combinazione di due tecniche analitiche. La misurazione di myo-inositolo e scyllo-inositolo, combinata con i dati isotopici, fornisce un insieme di dati che può essere modellato come una mistura di distribuzioni finite. L'algoritmo EM viene quindi utilizzato per analizzare questi dati, permettendo di distinguere in modo più preciso tra CRM genuino e adulterato.

4.2 Algoritmo EM e distribuzioni mistura

Come già spiegato nei capitoli precedenti, le distribuzioni mistura sono utilizzate per modellare una popolazione statistica composta da diverse sottopopolazioni. In questo contesto, vengono applicate per distinguere tra campioni di mosto concentrato rettificato genuini e adulterati.
L'idea è che ogni campione di CRM possa essere considerato come un'osservazione da una popolazione che è una mistura di diverse sottopopolazioni, cioè, nel nostro caso, campioni genuini e adulterati. Utilizzando le distribuzioni mistura, si cerca

di identificare le componenti di questa mistura e stimare i parametri delle distribuzioni di ciascuna componente.

Come già visto, esistono diversi metodi per stimare una distribuzione mistura. In questo lavoro è stato adottato l'approccio della massima verosimiglianza perché, in condizioni molto generali, offre alcune proprietà utili, come la consistenza degli stimatori, che convergono in probabilità ai veri valori dei parametri, e la normalità asintotica (Everitt, 1985). Tuttavia, la complessa dipendenza della funzione di verosimiglianza dai parametri da stimare causa difficoltà computazionali che non possono essere risolte esplicitamente. Per questo motivo, è necessario ricorrere a soluzioni approssimative tramite procedure iterative, come l'algoritmo EM.
L'algoritmo funziona iterativamente attraverso due fasi principali:

1. **Fase di Aspettazione (E-step):** Calcolare le probabilità di appartenenza di ciascun campione a ciascuna componente della mistura. In altre parole, si stima la probabilità che un dato campione sia genuino o adulterato basandosi sui parametri correnti della distribuzione.

2. **Fase di Massimizzazione (M-step):** Aggiornare i parametri delle distribuzioni mistura massimizzando la funzione di verosimiglianza completa con le probabilità di appartenenza calcolate nella fase E.

Questi due passaggi vengono ripetuti fino a quando viene soddisfatto un criterio di convergenza, come un miglioramento trascurabile della verosimiglianza oppure, come nel presente lavoro, una differenza assoluta massima tra tutti i parametri nelle due iterazioni successive inferiore a 10^{-5}. Nel caso dell'algoritmo EM, la soluzione è più rapida se le popolazioni sono ben separate e se i valori iniziali scelti sono appropriati (Titterington et al., 1985), la convergenza è assicurata, ma l'algoritmo potrebbe richiedere molte iterazioni o potrebbe convergere ad un massimo locale. In questo caso, per garantire stime di massima verosimiglianza coerenti, il myo- e scyllo-inositolo sono stati trasformati logaritmicamente.
Attraverso l'applicazione dell'algoritmo EM, sono stati stimati i parametri delle due popolazioni e l'attenzione è stata rivolta al gruppo non adulterato per derivare una regola di classificazione di interesse pratico. Questo criterio si basa sulla distanza generalizzata di Mahalanobis (Krzanowski, 1988), della quale si discuterà in seguito, che consente di classificare nuove osservazioni e di minimizzare l'errore di probabilità assegnando un'osservazione sconosciuta al gruppo dal quale la distanza su tutte le variabili è inferiore a un limite precedentemente stabilito a un certo livello di significatività.
Infine, per verificare se i CRM genuini non siano realmente adulterati, i valori di D/H_I, precedentemente trasformati tramite l'equazione

$$(\mathrm{D}/\mathrm{H}_I)_{\mathrm{WINE}} = -14.35 + 1.16(\mathrm{D}/\mathrm{H}_I)_{\mathrm{CRM}}$$

che permette di confrontare i valori D/H_I del CRM con quelli dei vini, sono stati confrontati con i valori dei campioni di vino genuini appartenenti alla banca dati italiana (Monetti et al., 1993, 1994). Utilizzando un'analisi della varianza (ANOVA) su D/H_I e considerando che i CRM sono principalmente prodotti con uve provenienti dalle regioni meridionali italiane, i candidati non adulterati sono stati confrontati con i vini genuini delle stesse regioni. Solo questo parametro è stato considerato poiché nei campioni della banca dati non è determinato il contenuto di poli-alcol e D/H_{II} è influenzato dalle caratteristiche dell'acqua utilizzata per la diluizione.

4.2.1 Scelta del modello

Come spiegato nei capitoli precedenti, un modello mistura può avere diverse forme, dal momento che può dipendere da come sono fatte le componenti, da come si differenziano tra loro e, ovviamente, anche dalla forma analitica. Per quanto riguarda quest'ultima, si è deciso, come già detto, di considerare una mistura con due componenti. Quello che può essere interessante, invece, è capire se si è di fronte a componenti che condividono medie e varianza oppure sono completamente differenti. Dal momento che vi è una casistica articolata di possibili combinazioni, tutte compatibili con una mistura a due componenti, è fondamentale utilizzare un criterio per individuare il modello più adeguato. Il criterio in questione è il **BIC**, Bayesian Information Criterion. Il BIC è particolarmente utile quando si cerca di bilanciare la complessità del modello (ossia il numero di parametri) con la bontà del modello nell'adattare i dati osservati. Il BIC viene calcolato utilizzando la formula

$$\mathrm{BIC} = -2 \cdot \ln(\hat{L}) + k \cdot \ln(n)$$

dove

- $\hat{L}$ è il valore massimo della verosimiglianza $L(\theta)$, che misura quanto bene il modello con i parametri stimati si adatta ai dati
- k è il numero di parametri del modello
- n è il numero di osservazioni.

Il primo termine, $-2 \cdot \ln(\hat{L})$, deriva direttamente dalla massima verosimiglianza stimata dal modello. Se il modello si adatta molto bene ai dati, la verosimiglianza sarà alta, e quindi $-2 \cdot \ln(\hat{L})$ sarà relativamente bassa. Questo rappresenta

la "qualità dell'adattamento" del modello. Il secondo termine, $k \cdot \ln(n)$, invece, penalizza la complessità del modello. In altre parole, quando si adatta un modello ai dati, un modello più complesso, cioè con più parametri k, tende a fornire una migliore aderenza ai dati, ovvero un $\ln(\hat{L})$ maggiore. Tuttavia, un modello con molti parametri potrebbe non generalizzare bene a nuovi dati, il che porta al fenomeno dell'overfitting. Il termine $k \cdot \ln(n)$ serve a penalizzare l'aggiunta di troppi parametri, in modo che il BIC non favorisca automaticamente i modelli più complessi. La penalizzazione cresce con:

- **Il numero di parametri** k: Più parametri ha il modello, maggiore è la penalizzazione. Questo riflette il fatto che ogni nuovo parametro aggiunge complessità al modello, e quindi c'è un costo per la sua inclusione.
- **Il numero di osservazioni** n: La penalizzazione dipende anche dalla dimensione del dataset attraverso il logaritmo del numero di osservazioni. Un dataset più grande può supportare l'uso di modelli più complessi senza incorrere in overfitting, quindi la penalizzazione per parametro diminuisce proporzionalmente.

In sintesi, il termine $k \cdot \ln(n)$ nel BIC garantisce che l'aggiunta di parametri sia giustificata solo se migliora sufficientemente l'aderenza del modello ai dati. Se l'aumento della complessità non comporta un miglioramento significativo della verosimiglianza, il BIC penalizzerà il modello più complesso, preferendo modelli più semplici.
L'idea alla base del BIC è di trovare un modello che equilibri il compromesso tra la bontà dell'adattamento e la complessità. Un BIC più basso indica un modello preferibile. Quando si confrontano più modelli, quello con il BIC più basso è generalmente considerato il migliore. La libreria utilizzata in questo lavoro, `Mclust`, è in grando di individuare diversi modelli, identificati da acronimi in cui:

- **E** (Equal) indica che la caratteristica è uguale per tutte le componenti
- **V** (Variable) indica che la caratteristica può variare tra le componenti
- **I** (Identity) indica che la matrice di covarianza è una matrice identità (nessuna correlazione tra le variabili).

I modelli sono i seguenti:

- **EII (Equal volume, Identity shape, Identity orientation)**: Volume uguale per tutti i cluster, forma sferica uguale per tutti i cluster, nessuna orientazione preferita (matrice identità).

- **VII (Variable volume, Identity shape, Identity orientation)**: Volume variabile tra i cluster, forma sferica uguale per tutti i cluster, nessuna orientazione preferita (matrice identità).

- **EEI (Equal volume, Equal shape, Identity orientation)**: Volume uguale per tutti i cluster, forma ellissoidale uguale per tutti i cluster, nessuna orientazione preferita (matrice identità).

- **VEI (Variable volume, Equal shape, Identity orientation)**: Volume variabile tra i cluster, forma ellissoidale uguale per tutti i cluster, nessuna orientazione preferita (matrice identità).

- **EVI (Equal volume, Variable shape, Identity orientation)**: Volume uguale per tutti i cluster, forma variabile tra i cluster, nessuna orientazione preferita (matrice identità).

- **VVI (Variable volume, Variable shape, Identity orientation)**: Volume variabile tra i cluster, forma variabile tra i cluster, nessuna orientazione preferita (matrice identità).

- **EEE (Equal volume, Equal shape, Equal orientation)**: Volume uguale per tutti i cluster, forma uguale per tutti i cluster, orientazione uguale per tutti i cluster.

- **VEE (Variable volume, Equal shape, Equal orientation)**: Volume variabile tra i cluster, forma uguale per tutti i cluster, orientazione uguale per tutti i cluster.

- **EVE (Equal volume, Variable shape, Equal orientation)**: Volume uguale per tutti i cluster, forma variabile tra i cluster, orientazione uguale per tutti i cluster.

- **VVE (Variable volume, Variable shape, Equal orientation)**: Volume variabile tra i cluster, forma variabile tra i cluster, orientazione uguale per tutti i cluster.

- **EEV (Equal volume, Equal shape, Variable orientation)**: Volume uguale per tutti i cluster, forma uguale per tutti i cluster, orientazione variabile tra i cluster.

- **VEV (Variable volume, Equal shape, Variable orientation)**: Volume variabile tra i cluster, forma uguale per tutti i cluster, orientazione variabile tra i cluster.

- **EVV (Equal volume, Variable shape, Variable orientation)**: Volume uguale per tutti i cluster, forma variabile tra i cluster, orientazione variabile tra i cluster.
- **VVV (Variable volume, Variable shape, Variable orientation)**: Volume variabile tra i cluster, forma variabile tra i cluster, orientazione variabile tra i cluster.

Nella Tabella 4.1 si nota come, nel caso in questione, il modello migliore sia il VVV, risultato confermato anche dalla Figura 4.2. Nella Figura 4.1, inoltre, è possibile vedere una panoramica del BIC nel caso in cui il numero di componenti sia diverso da 2.

	VVV,2	**VEV,2**	**VVE,2**
BIC	-1668.946	-1685.79694	-1685.83862
BIC diff	0.000	-16.85097	-16.89266

Tabella 4.1: Tabella con valori del BIC

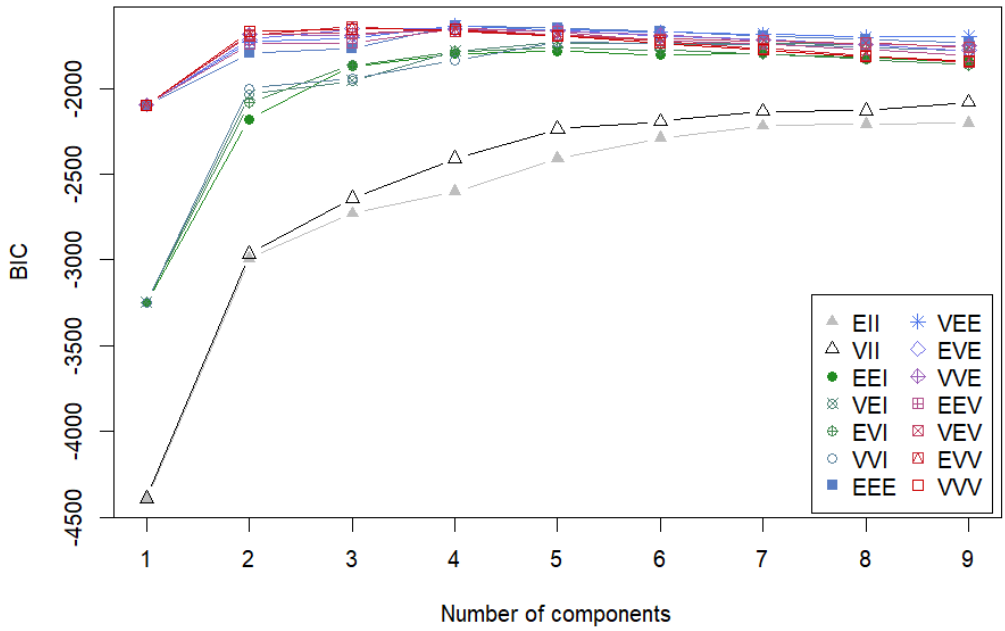


Figura 4.1: BIC completo

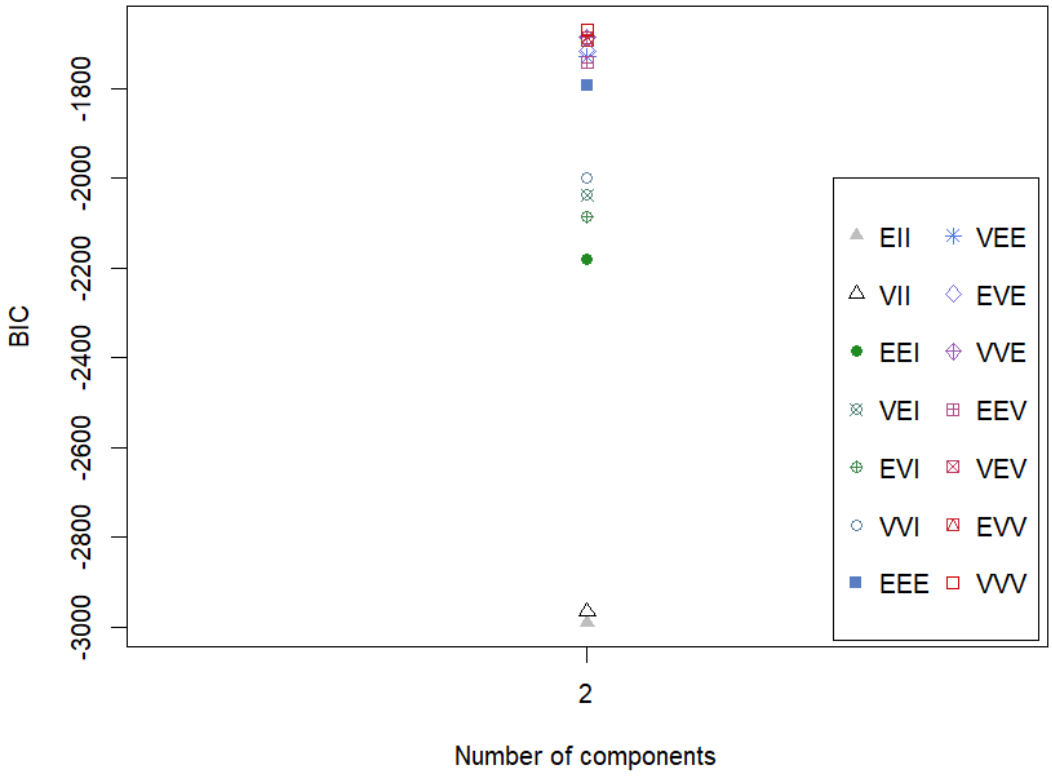


Figura 4.2: Risultati BIC

4.2.2 Identificazione delle sottopopolazioni tramite Clustering

Per individuare potenziali sottopopolazioni all'interno del dataset di vini, è stato utilizzato un approccio di clustering basato su modelli mistura gaussiana. Nello specifico, è stata utilizzata la libreria R di clustering `Mclust`, che ha permesso di suddividere i campioni in due gruppi distinti, corrispondenti a due sottopopolazioni di interesse. La libreria `Mclust` non solo identifica il numero ottimale di gruppi attraverso il criterio di informazione bayesiano (BIC), ma fornisce anche le probabilità di appartenenza di ciascun campione a ciascuna delle classi individuate. In questo caso, i campioni sono stati classificati in due gruppi, che possono essere interpretati come "Genuine" (non adulterato) e "Adulterated" (adulterato), in base alle caratteristiche chimiche analizzate. Di seguito è riportato il plot che visualizza la classificazione ottenuta. Ogni punto nel grafico rappresenta un campione, colo-

rato in base alla classe a cui è stato assegnato. Questo grafico permette di visualizzare chiaramente come i campioni si distribuiscono tra le due sottopopolazioni identificate.

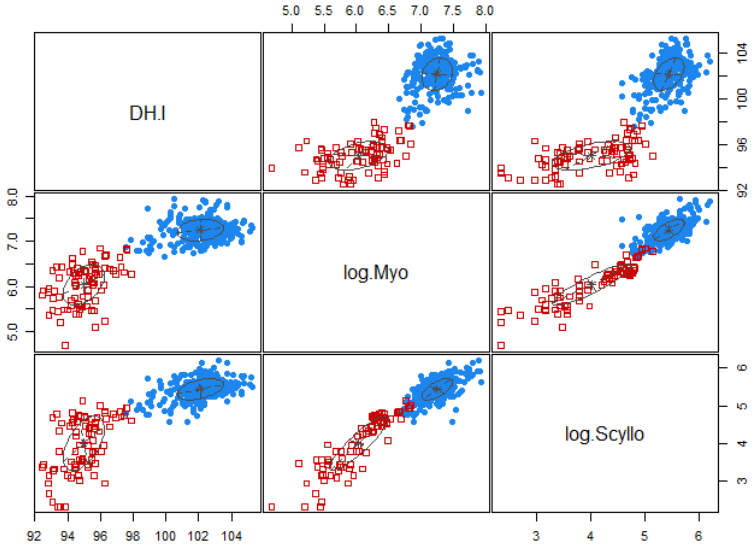


Figura 4.3: Classificazione

4.2.3 Analisi dei risultati

Nove campioni sono stati inizialmente esclusi dal set di dati: sette di essi erano adulterati poiché privi di myo-inositolo, uno aveva un contenuto nullo di scyllo-inositolo e l'ultimo era probabilmente adulterato con zucchero di canna, poli-alcoli bassi e alto D/H_I. Alla fine, dopo aver derivato la regola di classificazione, tutti i campioni sono stati reintrodotti nel set di dati per testare l'assegnazione.

Tra le quattro variabili, il cui comportamento è mostrato in Figura 4.4, D/H_{II} è stato scartato per il suo scarso potere discriminante, confermato anche dalla sua distribuzione marginale delle frequenze. Al contrario, D/H_I presentava una distribuzione bimodale, con un picco intorno a 95 ppm per il gruppo adulterato e 103 ppm per l'altro gruppo. Allo stesso modo, i poli-alcoli differivano nei due gruppi, con valori modali di 750 e 1400 mg/kg di zuccheri per il myo-inositolo e di circa 30 e 230 mg/kg di zuccheri per il scyllo-inositolo nei mosti adulterati e non adulterati, rispettivamente. Osservando la Figura 4.4, è importante notare che il comportamento degli istogrammi di frequenza è fortemente influenzato dal numero di classi scelto.
Nella figura è evidente che i campioni appartengono a due gruppi distinti in base

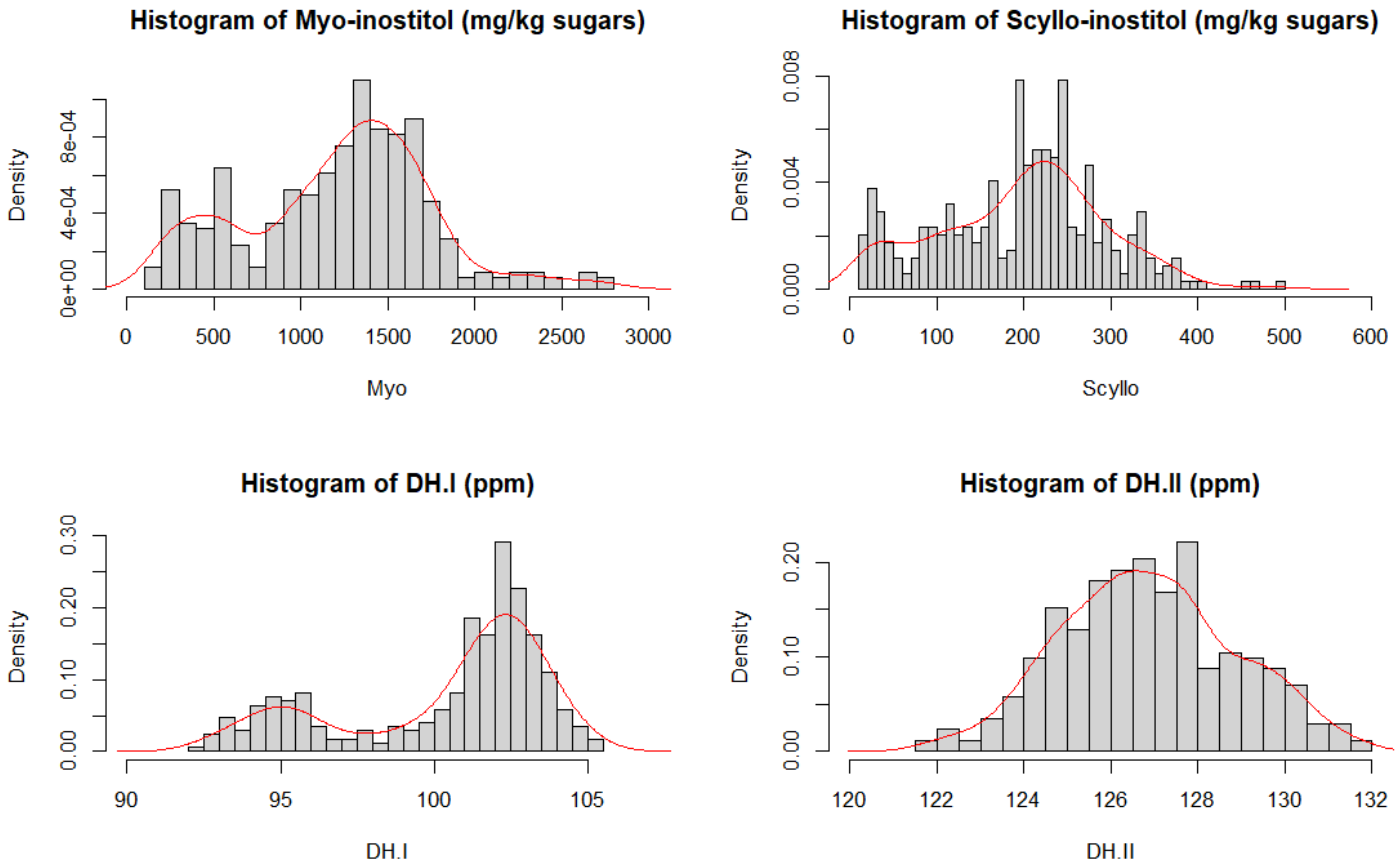


Figura 4.4: Distribuzioni di frequenza marginali e funzioni di densità di probabilità calcolate tramite l'algoritmo EM su dati non trasformati.

alla presenza o assenza di adulterazioni, ma questi gruppi non sono immediatamente identificabili a causa della sovrapposizione. Di conseguenza, senza informazioni preliminari, la prima analisi è stata orientata alla separazione dei gruppi tramite l'algoritmo EM.
Le stime iniziali necessarie per avviare l'algoritmo sono basate sia sulla distribuzione delle frequenze che sui grafici di probabilità di ciascuna variabile, mentre il numero delle popolazioni è stato determinato sulla base delle conoscenze preliminari del problema. I valori iniziali sono:

1. **Gruppo adulterato:** D/H_I = 95 ppm, myo-inositolo = 300 mg/kg di zuccheri, e scyllo-inositolo = 30 mg/kg di zuccheri;
2. **Gruppo genuino:** D/H_I = 103 ppm, myo-inositolo = 1400 mg/kg di zuccheri, e scyllo-inositolo = 231 mg/kg di zuccheri.

		ADULTERATED SAMPLES			GENUINE SAMPLES		
	EM iterations	Proportion (%)	Mean	Variance–Covariance Matrix	Proportion (%)	Mean	Variance–Covariance Matrix
D/H_I	25	25.85	95.21	2.2021	74.15	102.19	1.7221
myo-inositol	80	26.89	6.14	0.2897	73.11	7.25	0.0564
scyllo-inositol	91	30.68	4.29	0.7340	69.31	5.44	0.0711
D/H_I + *myo-inositol*	5	26.37	95.28 6.11	2.4211 0.4587 0.4587 0.2481	73.63	102.21 7.26	1.6524 0.0231 0.0231 0.0538
D/H_I + *scyllo-inositol*	11	24.95	95.09 4.04	1.8533 0.5765 0.5765 0.5416	75.04	102.14 5.44	1.8628 0.1662 0.1662 0.0780
myo-inositol + *scyllo-inositol*	41	24.26	6.05 4.03	0.2278 0.3237 0.3237 0.5666	75.73	7.25 5.43	0.0599 0.0465 0.0465 0.0815
D/H_I + *myo-inositol* + *scyllo-inositol*	11	24.15	94.99 6.04 4.01	1.5802 0.2449 0.4855 0.2449 0.2048 0.2923 0.4855 0.2923 0.5212	75.84	102.10 7.25 5.44	2.0109 0.061 0.1810 0.0611 0.0570 0.0434 0.1810 0.0434 0.079

Tabella 4.2: Parametri statistici (medie, matrici di varianza e covarianza e proporzioni) del gruppo adulterato e di quello genuino stimati attraverso l'algoritmo EM con diverse combinazioni di variabili

Per confermare che la soluzione non fosse un ottimo locale, sono stati utilizzati altri valori iniziali, ma la soluzione è risultata unica. Il processo di convergenza è stato relativamente rapido: ad esempio, nel caso trivariato sono state necessarie 11 iterazioni e la funzione di densità congiunta trivariata per le due popolazioni (adulterata, A, e genuina, G) è

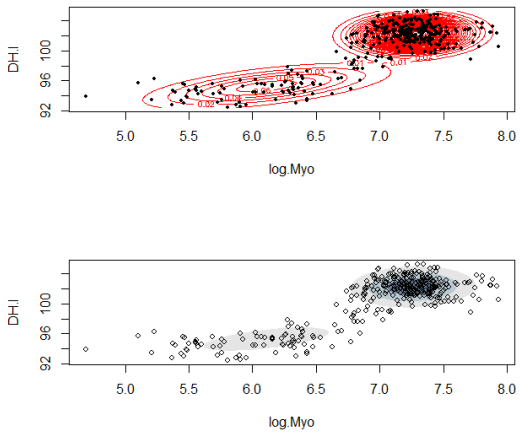
$$f(\mathbf{x}) = 0.2419 f_A(\mathbf{x}) + 0.7581 f_G(\mathbf{x})$$

dove

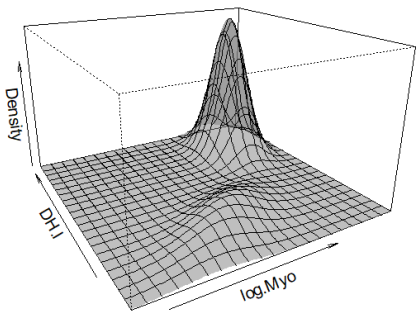
$$f_i(\mathbf{x}) = (2\pi)^{-m/2} |\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu_i)^\top \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right)$$

è l'i-esima componente della dentistà normale multivariata.
Poiché l'obiettivo principale era stimare i parametri della popolazione CRM non

adulterata, le proporzioni stimate per ciascun gruppo erano di importanza secondaria. Questo è particolarmente vero considerando che tali proporzioni erano influenzate dalle caratteristiche del campione. Nel caso dei campioni commerciali, la proporzione di campioni adulterati risultava elevata a causa di alcune aziende soggette a controlli più frequenti, poiché i loro campioni risultavano spesso anomali. Pertanto, le proporzioni osservate in questo studio non rappresentano le caratteristiche reali della produzione di CRM, ma sono specifiche ai campioni analizzati.
Nella Tabella 4.2 è possibile vedere come il D/H_I, il myo-inositolo e scyllo-inositolo hanno bisogno di un numero di iterazioni elevato, dal momento che la separazione tra i gruppi è progressivamente meno evidente; in ogni caso, per tenere conto di tutte le informazioni disponibili e ottenere una discriminazione robusta, tutte le variabili sono state utilizzate nelle analisi successive. La bontà dell'adattamento delle stime EM è mostrata nella Figura 4.4, dove le funzioni di densità teoriche, calcolate a partire dai dati non trasformati, sono sovrapposte agli istogrammi di frequenza empirici.

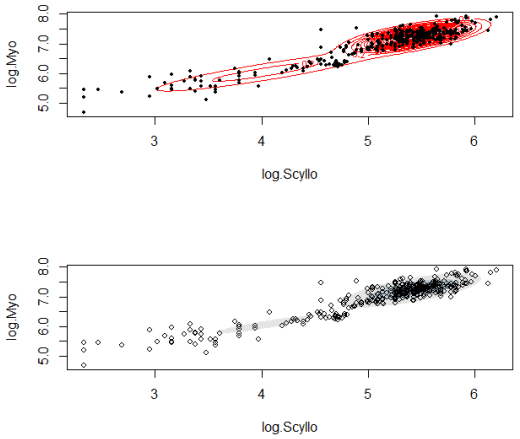


(a) Contour plot della densità stimata della mistura per D/H_I e *myo*-inositolo

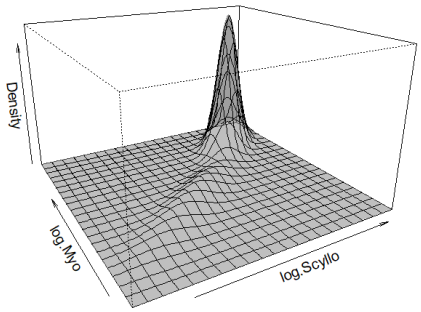


(b) Prospettiva

Figura 4.5: Due grafici comparativi

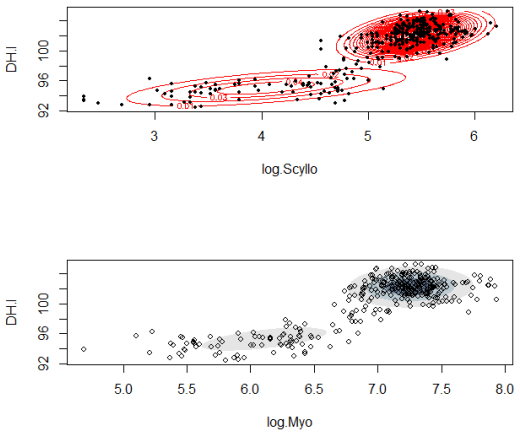


(a) Contour plot della densità stimata della mistura per *scyllo*-inositolo e *myo*-inositolo

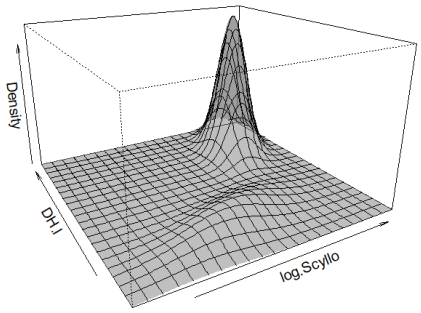


(b) Prospettiva

Figura 4.6: Due grafici comparativi



(a) Contour plot della densità stimata della mistura per D/H_I e *scyllo*-inositolo



(b) Prospettiva

Figura 4.7: Due grafici comparativi

Le Figure 4.5a, 4.6a, 4.7a mostrano le densità delle misture stimate sotto forma di un contour plot per il D/H_I-myo-inositolo e per i due polialcoli. Nel primo caso, i due gruppi sono ben separati perché il loro sovrapporsi è minimo. Infatti, nel gruppo adulterato c'è una sostanziale correlazione di r = 0,61 tra le due variabili;

nell'altro gruppo, con un coefficiente di r = 0,02, non c'è nulla: l'adulterazione introduce una dipendenza riducendo i valori di entrambe le variabili. Lo stesso effetto di diluizione è apprezzabile nella Figura 4.6a: i polialcoli sono fortemente correlati (r = 0,90) nel gruppo adulterato e molto meno nell'altro (r = 0,66) e lo stesso comportamento, con correlazioni rispettivamente pari a r = 0,58 e r = 0,43, è osservabile nella combinazione D/H_I-scyllo-inositolo.
Con tutti i parametri necessari disponibili per classificare campioni sconosciuti, è stato sviluppato un criterio basato sul CRM genuino. La distribuzione trivariata normale del CRM genuino, completamente definita dal vettore delle medie e dalla matrice varianza-covarianza, è stata utilizzata per derivare una regola di classificazione basata sulla distanza generalizzata di Mahalanobis, definita da

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}).$$

Questa distanza misura l'estensione di un campione dal suo centroide $\bar{\mathbf{x}}$, ovvero il punto nello spazio m-dimensionale in cui ogni coordinata rappresenta la media di una delle variabili attraverso tutti i campioni e, sapendo che d_i^2 è distribuito come χ_m^2, si può calcolare la distanza limite oltre la quale, con una certa probabilità, un campione non appartiene a una determinata popolazione.
Nel caso in questione, il livello di significatività è stato fissato al valore $\alpha = 5‰$ e, con $m = 3$ (poichè si stanno considerando tre variabili), la distanza limite corrispondente è $\chi^2_{3,\alpha=5‰} = 12.84$. La regola di allocazione per un nuovo campione, indicando con ν_i il vettore dei suoi valori, con π_g la popolazione del CRM genuino, e con π_a la popolazione degli adulterati, potrebbe essere espressa come

$$\text{assegna } \nu_i \text{ a } \pi_g \text{ se } d_i^2 \leq \chi^2_{3,\alpha=5‰}, \text{ altrimenti assegna } \nu_i \text{ a } \pi_a.$$

dove d_i^2 è la distanza di Mahalanobis generalizzata definita precedentemente e $\bar{\mathbf{x}}^T = [102.103, 7.249, 5.437]$ è il vettore medio di D/HI, myo-inositolo e scyllo-inositolo rispettivamente, mentre

$$\mathbf{S}^{-1} = \begin{pmatrix} 0.6444 & 0.7470 & -1.8869 \\ 0.7451 & 31.0228 & -18.7499 \\ -1.8858 & -18.7544 & 27.2819 \end{pmatrix}$$

è l'inversa della matrice di dispersione del CRM non adulterato (Tabella 4.2). Ad esempio, un campione con D/H_I, myo-inositolo e scyllo-inositolo, rispettivamente, pari a $\nu_j' = [100, 6.90, 5.30]$ (corrispondenti a un contenuto di 1000 e 200 mg/kg di zuccheri per i due polialcoli) avrà $d_j^2 = 5.91$, inferiore al valore limite di $\chi^2_{3,\alpha=5‰} = 12.84$, e sarà assegnato a π_g; al contrario, un campione sospetto con $\nu_i' = [98, 6.21, 4.38]$ (500 e 80 mg/kg di zuccheri per i due polialcoli) avrà $d_i^2 = 31.77$, superiore alla distanza limite, e sarà assegnato a π_a. Applicando questa regola all'intero set di dati, i risultati sono paragonabili a quelli ottenuti con

l'algoritmo EM: 261 campioni (75.87%) sono stati classificati come genuini e 83 (24.13%) come adulterati.
Di seguito, per completezza, sono rappresentate le distribuzioni e gli istogrammi nel caso in cui si prendano in considerazione le tre variabili separatamente.

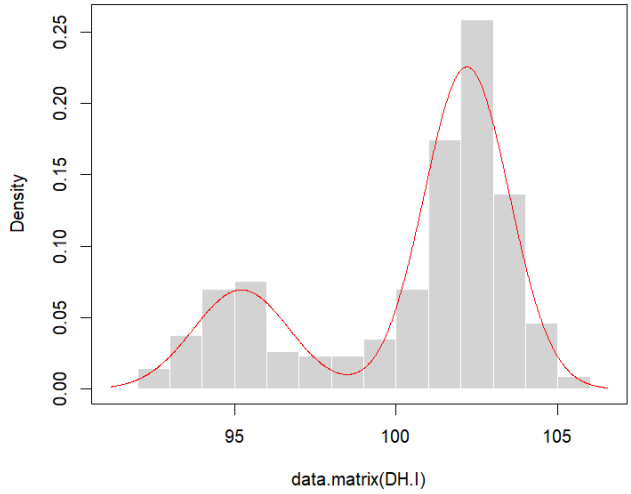


Figura 4.8: Densità stimata per DH_I

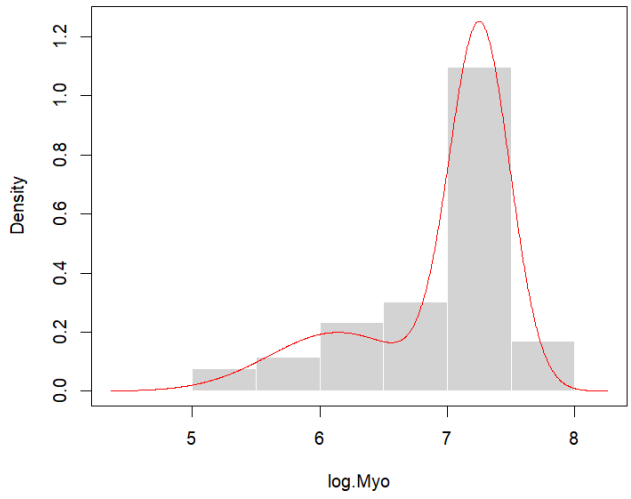


Figura 4.9: Densità stimata per *myo*-inositolo

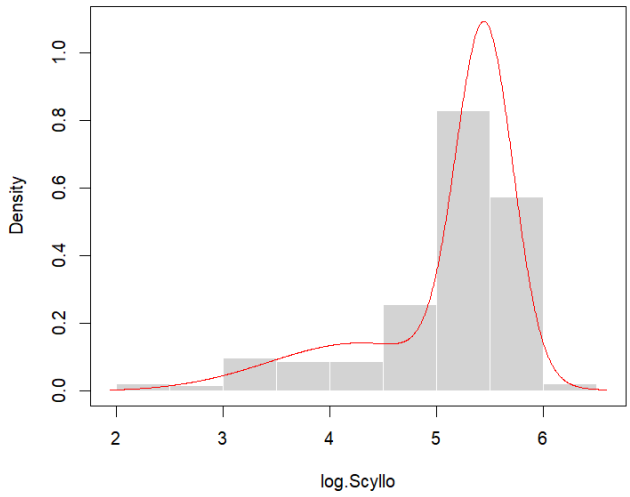


Figura 4.10: Densità stimata per *scyllo*-inositolo

4.3 Script R

```
library(mclust)
library(ggplot2)

data=read.csv(file.choose(), header=TRUE, sep=";")
attach(data)

# Histograms e EDA

par(mfrow = c(2, 2))
hist(Myo, nclass=35, freq = FALSE, main="Histogram of Myo-
   inostitol (mg/kg sugars)",
     xlim=c(0,3000)) #, ylim=c(0,40))
lines(density(Myo), col="red")

hist(Scyllo, nclass=35, freq = FALSE, main="Histogram of Scyllo-
   inostitol (mg/kg sugars)",
     xlim=c(0,600)) #, ylim=c(0,40))
lines(density(Scyllo), col="red")

hist(DH.I, nclass=35, freq = FALSE, main="Histogram of DH.I (ppm)"
   ,
     xlim=c(90, 107))#, ylim=c(0,40))
lines(density(DH.I), col="red")

hist(DH.II, nclass=35, freq = FALSE, main="Histogram of DH.II (ppm
   )",
     xlim=c(120, 132)) #, ylim=c(0,40))
lines(density(DH.II), col="red")

# Trasformazioni logaritmiche e preparazione del data.frame

DH.I=data[,3]
log.Myo=log(data[,1])
log.Scyllo=log(data[,2])

wines=data.frame(DH.I, log.Myo, log.Scyllo)

# Scelta del modello via BIC (Bayesian Information Criterion)

BIC <- mclustBIC(wines, G=2)

par(mfrow = c(1,1))
plot(BIC)

summary(BIC)
```

```
# Stima dei parametri della mistura a tre componenti (log(Myo),
   log(Scyllo) e DH.I) via EM-algorithm

winesMclust <- Mclust(wines, x=BIC, G=2)
summary(winesMclust, parameters = TRUE)

# Identificazione delle sottopopolazioni

class=winesMclust$classification

wines.class=data.frame(wines, class, winesMclust$z)
colnames(wines.class)[c(5,6)] <- c("z.i1","z.i2")
wines.class
plot(winesMclust, what = "classification")

# Sotto-analisi con differenti combinazioni di coppie di variabili
    o singole variabili

# log(Myo) vs. DH.I
data.1 = data.frame(log.Myo, DH.I)

mod.1 = densityMclust(data.1, G=2, x=BIC, plot=FALSE)
summary(mod.1, parameter=TRUE)

par(mfrow = c(2,1))
plot(mod.1, what = "density", data = data.1, drawlabels = TRUE,
   points.pch = 20, col="red", nlevels=30)
#plot(mod.1, what = "density", type = "hdr", data=data.1)
plot(mod.1, what = "density", type = "persp")  # Perspective plot

# log(Myo) vs.log(Scyllo)

data.2=data.frame(log.Scyllo, log.Myo)

mod.2 = densityMclust(data.2, G=2, x=BIC, plot=FALSE)
summary(mod.2, parameter=TRUE)

plot(mod.2, what = "density", data = data.2, drawlabels = TRUE,
   points.pch = 20, col="red", nlevels=20)
#plot(mod.2, what = "density", type = "hdr", data=data.2)
plot(mod.2, what = "density", type = "persp")

# log(Scyllo) vs. DH.I

data.3 = data.frame(log.Scyllo, DH.I)

mod.3 = densityMclust(data.3, G=2, x=BIC, plot=FALSE)
summary(mod.3, parameter=TRUE)
```

```
plot(mod.3, what = "density", data = data.3, drawlabels = TRUE,
   points.pch = 20, col="red", nlevels=30)
#plot(mod.1, what = "density", type = "hdr", data=data.1)
plot(mod.3, what = "density", type = "persp")  # Perspective plot

# DH.I

mod.4 <- densityMclust(data.matrix(DH.I), modelNames = "V", G=2,
   plot=FALSE)
summary(mod.4, parameter=TRUE)

par(mfrow = c(1,1))
plot(mod.4, what = "density", data = DH.I, col="red")
#plot(mod.4, what = "density", type = "persp")

# log(Myo)

mod.5 <- densityMclust(log.Myo, modelNames = "V", G=2, plot=FALSE)
summary(mod.5, parameter=TRUE)

plot(mod.5, what = "density", data = log.Myo, col="red")
#plot(mod.5, what = "density", type = "persp")

# log(Scyllo)

mod.6 <- densityMclust(log.Scyllo, modelNames = "V", G=2, plot=
   FALSE)
summary(mod.6, parameter=TRUE)

plot(mod.6, what = "density", data = log.Scyllo, col="red")
#plot(mod.6, what = "density", type = "persp")
```

Conclusioni

Lo studio presentato nell'ultimo capitolo ha dimostrato l'efficacia dell'applicazione dell'algoritmo EM combinato con l'analisi delle distribuzioni mistura per l'identificazione e la quantificazione delle adulterazioni nei mosti concentrati rettificati d'uva (CRM). La metodologia proposta, che integra l'analisi isotopica con la misurazione dei polialcoli myo-inositolo e scyllo-inositolo, ha permesso di superare i limiti dei metodi tradizionali, offrendo una maggiore precisione nel distinguere tra campioni genuini e adulterati. L'algoritmo EM si è rivelato uno strumento potente per modellare la popolazione di campioni di CRM come una mistura di sottopopolazioni, consentendo di stimare in modo affidabile i parametri delle distribuzioni delle componenti genuina e adulterata. La combinazione dei dati isotopici e dei polialcoli ha fornito una base solida per sviluppare una regola di classificazione basata sulla distanza di Mahalanobis, che ha dimostrato un'elevata capacità discriminante. L'approccio proposto non solo migliora la capacità di rilevare le adulterazioni nei CRM, ma rappresenta anche un esempio significativo di come le tecniche statistiche avanzate possano essere applicate in ambiti pratici come la chimica enologica. La possibilità di identificare adulterazioni con maggiore precisione contribuisce a garantire la qualità e l'integrità dei prodotti enologici, tutelando sia i produttori onesti che i consumatori.

In un contesto più ampio, l'importanza delle distribuzioni mistura e dell'algoritmo EM si estende ben oltre il caso specifico dei CRM. Questi strumenti matematici offrono una straordinaria flessibilità e potenza nella modellizzazione di dati complessi, particolarmente in situazioni in cui la popolazione osservata può essere considerata una combinazione di più sottopopolazioni. La capacità dell'algoritmo EM di gestire dati incompleti o parzialmente osservati, insieme alla sua applicazione in una vasta gamma di discipline, ne fanno uno strumento indispensabile in statistica moderna e nelle scienze applicate.

Dunque, le distribuzioni mistura e l'algoritmo EM non solo forniscono una base teorica robusta per la classificazione e la stima di parametri in contesti complessi, ma rappresentano anche una risorsa pratica fondamentale per affrontare problemi reali in diverse aree, dalla biostatistica all'ingegneria, dalla finanza alla chimica.

Bibliografia

[1] A. P. Dempster, N. M. Laird e D. B. Rubin. «Maximum Likelihood from Incomplete Data Via the EM Algorithm». In: *Journal of the Royal Statistical Society Series B (Methodological)* (1977).

[2] B. Everitt. *Finite mixture distributions*. Springer Science & Business Media, 2013.

[3] P. J. Green. «Introduction to finite mixtures». In: *Chapman and Hall/CRC eBooks*. Chapman e Hall/CRC, 2019.

[4] B. G. Lindsay. *Mixture models: Theory, Geometry, and Applications*. IMS, 1995.

[5] G. J. McLachlan e T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

[6] G. J. McLachlan, S. X. Lee e S. I. Rathnayake. «Finite mixture models». In: *Annual Review of Statistics and Its Application* (2019).

[7] A. Monetti et al. «Sugar Adulterations Control in Concentrated Rectified Grape Musts by Finite Mixture Distribution Analysis of the myo- and scyllo-Inositol Content and the D/H Methyl Ratio of Fermentative Ethanol». In: *Journal of Agricultural and Food Chemistry* (1996).

[8] L. Scrucca et al. «mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models». In: *The R Journal* (2016).

[9] D. M. Titterington, A. F. M. Smith e U. E. Makov. *Statistical analysis of finite mixture distributions*. 1985.