UNIVERSITÀ
DI TRENTO

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

Giuseppe Nicola Liso

Supervisore: Pier Luigi Novi Inverardi

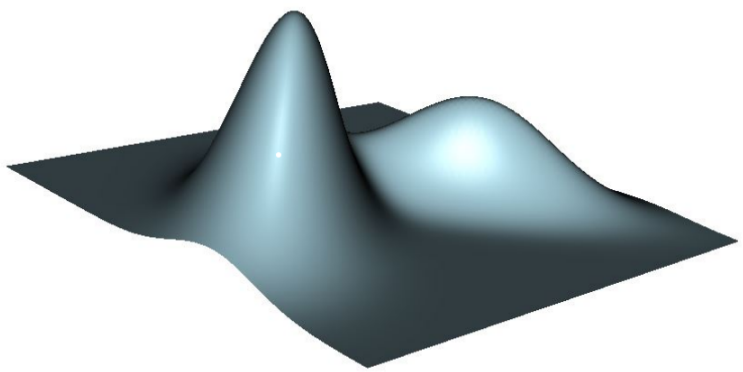
16 Settembre 2024

Indice

Introduzione e definizione delle distribuzioni mistura finite

Motivazione alle distribuzioni mistura finite

Le distribuzioni mistura finite sono fondamentali nell'analisi statistica per la loro capacità di rappresentare dati complessi provenienti da diverse sottopopolazioni. 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. Trovano applicazioni in numerosi campi, grazie proprio alla loro flessibilità, versatilità e propensione all'adattamento ai diversi modelli della realtà.



Motivazione alle distribuzioni mistura finite

Si dividono in:

1. *applicazioni dirette*: si presume l'esistenza di sottopopolazioni nei dati e l'obiettivo è identificare e analizzare queste sottopopolazioni;
2. *applicazioni indirette*: utilizzano le misture come strumento matematico per modellare i dati, ad esempio per generare outlier per verificare la robustezza dei metodi di stima.

Definizione

Una *mistura di distribuzioni finita* è una distribuzione, la cui funzione di densità è data da una combinazione di densità di altre variabili casuali come segue:

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

dove $f_i(\mathrm{x}|\boldsymbol{\theta}_i)$ rappresenta le densità di probabilità delle componenti, $\boldsymbol{\Psi}$ indica il vettore di parametri che si vuole stimare e π_i sono i pesi associati a ciascuna componente, tali che:

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

Metodi di stima

Metodo di massima verosimiglianza

Definizione

Considerato un campione casuale di n osservazioni indipendenti derivanti dalla mistura, la funzione di verosimiglianza è data dalla quantità

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

in cui $\boldsymbol{\Psi}$ indica il vettore di parametri che si vuole stimare.

La funzione di verosimiglianza 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.

Dal momento che si suppone di avere un problema regolare, è possibile affrontare il problema della massimizzazione della funzione di verosimiglianza derivando L_0 rispetto alle componenti di $\boldsymbol{\Psi}$ e ponendo le derivate uguali a zero per ottenere le così dette *equazioni normali*

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

I problemi riscontrabili nel caso di distribuzioni mistura sono:

1. le equazioni normali non sono sempre risolvibili esplicitamente nei parametri Ψ_i;
2. la funzione di verosimiglianza e le derivate parziali $\frac{\partial \mathscr{L}}{\partial \Psi_i}$ non sono limitate.

Dati completi e dati incompleti

I dati completi, ovvero classificati, 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

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

Poichè in presenza di dati mancanti la stima dei parametri del modello mediante massima verosimiglianza può diventare più complessa, viene in aiuto una classe di algoritmi iterativi denominati EM, la cui idea è quella di completare virtualmente $\boldsymbol{Z}$ per poi aggiornare i parametri del modello.

Algoritmo EM

L'algoritmo *Expectation-Maximization* (EM) è un algoritmo iterativo composto da:

- Passo E: si calcolano i valori attesi condizionati dei dati mancanti rispetto ai dati osservati e alle corrette stime dei parametri di interesse e si sostituiscono i valori mancanti con quelli attesi.
- Passo M: si calcolano le stime di massima verosimiglianza dei parametri sui dati completi.

Definizione

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 $\boldsymbol{y}$. 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:

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

- Pregi dell'algoritmo EM: non serve calcolare e invertire matrici, è molto semplice da costruire e da implementare in un calcolatore.
- Difetti dell'algoritmo EM: l'algoritmo risulta efficiente quando il passo E si può calcolare direttamente e la velocità di convergenza può risultare molto lenta.

Un'applicazione a dati reali

Contesto del CRM e delle adulterazioni

Un esempio di **applicazione diretta** delle distribuzioni mistura è quello legato all'adulterazione del CRM. Il mosto concentrato rettificato (CRM) può essere adulterato con zuccheri esogeni. Lo studio propone un nuovo metodo per rilevare queste adulterazioni combinando l'analisi isotopica e la misurazione di polialcoli (myo-inositolo e scyllo-inositolo), utilizzando un algoritmo EM per distinguere con precisione il CRM genuino da quello adulterato.

Algoritmo EM e distribuzioni mistura

Per stimare i parametri della mistura è stato utilizzato l'algoritmo EM come segue:

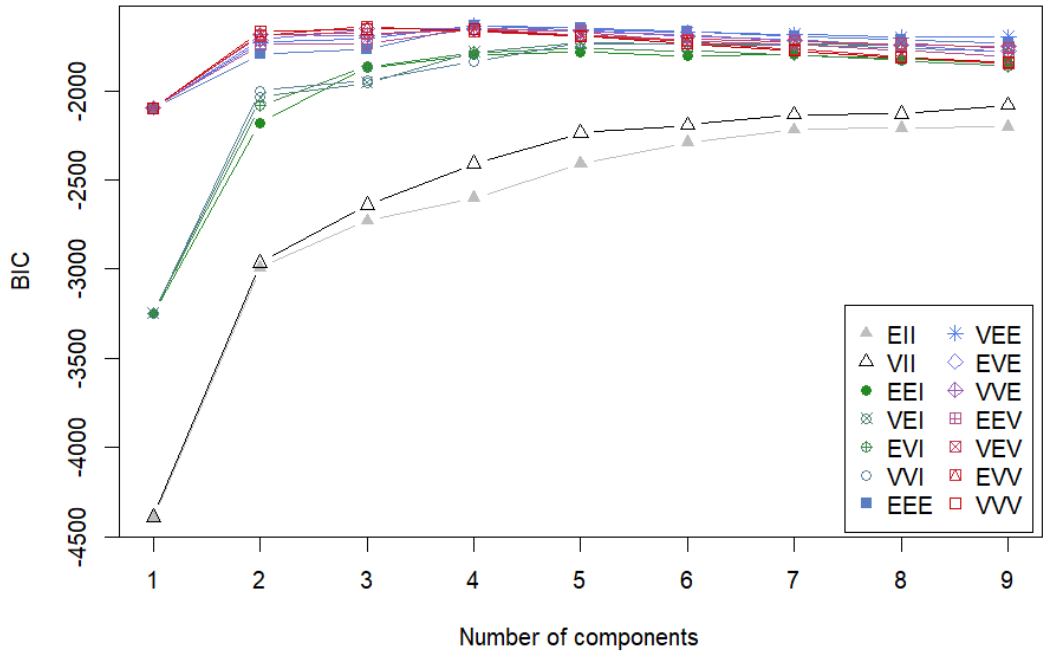
1. **Fase di Aspettazione (E-step):** Calcolare le probabilità di appartenenza di ciascun campione a ciascuna componente della mistura;
2. **Fase di Massimizzazione (M-step):** Aggiornare i parametri delle distribuzioni mistura massimizzando la funzione di verosimiglianza completa con le probabilità di appartenenza calcolate nella fase E.

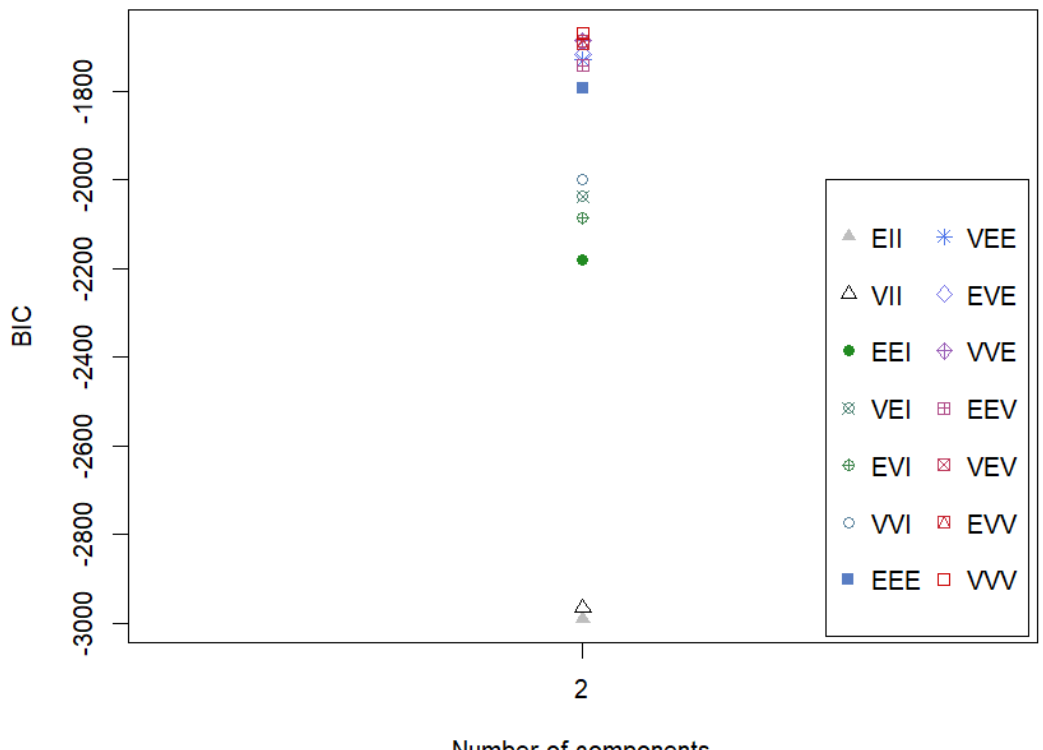
Scelta del modello

Il *Bayesian Information Criterion* (BIC) è un criterio utilizzato per scegliere il miglior modello tra diversi modelli statistici. Ha l'obiettivo di bilanciare due aspetti fondamentali:

- **Qualità dell'adattamento**: Un buon modello deve adattarsi bene ai dati osservati.
- **Semplicità del modello**: Aggiungere troppi parametri può rendere un modello complesso e portare al rischio di *overfitting*.

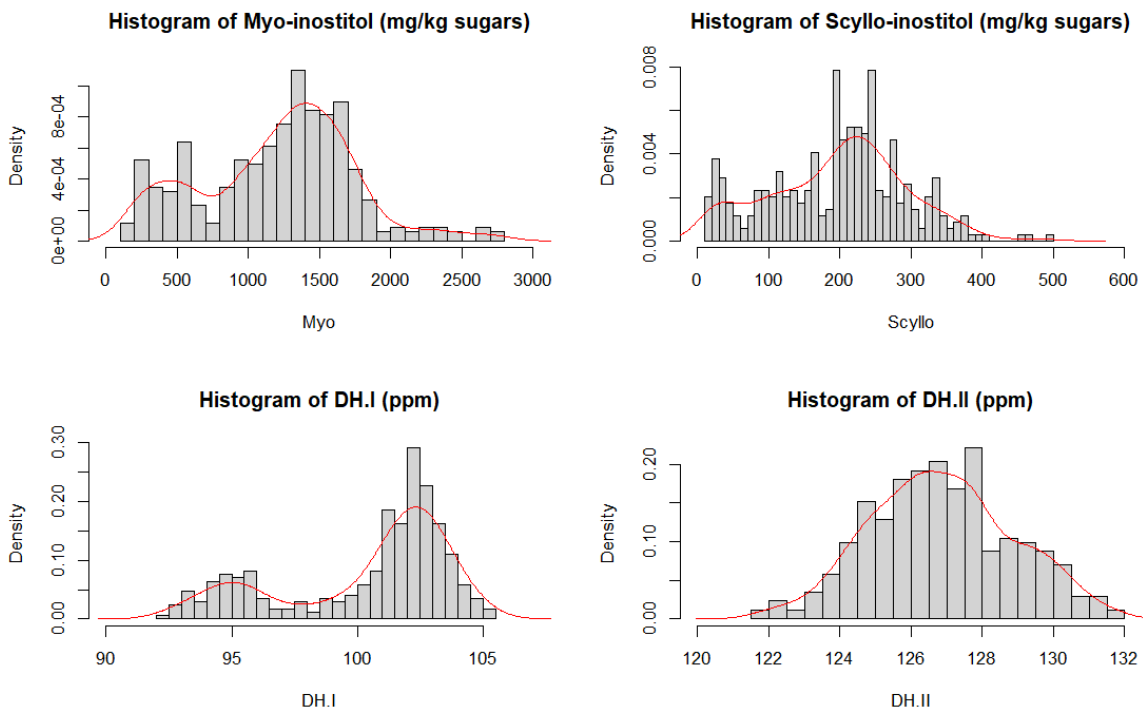
In sintesi, il BIC cerca di trovare un equilibrio tra quanto bene un modello si adatta ai dati e quanto è semplice. Un BIC più piccolo indica un modello migliore. In questo caso, il modello VVV, che permette a volume, forma e orientamento di variare per ogni gruppo, si adatta meglio ai dati.


BIC
-2000
-2500
-3000
-3500
-4000
-4500
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV


BIC
-1800
-2000
-2200
-2400
-2600
-2800
-3000
2
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV

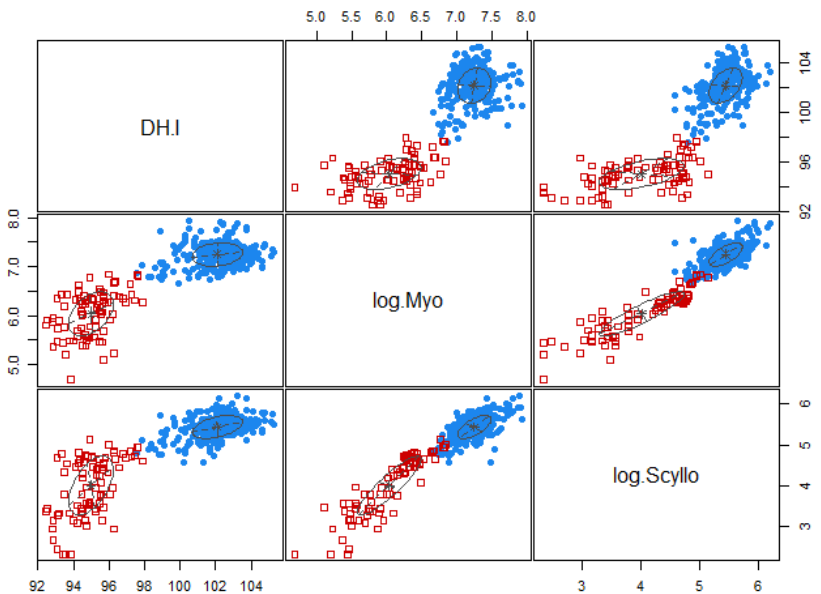
Analisi dei risultati

Tra le quattro variabili prese in considerazione, cioè *myo-inositolo*, *scyllo-inositolo*, D/H_I e D/H_{II}, il cui comportamento è mostrato in figura, D/H_{II} è stato scartato per il suo scarso potere discriminante, confermato anche dalla sua distribuzione marginale delle frequenze.



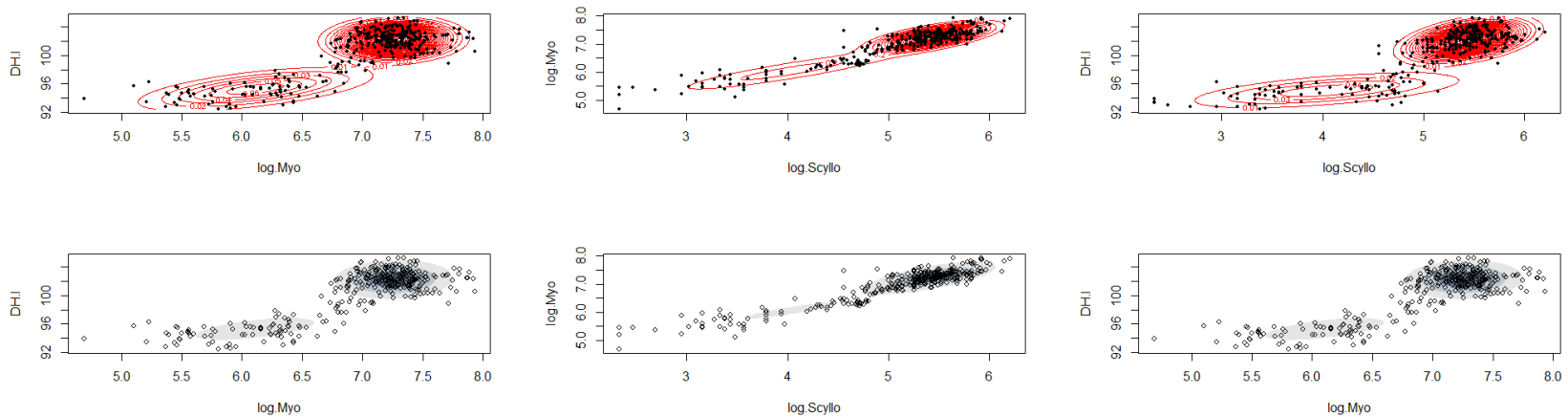
Identificazione delle sottopopolazioni tramite Clustering

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.



		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

Nella Tabella è possibile vedere come il D/H_I, il myo-inositolo e lo scyllo-inositolo hanno bisogno di un numero di iterazioni elevato, mentre nel modello trivariato la convergenza si raggiunge con sole 11 iterazioni; 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.



Conclusioni

Lo studio descritto ha dimostrato l'efficacia dell'algoritmo EM combinato con l'analisi delle distribuzioni a mistura per identificare e quantificare adulterazioni nei mosti concentrati rettificati d'uva (CRM). Questo approccio, applicato anche ad altre discipline, sottolinea la potenza dell'algoritmo EM e delle distribuzioni a mistura nella modellizzazione di dati complessi e nella risoluzione di problemi pratici.

Bibliografia

- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*.
- Everitt, B. (2013). *Finite Mixture Distributions*. Springer Science & Business Media.
- McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application*.
- Green, P. J. (2019). Introduction to Finite Mixtures. In *Chapman and Hall/CRC eBooks*.
- Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. IMS.
- Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*.
- Monetti, A., Versini, G., Dalpiaz, G., and Reniero, F. (1996). Sugar Adulterations Control in Concentrated Rectified Grape Musts. *Journal of Agricultural and Food Chemistry*.
- McLachlan, G. J., and Krishnan, T. (2007). *The EM Algorithm and Extensions*. John Wiley & Sons.

Ringraziamenti

Grazie per l'attenzione