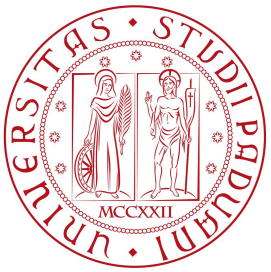
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per l'Economia e l'Impresa



Analisi dei gruppi basata sul modello in presenza di dati mancanti: un'applicazione ad indicatori macroeconomici

Relatrice: Prof.ssa Manuela Cattelan
Dipartimento di Scienze Statistiche

Laureanda: Alice Cappella
Matricola n. 1226863

Anno Accademico 2021/2022

A mio nonno Romano,
maestro di vita

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Raggruppare persone, animali o più in generale oggetti che condividono caratteristiche simili in gruppi rappresenta una delle abilità umane più primitive. Da un punto di vista prettamente statistico, questo raggruppamento può essere effettuato con una serie di metodologie che rientrano in quella che viene chiamata *cluster analysis* o *analisi dei gruppi*.

L'analisi dei gruppi ha differenti obiettivi, tutti legati alla suddivisione di una collezione di oggetti o di unità statistiche in gruppi chiamati *cluster* in modo tale che le unità facenti parte di un gruppo siano più simili rispetto alle unità appartenenti ad un gruppo differente.

Questa metodologia statistica ha applicazioni in molte discipline, dalla biologia, la medicina, la psicologia fino all'economia, al marketing ma anche all'elaborazione di immagini.

La *cluster analysis* viene spesso utilizzata anche come semplice analisi esplorativa, per comprendere se i dati possano o meno consistere in una serie di gruppi distinti.

L'obiettivo di questa tesi è quello di descrivere una delle metodologie dell'analisi dei gruppi, il *model-based clustering*, ed applicare quest'ultima ad un dataset contenente importanti indicatori macroeconomici relativi alla maggior parte dei Paesi europei. Questo permetterà di capire se è possibile ottenere un raggruppamento ragionevole in base alla situazione economica che caratterizza i vari Paesi europei considerati nel dataset.

Nel capitolo 1 viene presentato il dataset economico e vengono descritte tutte le variabili presenti in esso. Verrà svolta una prima analisi esplorativa e verranno esposte alcune problematiche del dataset, in particolare la presenza di dati mancanti.

Nel capitolo 2 viene trattata a livello teorico la *cluster analysis* e le metodologie utilizzate per la sua applicazione. Dopo aver illustrato alcune limitazioni legate alle principali procedure di *clustering*, verrà presentata e descritta in modo più approfondito la metodologia dell'analisi dei gruppi basata su modello e la sua applicazione in un contesto

in cui vi siano valori mancanti.

Gli ultimi due capitoli sono dedicati all'implementazione della metodologia descritta nel capitolo 2.

In particolare, nel capitolo 3 verranno presentati e commentati i risultati ottenuti dall'applicazione del *model-based clustering* al dataset economico. Mentre nel capitolo 4, verrà effettuato uno studio di simulazione considerando differenti scenari in modo tale da poter fare una valutazione più puntuale sulla performance dell'analisi dei gruppi basata su modello e dei pacchetti utilizzati per la sua implementazione.

Capitolo 1

Il dataset

In questo capitolo verrà presentato il dataset utilizzato con l'obiettivo di ottenere un raggruppamento dei Paesi europei tenendo in considerazione alcuni dei più importanti indicatori macroeconomici. In questo modo i Paesi appartenenti ad uno stesso gruppo saranno più simili dal punto di vista economico. La descrizione del dataset verrà poi accompagnata da un'analisi esplorativa dello stesso.

1.1 Indicatori macroeconomici

Il dataset costruito per questa tesi contiene una serie di importanti indicatori macroeconomici relativi all'anno 2019 per la maggior parte dei Paesi europei; in particolare, include 18 variabili quantitative per 39 Paesi europei. Sono stati considerati i valori relativi all'anno 2019 in modo tale che la suddivisione dei Paesi non fosse influenzata dalla pandemia SARS-CoV-2. Inoltre, dati più recenti costituiscono ancora stime o previsioni e, di conseguenza, non avrebbero reso l'analisi del tutto accurata.

Gli indicatori considerati sono fondamentali per effettuare analisi macroeconomiche e fornire quindi un quadro completo della situazione economica dei Paesi considerati, oltre che per prendere importanti decisioni di natura politica.

Il dataset include il PIL (GDP), acronimo di Prodotto Interno Lordo, ai prezzi di mercato. Si tratta dell'indicatore maggiormente utilizzato per sintetizzare la situazione economica di un Paese e rappresenta il risultato finale della produzione delle unità produttrici residenti in un dato periodo temporale, che solitamente corrisponde all'anno solare.

È quindi l'insieme di beni e servizi prodotti in un Paese, non considerando i consumi intermedi, ossia tutti quei beni e servizi che sono stati utilizzati come input nel processo produttivo.

La popolarità di questo indicatore è data dal fatto che oltre a sintetizzare la produzione economica di un Paese in un'unica misura facilmente interpretabile, è basata su metodologie rigide condivise a livello internazionale rendendolo confrontabile per diverse realtà economiche.

È necessario specificare che non tutte le attività economiche rientrano nel calcolo di questo indicatore. Di conseguenza, nonostante il PIL costituisca sicuramente uno dei più importanti indicatori economici, non è esaustivo per determinare il livello di benessere economico di una società.

Il PIL ai prezzi di mercato può essere calcolato mediante tre approcci differenti:

1. Metodo del valore aggiunto:

$$\begin{aligned} PIL = {} & \textit{valore aggiunto} \\ & + \textit{imposte sui prodotti} - \textit{sussidi sui prodotti}, \end{aligned} \tag{1.1}$$

2. Metodo della spesa:

$$\begin{aligned} PIL = {} & \textit{spesa per consumi finali} + \textit{investimenti lordi} \\ & + \textit{esportazioni} - \textit{importazioni}, \end{aligned} \tag{1.2}$$

3. Metodo del reddito:

$$\begin{aligned} PIL = {} & \textit{compensazione dei lavoratori} \\ & + \textit{risultato lordo di gestione e reddito misto lordo} \\ & + \textit{imposte sulla produzione e importazioni} \\ & - \textit{sussidi sulla produzione e importazioni}. \end{aligned} \tag{1.3}$$

Tutti e tre gli approcci portano allo stesso valore del PIL.

Vengono poi considerate nel dataset le componenti del PIL ossia:

- GVA: valore aggiunto lordo, dato dalla differenza tra il valore della produzione finale e i consumi intermedi.
- $T.S\_P$: tasse meno sussidi sui prodotti.
- FCE_GG: spesa per consumi finali delle pubbliche amministrazioni.
- FCE_NPISH_H: spesa per consumi finali delle famiglie e delle istituzioni sociali private al servizio delle famiglie.
- AIC: spesa per consumi individuali effettivi.
- $GFCF$: investimenti fissi lordi, dato dalle acquisizioni, al netto delle cessioni, di capitale fisso effettuato da produttori residenti a cui vengono aggiunti gli incrementi di valore dei beni non prodotti.
- $CI\_A.DV$: variazioni delle scorte e acquisizioni meno cessioni di oggetti di valore. La variazione delle scorte è calcolata come differenza tra il valore delle entrate nel magazzino e quello delle uscite dal magazzino. A questa variazione viene sommata la differenza tra acquisizioni e cessioni di oggetti di valore, ossia quei beni non finanziari utilizzati in modo secondario per la produzione o per il consumo.
- E_GS: esportazioni di beni e servizi.
- I_GS: importazioni di beni e servizi.
- CE: compensazione dei lavoratori, ovvero il costo che il datore di lavoro deve sostenere per remunerare il lavoratore per le attività manuali ed intellettuali svolte.
- GOS_GMI: risultato lordo di gestione e reddito misto lordo, ossia la parte del valore aggiunto prodotto destinata a remunerare i fattori produttivi diversi dal lavoro dipendente impiegati nel processo di produzione.
- $T.S\_PI$: tasse meno sussidi sulla produzione e sulle importazioni.

Altri due importanti indicatori sono la bilancia dei pagamenti e l'indice dei prezzi al consumo armonizzato.

La bilancia dei pagamenti (BOP) riassume le transizioni economiche avvenute tra residenti e non residenti di un determinato Paese in un dato periodo di tempo. Si tratta di un importante indicatore economico in quanto consente di valutare la posizione economica (di debito o di credito) di un Paese nei confronti del resto del mondo.

Mentre l'indice dei prezzi al consumo armonizzato (*HICP*) è un indicatore fortemente legato all'*inflazione*, ossia il processo generalizzato di aumento dei prezzi di un insieme rappresentativo di beni e servizi appartenenti al cosiddetto *paniere dei consumatori*. Misura quindi la variazione in termini percentuali dei prezzi dei beni appartenenti al paniere. Viene calcolato secondo un approccio armonizzato a livello europeo consentendo il confronto dell'inflazione tra Paesi. La stabilità dei prezzi, intesa come aumento annuo dell'indice *HICP* e definita dalla Banca Centrale Europea, è del 2%.

Sono infine presenti gli indicatori del mercato del lavoro. La forza lavoro e, più in generale, il mercato del lavoro sono aspetti fondamentali per l'economia di ogni Paese. In particolare, nel dataset sono riportati:

- *EMPL*: tasso di occupazione, ossia il numero di individui con un'attività lavorativa all'interno di un Paese. Il tasso di occupazione, insieme ad indicatori come il PIL e l'indice dei prezzi al consumo (armonizzato), forniscono una sintesi della situazione economica di un Paese. Maggiore è il tasso di occupazione migliore sarà l'uso della manodopera disponibile. A livello sociale, l'aumento di questo tasso comporta un aumento del reddito totale percepito che contribuisce a favorire la crescita economica.

- *LFA*: tasso di attività, ossia la proporzione di individui attivi all'interno di una popolazione, dove per individui attivi si intende la parte della popolazione in grado di svolgere un'attività lavorativa.

- *UNEMPL*: tasso di disoccupazione, ossia la proporzione di disoccupati di età compresa tra i 15 e i 74 anni all'interno della forza lavoro. Costituisce un importante indicatore economico e sociale e rappresenta la manodopera disponibile, in termini di persone, rimasta inutilizzata.

È necessario fare delle specificazioni su alcune variabili presenti nel dataset.

Innanzitutto, nel calcolo del PIL mediante il metodo della spesa (formula 1.2) viene coinvolta la *spesa per consumi finali*. Quest'ultima è data dalla somma delle variabili *FCE_GG* e *FCE_H_NPISH*.

In secondo luogo, la somma delle variabili *GFCF* e *CI_A.DV* forma gli *investimenti lordi* o *formazione lorda del capitale*. Gli investimenti lordi, necessari per il calcolo del PIL con il metodo del reddito (formula 1.3), consistono nel valore dei beni materiali acquisiti dalle unità produttive che procureranno reddito.

Infine si precisa che il PIL, le sue componenti e il bilancio dei pagamenti sono variabili espresse in milioni di Euro, mentre l'indice dei prezzi al consumo armonizzato e tutti gli indicatori del mercato del lavoro sono espressi in termini percentuali.

La quasi totalità dei dati presenti all'interno del dataset sono stati reperiti dal database di *Eurostat*. Eurostat provvede al rilascio di molte informazioni a livello economico, demografico e sociale con riferimento ai Paesi europei. Tuttavia, per quanto riguarda gli indicatori del mercato del lavoro, in fase di costruzione del dataset i valori per la Gran Bretagna non sono risultati disponibili a causa di una mancata deliberazione di questi dati ad Eurostat da parte del Paese. Nonostante questo è stato possibile recuperare queste informazioni tramite *ONS* (*Office for Nation Statistics*) che provvede a rilasciare una collezione di statistiche sulla situazione economica e sociale solo relativamente al Regno Unito.

1.2 Analisi esplorativa dei dati

Dalla tabella 1.1 emerge che, come anticipato nel paragrafo precedente, le variabili non sono espresse nella stessa unità di misura. Sarà quindi necessario, in fase di applicazione della *cluster analysis*, standardizzare i dati.

Dai valori delle statistiche di riepilogo si evidenzia che la variabile *CI.A_DV* può assumere anche valori negativi. Questo è dovuto al fatto che questa variabile è frutto di una differenza. Anche la variabile *BOP* assume valori negativi, un saldo della bilancia dei pagamenti negativo riflette una situazione di indebitamento.

A differenza di tutte le altre variabili *HICP* e tutti i tassi del mercato del lavoro, essendo espressi in percentuale, potranno assumere valori in $[0, 100]$.

Inoltre, dall'ultima colonna della tabella, si nota la presenza di dati mancanti per alcune variabili.

Variabile	**Min**	$\mathbf{q_{0.25}}$	**Mediana**	**Media**	$\mathbf{q_{0.75}}$	**Max**	**NA**
GDP	4951	29206	183250	470938	477554	3473260	0
GVA	4022	25406	158762	422287	425243	3129717	0
FCE_GG	881	6348	36575	96135	94122	703156	1
FCE_NPISH_H	3534	18616	104995	263690	237923	1805463	1
AIC	3944	20582	129897	320618	302871	2256364	2
GFCF	1352	7700	39194	106149	116184	742361	1
CI_A.DV	-7271	27	1343	3052	3560	25943	1
E_GS	2068	21259	94464	224649	278691	1620957	1
I_GS	3218	21559	97300	210496	254667	1424620	1
CE	3495	24735	98912	255480	222715	1853274	5
GOS_GMI	2038	20805	100628	220829	250538	1276923	5
T.S_PI	208.6	6258	23769	61255	58517	343063	5
T.S_P	196	3800	21372	48705	46746	343543	0
BOP	-77684	-737	850	10197	7431	262904	1
HICP	0.3	0.8	1.7	2.1	2.3	15.2	5
LFA	58.4	69.5	74.7	74	77.7	87.3	4
EMPL	50.3	64.3	70.5	69.1	75.1	84.1	4
UNEMPL	2	2.4	5.5	6.9	7.1	17.9	4
Totale							41

Tabella 1.1: Statistiche descrittive delle variabili macroeconomiche: minimo (min), primo quartile ($q_{0.25}$), mediana, media, terzo quartile ($q_{0.75}$), valori mancanti (NA)

Il grafico riportato nella figura 1.1 fornisce una rappresentazione della matrice di correlazione. Si sottolinea che le celle di colore bianco indicano le correlazioni che assumono un valore vicino allo 0.

È prevedibile che le componenti del PIL siano fortemente correlate positivamente con quest'ultimo; le componenti infatti entrano additivamente nel calcolo del PIL. Per evitare problemi di perfetta collinearità verrà pertanto eliminata la variabile *GDP* prima dell'applicazione della *cluster analysis*. Due variabili perfettamente correlate non portano più informazione di quanto non ne faccia una sola delle due. Pertanto eliminare una delle due variabili che riportano una perfetta correlazione non comporta una grande perdita di informazione.

È poi possibile notare che il tasso di occupazione è altamente correlato negativamente con il tasso di disoccupazione. Tant'è vero che all'aumentare del tasso di occupazione il tasso di disoccupazione decresce e viceversa. Inoltre, il tasso di occupazione risulta essere correlato negativamente con il tasso di attività.

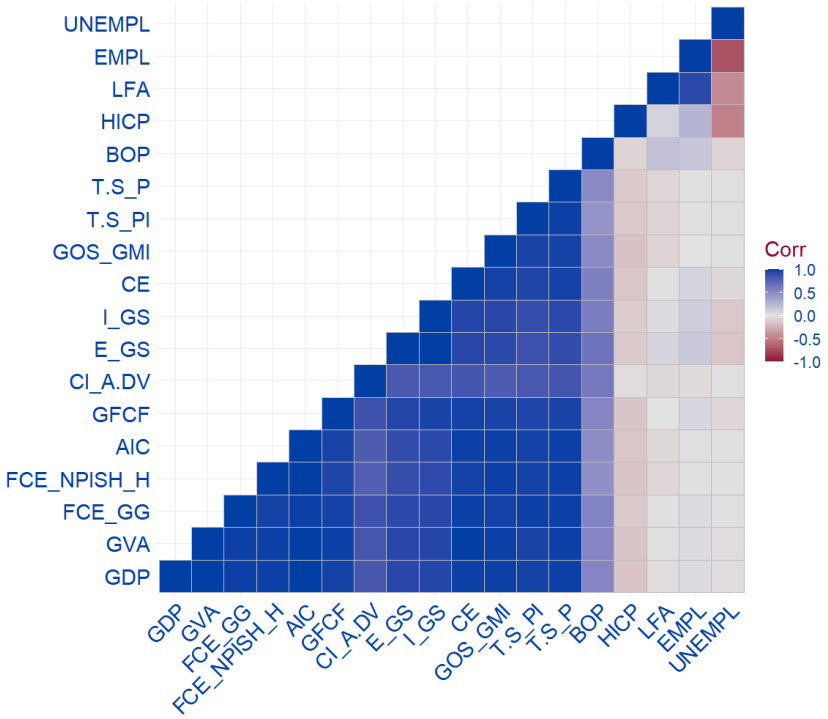


Figura 1.1: Rappresentazione grafica della matrice di correlazione

1.2.1 Analisi grafica univariata

Per iniziare a conoscere la situazione economica dei Paesi considerati nel dataset e quindi i valori delle variabili per ogni Paese vengono di seguito riportate alcune rappresentazioni grafiche univariate.

Come ci si poteva aspettare, dalla figura 1.2 emerge che la variabile *GDP* assume il suo valore massimo per la Germania. Quest'ultima, infatti, risulta essere la prima economia a livello europeo. Seguono Regno Unito, Francia e Italia.

In quanto all'indice dei prezzi al consumo armonizzato (figura 1.3), che caratterizza un aumento dei prezzi dei beni appartenenti al paniere (inflazione) o un eventuale decremento (deflazione), emerge dal grafico che, a differenza di tutti gli altri Paesi presenti nel dataset, vi è stato un netto aumento dell'inflazione in Turchia dall'anno 2018 all'anno 2019.

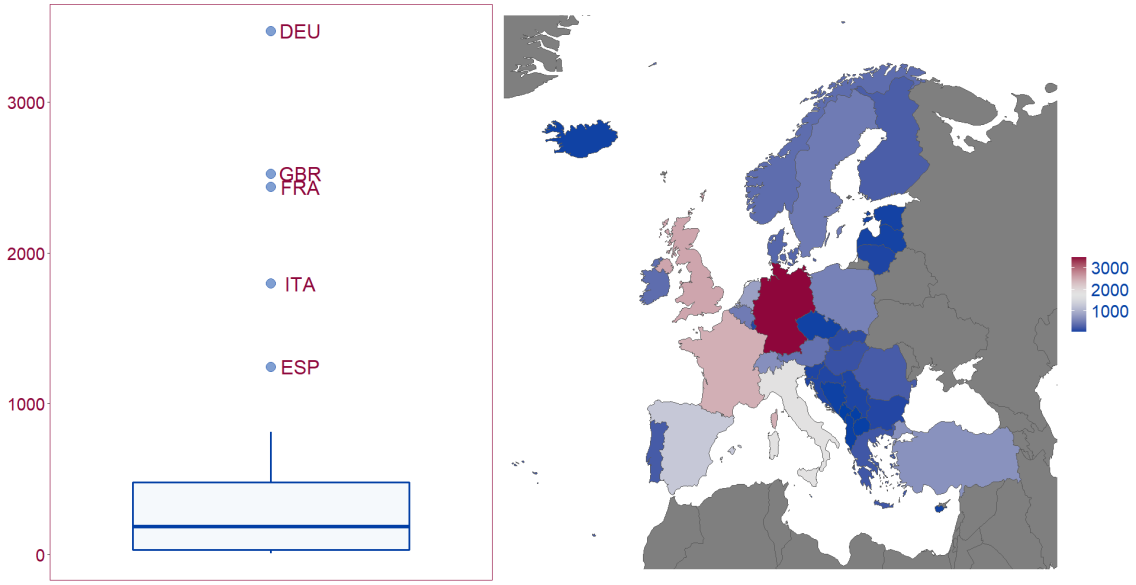


Figura 1.2: Boxplot (a sinistra) e rappresentazione della distribuzione geografica (a destra) della variabile *GDP* espressa in migliaia di milioni di Euro

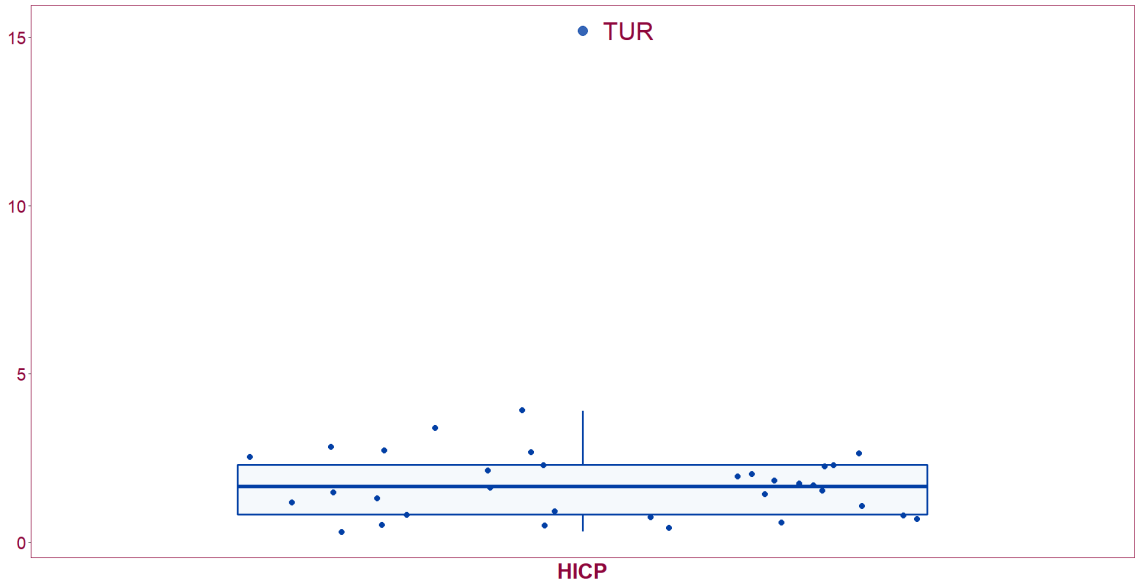


Figura 1.3: Boxplot della variabile *HICP*

Il Paese con il maggior tasso di occupazione è l'Islanda, seguita da Svizzera e Olanda. Mentre quelli con il valore minore di questo tasso risultano, nell'ordine, Turchia, Macedonia, Montenegro e Grecia.

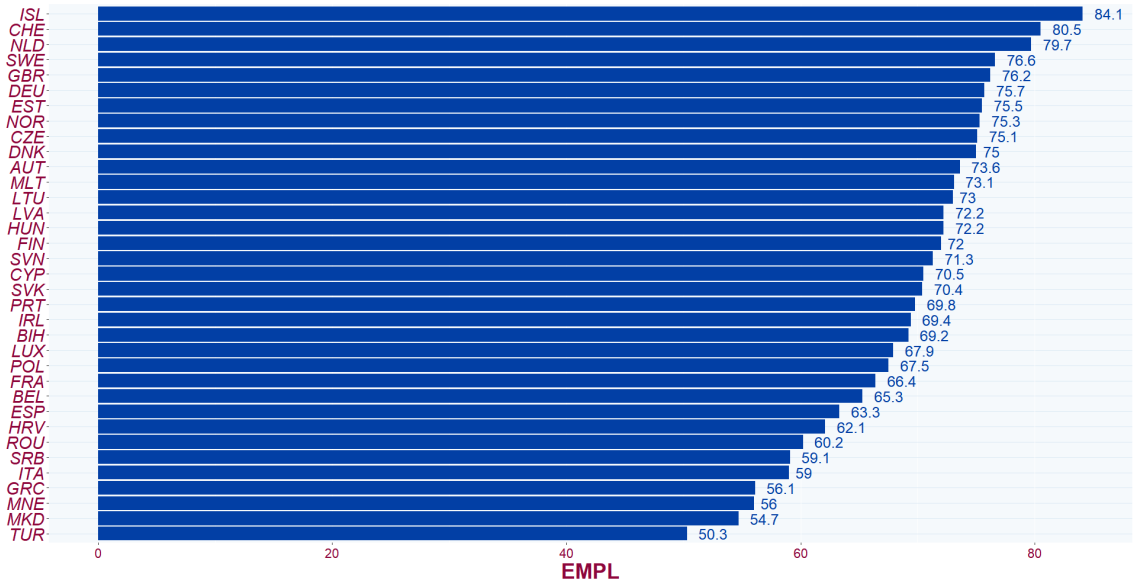


Figura 1.4: Rappresentazione grafica della variabile *EMPL*

Infine, ricordando che all'aumentare del tasso di occupazione il tasso di disoccupazione diminuisce, si nota che i Paesi con il maggior tasso di disoccupazione sono Grecia, Macedonia, Montenegro, Spagna e Turchia.

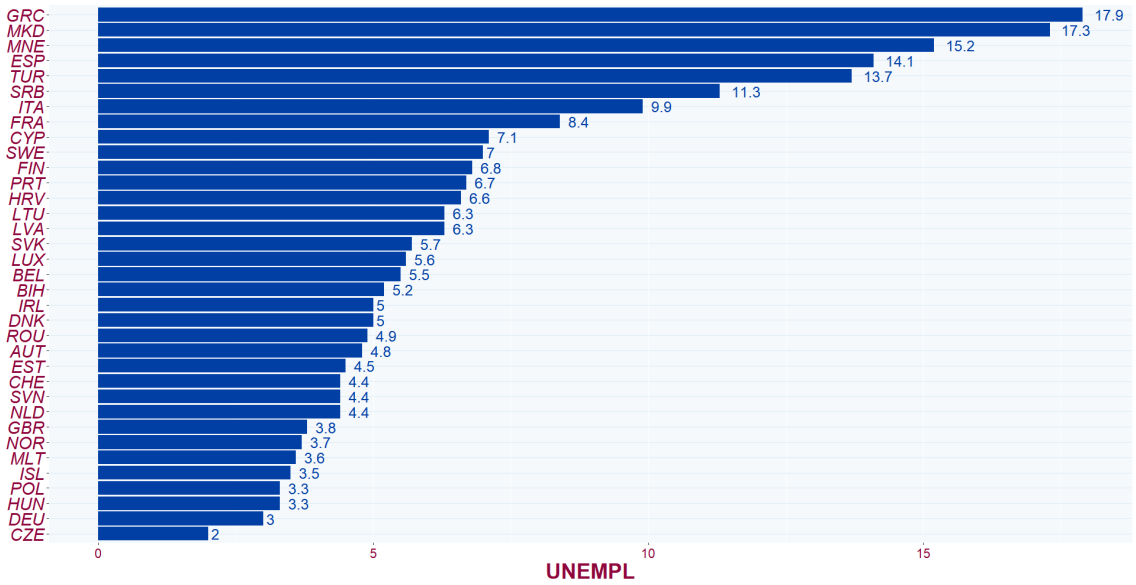


Figura 1.5: Rappresentazione grafica della variabile *UNEMPL*

Nei grafici in figura 1.3, 1.4 e 1.5 non sono stati considerati alcuni Paesi in quanto in corrispondenza delle variabili rappresentate riportano un valore mancante. In particolare, nella figura 1.3 non sono presi in considerazione Albania, Bosnia e Herzegovina, Kosovo, Liechtenstein e Montenegro, mentre nei grafici rappresentati nelle figure 1.4 e 1.5 Albania, Bulgaria, Kosovo e Liechtenstein.

1.2.2 Analisi grafica bivariata

Vengono di seguito riportate alcune rappresentazioni grafiche bivariate che consentono di comprendere la relazione esistente tra le variabili presenti nel dataset.

Dalla rappresentazione grafica della matrice di correlazione (figura 1.1) è emerso che le variabili *GDP*, ossia il PIL, e *BOP*, ovvero la bilancia dei pagamenti, sono correlate, anche se non fortemente, in senso positivo. Questo si riflette nel grafico riportato in figura 1.6.

Valori bassi del PIL corrispondono effettivamente a valori bassi della variabile relativa al bilancio dei pagamenti. Analogamente per i valori più elevati, come nel caso della Germania. Quello che si può infine notare è che in un caso il valore elevato del PIL non corrisponde ad un valore positivo di *BOP*. Questa è la situazione che si riscontra per il Regno Unito che, nonostante assuma il secondo valore più alto del PIL, si trova in una condizione di indebitamento.

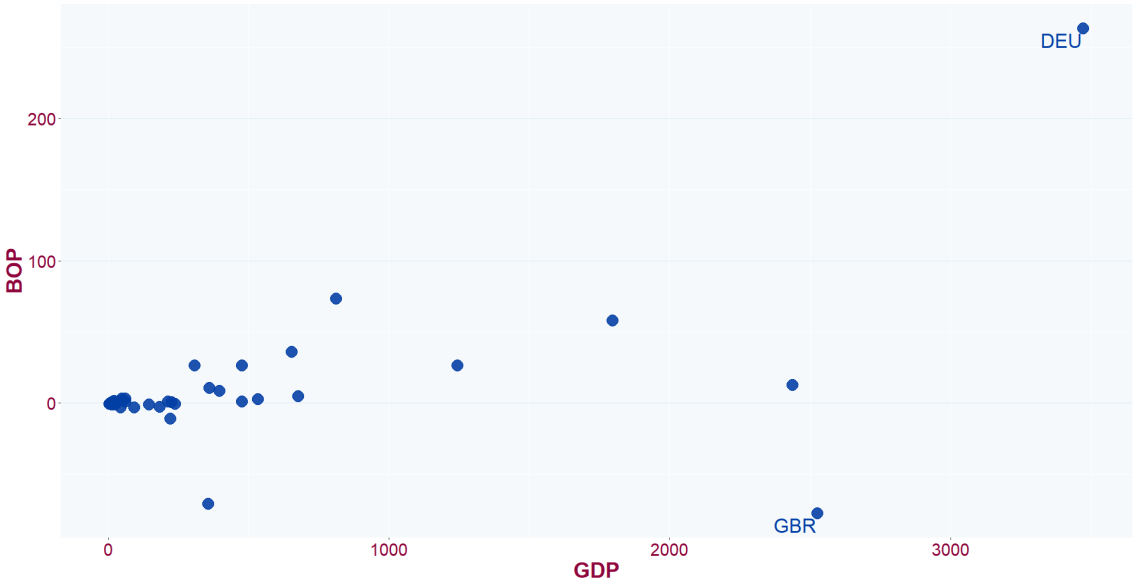


Figura 1.6: Grafico di dispersione delle variabili *GDP* e *BOP* espresse in migliaia di milioni di Euro

Si sottolinea inoltre che il Liechtenstein non è stato considerato per la rappresentazione grafica in figura 1.6 in quanto in corrispondenza della variabile *BOP* presenta un valore mancante.

La correlazione negativa tra il tasso di disoccupazione e il tasso di attività è evidenziato anche dal grafico 1.7. È chiaro che all'aumentare della proporzione di individui attivi in una popolazione, e quindi all'aumentare del tasso di attività, vi sia una diminuzione del tasso di disoccupazione. Queste due dimensioni sono strettamente legate ed è da sottolineare che i Paesi che presentano un alto tasso di attività, corrispondente ad un basso valore del tasso di disoccupazione, saranno caratterizzati da un buon impiego della manodopera disponibile.

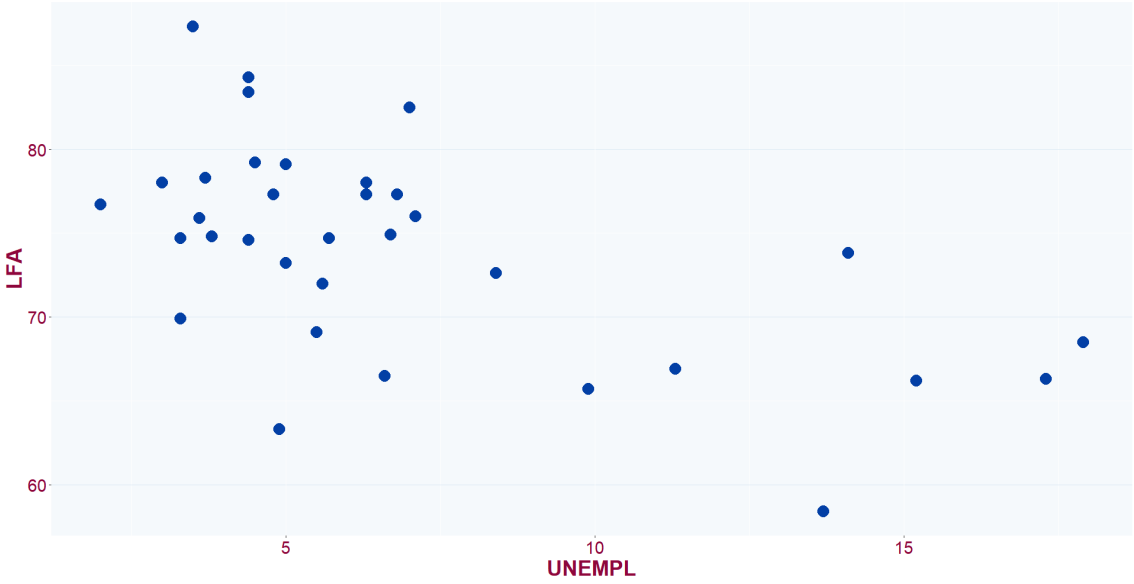


Figura 1.7: Grafico di dispersione delle variabili *UNEMPL* e *LFA*

Nel grafico in figura 1.7 non sono stati considerati Albania, Bosnia e Herzegovina, Kosovo, Liechtenstein e Montenegro in quanto in corrispondenza delle variabili rappresentate riportano un valore mancante.

1.3 Dati mancanti

In corrispondenza di alcune variabili non tutti i valori risultano disponibili. Ne consegue la presenza di alcuni valori mancanti. Rispetto al totale dei valori presenti nel dataset, i valori mancanti rappresetano all'incirca il 6%.

La figura 1.8 rappresenta graficamente il pattern dei valori mancanti all'interno del dataset.

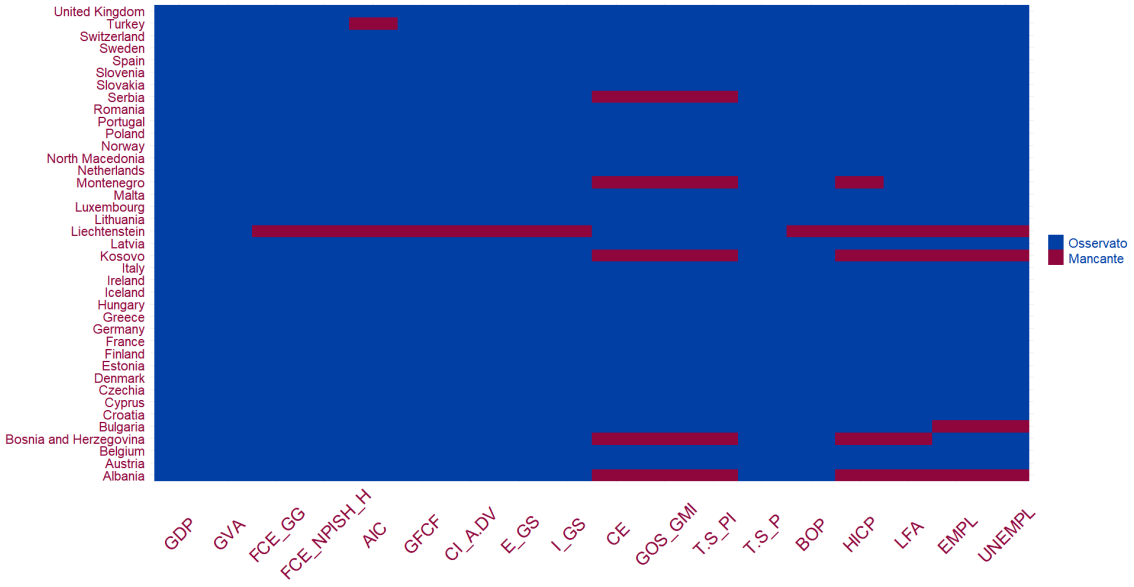


Figura 1.8: Distribuzione dei valori mancanti nel dataset

Nel contesto dei dati mancanti è usuale parlare di *listwise deletion* o *analisi dei casi completi*, un metodo di trattamento di dati mancanti che prevede di eliminare le unità statistiche per cui, in almeno una variabile, sia presente un valore mancante. Tuttavia, usufruire di questa metodologia per trattare i valori mancanti comporterebbe una perdita di informazione.

Si noti inoltre come i dati mancanti non siano distribuiti in modo casuale all'interno della matrice dei dati ma piuttosto come siano concentrati su alcuni Paesi e variabili. In particolare, sono presenti molti valori mancanti in corrispondenza di Liechtenstein, Kosovo e Albania.

È da sottolineare che i Paesi membri dell'Unione Europea e gli stati EFTA (*European Free Trade Association*), come definito nel *Programma di trasmissione dei dati* del *Sistema europeo dei conti SEC 2010*, hanno l'obbligo legale di fornire i dati ad Eurostat. Questo, eccetto nel caso del Liechtenstein che fa parte degli stati EFTA, può giustificare

la presenza dei dati mancanti in corrispondenza di alcune variabili per Serbia, Montenegro, Kosovo, Bosnia e Herzegovina e Albania.

Inoltre, è possibile notare come molti dei valori mancanti presenti nel dataset siano concentrati nelle variabili *CE, GOS_GMI, T.S_PI, HICP* e gli indicatori del mercato del lavoro.

Per evitare una perdita di informazione, sarà quindi necessario adottare una metodologia che consenta di effettuare l'imputazione dei dati mancanti.

Nell'ambito della *cluster analysis* basata su modello esistono diversi algoritmi che possono essere utilizzati in presenza di valori mancanti. In questa tesi verrà utilizzato l'algoritmo *EM*, un metodo iterativo che permette di ottenere le stime di massima verosimiglianza in presenza di dati incompleti.

Capitolo 2

Cluster Analysis

Le prime procedure di clustering risalgono agli inizi del Novecento, in particolare il termine "cluster analysis" fu coniato nel 1939 da Tryon in ambito psicologico (Tryon, 1939).

La *cluster analysis o analisi dei gruppi* racchiude una serie di metodologie statistiche multivariate che ricercano una struttura all'interno dei dati in modo da suddividere le osservazioni in gruppi, chiamati *cluster*. Non conoscendo a priori il numero di gruppi da formare, la cluster analysis è assimilabile ad un'analisi esplorativa.

L'idea elementare alla base del *clustering* è quella di formare dei gruppi le cui unità siano omogenee al loro interno ed eterogenee al loro esterno. Per determinare l'omogeneità e l'eterogeneità delle unità statistiche sono necessarie delle misure di dissimilarità, che consentano di quantificare la distanza tra le unità stesse. Si formeranno quindi dei gruppi in modo tale da minimizzare la distanza delle unità statistiche appartenenti ad un determinato gruppo ma che, allo stesso tempo, massimizzi la distanza tra unità statistiche di gruppi differenti.

Difficilmente svolgendo la *cluster analysis* è possibile esaminare tutti i possibili raggruppamenti. Per ovviare a questo problema sono stati sviluppati diversi algoritmi che, anche se non riescono a determinare qual è il miglior raggruppamento, consentono di effettuare una suddivisione ragionevole.

Esistono diversi algoritmi di *clustering* e la scelta del metodo da utilizzare viene effettuata sulla base degli scopi della ricerca e della tipologia dei dati. Nelle procedure di *cluster analysis* basate sul concetto di distanza è possibile effettuare una suddivisione tra *metodi di partizione* e *metodi gerarchici*. Tra le altre metodologie, si trovano i *metodi basati su modello*, di cui si parlerà nel paragrafo 2.2.

2.1 *k*-means e metodi gerarchici

Nei *metodi di partizione* o *metodi non gerarchici* è necessario definire a priori il numero di gruppi da formare e successivamente allocare le unità statistiche in modo ottimale secondo un criterio prestabilito. Rientra in questa categoria il più noto algoritmo di *clustering*, chiamato algoritmo k-means (MacQueen, 1967). Questo algoritmo consiste nell'allocare ogni unità al gruppo il cui centroide (media) risulta più vicino. In particolare, si compone di tre passi:

1. Le unità vengono suddivise in k cluster iniziali.

2. Una volta calcolato il centroide dei k cluster iniziali si verifica se ogni unità risulta essere più vicina al centroide del cluster a cui appartiene o ad un cluster differente, solitamente utilizzando la distanza euclidea. Nel secondo caso, l'unità viene spostata e vengono ricalcolati i centroidi.

3. Viene ripetuto il passo 2 fino a quando non vi sono più spostamenti di unità.

Si parla di *metodi gerarchici* poiché tramite essi viene formata una sequenza di partizioni nidificate. In base a come queste suddivisioni gerarchiche vengono formate si possono distinguere due differenti approcci:

- *Approccio "botton-up"* o *agglomerativo*: inizialmente vengono formati tanti gruppi quante sono le singole unità statistiche. Queste ultime verranno poi raggruppate in base alla loro vicinanza. L'unione delle unità continua fino a quando non è stato formato un solo gruppo.

- *Approccio "top-down"* o *divisivo*: al contrario dell'*approccio agglomerativo*, inizialmente si avrà un singolo gruppo contenente tutte le unità statistiche. Da questo unico cluster iniziale, si formano gruppi sempre più piccoli escludendo le unità da un cluster in base alla sua distanza da esso. Questa suddivisione procede fino a quanto non sono stati formati tanti gruppi quante sono le unità statistiche.

Entrambi i metodi basano la formazione dei gruppi in base a criteri legati alla distanza tra le osservazioni. Tra i criteri più noti vi sono: il *legame singolo*, il *legame completo* e il *legame medio*.

I gruppi formati con i *metodi gerarchici* possono essere rappresentati mediante il *dendrogramma*. Il dendrogramma in figura 2.1 rappresenta quanto ottenuto dall'applicazione dei metodi gerarchici con distanza di Canberra, una versione pesata della distanza di Manhattan, e il legame completo mediante la funzione *hclust* del pacchetto R *stats* (R Core Team, 2021).

In base alla distanza con cui i gruppi sono formati è possibile vedere dove tagliare il dendrogramma. L'ideale è trovare delle barre orizzontali molto alte in modo tale che vi sia una buona distanza tra un gruppo ed un altro. In questo caso non si evidenzia una grande distanza tra il primo e il secondo gruppo.

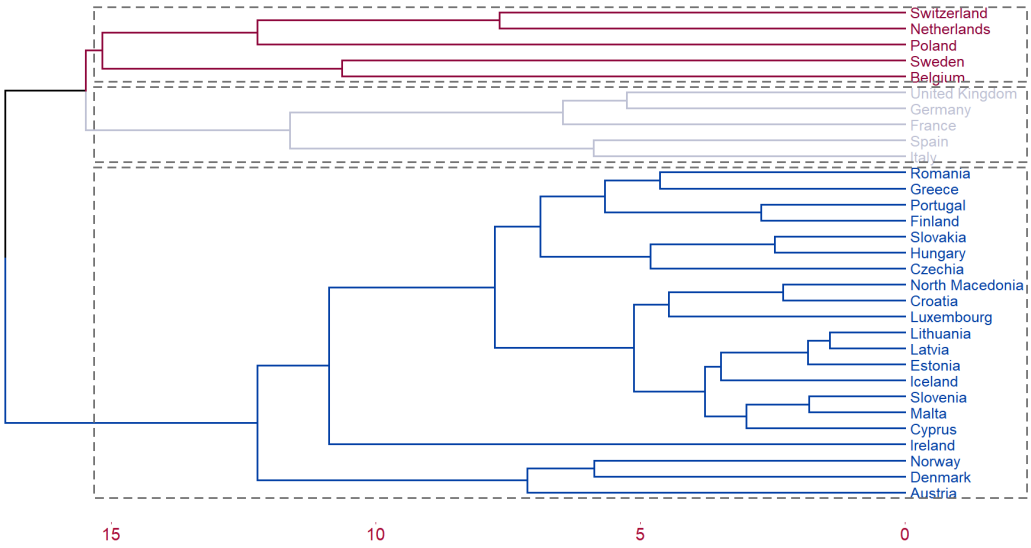


Figura 2.1: Dendrogramma

Dalla figura 2.1 si può vedere che ad una distanza superiore a 15, vengono formati tre gruppi. Si può inoltre notare come il terzo cluster includa molti dei Paesi presenti nel dataset. In questo gruppo, senza alcun dubbio, Austria, Danimarca e Norvegia presentano una situazione economica differente rispetto al resto dei Paesi inclusi, non rendendo del tutto ragionevole il raggruppamento ottenuto.

2.2 Limitazioni delle procedure classiche

Le procedure di *clustering* descritte brevemente sopra presentano alcune limitazioni. In particolare, la suddivisione finale è influenzata notevolmente dalla partizione iniziale adottata. Motivo per il quale questi metodi vengono applicati più volte confrontandone i risultati. Inoltre, i metodi di partizione necessitano di stabilire preventivamente il numero di cluster che devono essere formati.

Per quanto riguarda l'algoritmo *k-means*, nonostante goda di alcuni vantaggi come il suo basso costo computazionale, forma gruppi di forma sferica, rendendo complicata la suddivisione quando i gruppi assumono forme e dimensioni differenti. Cluster di forma sferica implicano che le variabili abbiano un'uguale varianza e che siano tra loro incorrelate. Come si è visto nella *figura 1.1* nel dataset considerato la maggior parte delle variabili sono fortemente correlate, non rendendo quindi l'algoritmo *k-means* il più adatto per la creazione dei gruppi.

Lo svantaggio principale dei *metodi gerarchici*, oltre all'onere a livello computazionale, consiste nel fatto che una volta allocata un'unità questa non possa più essere collocata in un cluster differente. Ne consegue che raggruppamenti errati non potranno più essere corretti. Inoltre, i risultati ottenuti dall'applicazione dei *metodi gerarchici* cambiano a seconda della tipologia di distanza e di legame adattati.

2.3 Model-based clustering

Come si è parlato nel paragrafo precedente, le procedure classiche per effettuare la cluster analysis, quali i metodi gerarchici e i metodi di partizione, presentano delle limitazioni nella formazione dei gruppi.

Per questo motivo nuove procedure più avanzate sono state introdotte. I *modelli a mistura finita* sono stati studiati e proposti per essere utilizzati nel contesto della *cluster analysis*. Il problema della scelta del numero di cluster da formare e del metodo di *clustering* da utilizzare si è quindi convertito in un problema legato alla scelta del modello che spieghi adeguatamente come le osservazioni siano state generate.

Nel *model-based clustering* si assume che i dati provengano da una combinazione di distribuzioni di probabilità sottostante i dati, chiamata *mistura di distribuzioni*. Ogni componente della *mistura* rappresenta un singolo cluster.

Sia $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, dove $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}), i = 1, \ldots, N$, un campione di osservazioni indipendenti ed identicamente distribuite. Supponendo che la popolazione possa essere suddivisa in K cluster, $\mathbf{x}$ è modellato come proveniente dalla *mistura di distribuzioni*

$$f(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}|\boldsymbol{\theta}_k) \tag{2.1}$$

dove p_k sono chiamate *mixing proportions* o *proporzioni della mistura* e rispettano i vincoli

$$\sum_{k=1}^{K} p_k = 1, \tag{2.2}$$

$$0 \leq p_k \leq 1 \quad \forall k = 1, ..., K, \tag{2.3}$$

mentre f_k e $\boldsymbol{\theta}_k$ indicano rispettivamente la funzione di densità della k-esima componente della mistura e i parametri che la caratterizzano.

La funzione di verosimiglianza per il modello di mistura con K componenti è:

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{N} \sum_{k=1}^{K} p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k). \tag{2.4}$$

2.4 Stima dei parametri e algoritmo *EM*

Nell'ambito dei modelli di mistura sono stati proposti due approcci per la stima dei parametri: *mixture approach* e *classification approach*. In questo paragrafo verrà descritto il primo di questi.

Nel *mixture approach* gli elementi del vettore dei parametri $\boldsymbol{\theta}$ vengono stimati in modo da massimizzare la funzione di log-verosimiglianza

$$l(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}) \right) \tag{2.5}$$

solitamente utilizzando l'algoritmo *EM* (Dempster, Laird e Rubin, 1977).

L'algoritmo *EM* (*Expectation-Maximation*) è un algoritmo iterativo che permette di calcolare le stime di massima verosimiglianza in presenza di *dati incompleti*. I dati completi sono indicati con $\mathbf{y}_i = (\mathbf{x}_i, \mathbf{z}_i)$ con $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iK}), i = 1, ..., N$ dove

$$z_{ik} = \begin{cases} 1 \text{ se } \mathbf{x}_i \text{ appartiene al } k-\text{esimo cluster}, \\ 0 \text{ altrimenti.} \end{cases} \tag{2.6}$$

Assunzioni importanti riguardano la densità di $\mathbf{x}_i$ dato $\mathbf{z}_i$, che assume la forma

$$f(\mathbf{x}_i|\mathbf{z}_i) = \prod_{k=1}^{K} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)^{z_{ik}} \tag{2.7}$$

e le $\mathbf{z}_i$, considerate come indipendenti ed identicamente distribuite secondo una distribuzione multinomiale con un'estrazione, K categorie e probabilità $p_1, \ldots, p_K$.

La funzione di log-verosimiglianza del modello di mistura $l(\boldsymbol{\theta}|\mathbf{x})$ è difficile da massimizzare. Per cui viene definita la funzione di log-verosimiglianza per i *dati completi*:

$$l(\boldsymbol{\theta}_k, p_k, z_{ik}|\mathbf{x}) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log(p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)). \tag{2.8}$$

Tuttavia, anche questa quantità è complicata da massimizzare in quanto i valori z_{ik} sono ignoti. Nell'algoritmo viene per questo considerata come quantità da massimizzare il valore atteso della funzione di log-verosimiglianza dei dati completi condizionata ai valori assunti dai parametri alla t-esima iterazione dell'algoritmo, indicato con $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E[l(\boldsymbol{\theta}_k, p_k, z_{ik}|\mathbf{x})|\boldsymbol{\theta}^{(t)}]$. Procedendo in questo modo, in tutti casi in cui era necessario esplicitare le z_{ik} sarà possibile prendere il suo valore atteso

$$\begin{aligned} \widehat{z}_{ik} = E[z_{ik}|\mathbf{x}_i, \boldsymbol{\theta}] &= \sum_{z_{ik}\in\{0,1\}} z_{ik} Pr(z_{ik}|\mathbf{x}_i, \boldsymbol{\theta}) \\ &= Pr(z_{ik} = 1|\mathbf{x}_i, \boldsymbol{\theta}) = Pr(\mathbf{z}_k = k|\mathbf{x}_i, \boldsymbol{\theta}) \end{aligned} \tag{2.9}$$

dove $i = 1, ..., N, k = 1, ..., K$. La quantità $\widehat{z}_{ik}$ viene chiamata *responsability* del k-esimo cluster per l'i-esima osservazione e può essere vista come una probabilità a posteriori che $\mathbf{x}_i$ sia generata dal cluster k.

Assumendo per un momento che $\boldsymbol{\theta}$ non sia ignoto, utilizzando il *teorema di Bayes* si ottiene che

$$\begin{aligned}\widehat{z}_{ik} = Pr(\mathbf{z}_k = k|\mathbf{x}_i, \boldsymbol{\theta}) &= \frac{Pr(\mathbf{x}_i|\mathbf{z}_i = k, \boldsymbol{\theta}) \cdot Pr(\mathbf{z}_i = k)}{\sum_{j=1}^{K} p_j f_j(\mathbf{x}|\boldsymbol{\theta}_j)} \\ &= \frac{f_k(\mathbf{y}_i|\boldsymbol{\theta}) \cdot p_k}{\sum_{j=1}^{K} p_j f_j(\mathbf{x}|\boldsymbol{\theta}_j)}\end{aligned} \tag{2.10}$$

per $i = 1, ..., N, k = 1, ..., K$.

È possibile ora illustrare la procedura utilizzata dall'algoritmo. L'algoritmo *EM* partendo da un valore iniziale per i parametri $\boldsymbol{\theta}^{(0)}$ alterna due passi, l'*expectation step* e il *maximization step*. Alla t-esima iterazione

- Nell'*expectation step* o *E-step* vengono calcolate le quantità

$$\widehat{z}_{ik} = \frac{\widehat{p}_k f_k(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}_k^{(t)})}{\sum_{j=1}^{K} \widehat{p}_j f_j(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}_j^{(t)})}, \tag{2.11}$$

  dove $i = 1, ..., N, k = 1, ..., K$, affinché sia possibile calcolare il valore atteso della funzione di log-verosimiglianza dei dati completi condizionata ai valori assunti dai parametri alla t-esima iterazione $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

- Nel *maximization step* o *M-step* viene effettuato un aggiornamento dei parametri $\boldsymbol{\theta}^{(t+1)}$ utilizzando le $\widehat{z}_{ik}$ ottenute nell'*E-step* in modo tale da massimizzare $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{2.12}$$

  per cui

$$Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \forall \boldsymbol{\theta} \tag{2.13}$$

  I passi si alternano fino a quando

$$l(\boldsymbol{\theta}^{(t+1)})) - l(\boldsymbol{\theta}^{(t)})) < \varepsilon \tag{2.14}$$

per un ε arbitrario positivo e sufficientemente piccolo. Viene raggiunta in questo modo la convergenza.

Con questo approccio la suddivisione delle unità deriva direttamente dalle stime di massima verosimiglianza dei parametri del modello di mistura. Indicato con z_{ik}^* il valore $\widehat{z}_{ik}$ che massimizza la funzione di log-verosimiglianza $l(\boldsymbol{\theta}|\mathbf{x})$ del modello di mistura con K componenti, l'i-esima osservazione $\mathbf{x}_i$ verrà collocata nel cluster con la maggiore probabilità stimata, ossia

$$\left\{ j | z_{ij}^* = \max_k z_{ik}^* \right\}. \tag{2.15}$$

In questo modo $1 - \max_k z_{ik}^*$ rappresenta una misura di incertezza nella classificazione.

L'algoritmo *EM* quando viene utilizzato nell'ambito della cluster analysis presenta tuttavia alcune limitazioni. Innanzitutto, potrebbe convergere molto lentamente. Ad ogni modo, l'algoritmo lavora in modo adeguato qualora i dati si adattassero bene al modello e i valori di partenza per i parametri risultassero ragionevoli. Oltre a ciò, in relazione ad un *modello di mistura Gaussiana*, l'algoritmo *EM* non è adatto quando la matrice di varianze e covarianze associata ad una o più componenti è singolare o quasi singolare[1]. In questo caso l'algoritmo potrebbe arrestarsi oppure fornire risultati inaccurati nella situazione in cui ci siano troppe componenti della mistura.

Sono state proposte diverse versioni dell'algoritmo *EM*, tra questi troviamo l'algoritmo *SEM* (*stochastic EM*) in cui le $\widehat{z}_{ik}$ non vengono stimate ma simulate e l'algoritmo *CEM* (*classification EM*) in cui le $\widehat{z}_{ik}$ calcolate nell'*E-step* vengono convertite in una classificazione discreta prima di procedere con il passo di massimizzazione.

2.5 Modelli di mistura Gaussiani

Nella maggior parte delle applicazioni si assume che tutte le componenti della mistura provengano dalla stessa famiglia di distribuzioni. Il modello di mistura più noto è quello che prevede una mistura di distribuzioni normali multivariate, chiamato *Gaussian Mixture Model* (*GMM*) o *modello di mistura Gaussiana*. In un modello di mistura Gaussiana con K componenti, la k-esima componente della mistura di distribuzioni $f_k(\mathbf{x})$ consiste in una distribuzione normale p-variata con vettore media $\boldsymbol{\mu}_k$ e matrice di varianze e covarianze $\boldsymbol{\Sigma}_k, k = 1, ..., K$.

[1]Una matrice singolare è una matrice quadrata il cui determinante è pari a 0. Mentre una matrice quasi singolare è una matrice quadrata con determinante che assume un valore prossimo a 0.

La mistura di distribuzioni normali con K componenti è:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_k|} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}.$$

I cluster formati da questo modello assumono una forma ellissoidale, sono centrati in $\boldsymbol{\mu}_k$ e hanno una densità maggiore nei punti vicino alla media. Forma, dimensione[2] e orientamento dei cluster sono determinati dalla matrice di varianze e covarianze $\boldsymbol{\Sigma}_k$. In Benfield e Raftery (1993) venne sviluppato un criterio generale che consente di controllare le caratteristiche geometriche dei cluster, quali forma, dimensione e orientamento. Questo criterio consiste nella riparametrizzazione della matrice di varianze e covarianze tramite la decomposizione a valori singolari. La struttura per $\boldsymbol{\Sigma}_k$ è la seguente:

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{'} \tag{2.16}$$

dove $\mathbf{D}_k$ è la matrice ortogonale degli autovettori di $\boldsymbol{\Sigma}_k$, $\lambda_k = |\boldsymbol{\Sigma}_k^{\frac{1}{p}}|$ e $\mathbf{A}_k$ consiste in una matrice diagonale tale che $|\mathbf{A}_k| = 1$ i cui elementi sulla diagonale corrispondono agli autovalori normalizzati di $\boldsymbol{\Sigma}_k$ disposti in ordine decrescente. Quindi $\mathbf{D}_k$ regola l'orientamento della k-esima componente della mistura, o analogamente del k-esimo cluster, λ_k governa la sua dimensione mentre $\mathbf{A}_k$ controlla la sua forma. Queste caratteristiche geometriche della distribuzione sono stimate dai dati e possono variare tra cluster oppure rimanere le stesse per tutti i gruppi.

La decomposizione a valori singolari della matrice di varianze e covarianze $\boldsymbol{\Sigma}_k$ riesce a modellare molteplici situazioni. In Celeux e Govaert (1995), sulla base del lavoro di Benfield e Raftery, vennero proposti 14 modelli di mistura Gaussiani che differiscono tra loro per come la matrice di varianze e covarianze $\boldsymbol{\Sigma}_k$ è definita. I nomi e le caratteristiche geometriche assunte dai modelli sono riassunti nella tabella 2.1.

[2]Per dimensione si fa riferimento al volume occupato dal cluster nello spazio p-dimensionale.

Modello	$\boldsymbol{\Sigma}_k$	**Orientamento**	**Volume**	**Forma**
EII	$\lambda\mathbf{I}$	Non definito	Uguale	Sferica
VII	$\lambda_k\mathbf{I}$	Non definito	Variabile	Sferica
EEI	$\lambda\mathbf{A}$	Allineato con gli assi	Uguale	Uguale
VEI	$\lambda_k\mathbf{A}$	Allineato con gli assi	Variabile	Uguale
EVI	$\lambda\mathbf{A}_k$	Allineato con gli assi	Uguale	Variabile
VVI	$\lambda_k\mathbf{A}_k$	Allineato con gli assi	Variabile	Variabile
EEE	$\lambda\mathbf{DAD}'$	Uguale	Uguale	Uguale
VEE	$\lambda_k\mathbf{DAD}'$	Uguale	Variabile	Uguale
EVE	$\lambda\mathbf{DA}_k\mathbf{D}'$	Uguale	Uguale	Variabile
EEV	$\lambda\mathbf{D}_k\mathbf{AD}_k'$	Variabile	Uguale	Uguale
VVE	$\lambda_k\mathbf{DA}_k\mathbf{D}'$	Uguale	Variabile	Variabile
VEV	$\lambda_k\mathbf{D}_k\mathbf{AD}_k'$	Variabile	Variabile	Uguale
EVV	$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k'$	Variabile	Uguale	Variabile
VVV	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k'$	Variabile	Variabile	Variabile

Tabella 2.1: Modelli di mistura Gaussiani

2.6 Selezione del modello

Con riferimento alla *cluster analysis*, dopo aver scelto il metodo da utilizzare per effettuare il raggruppamento, il problema più rilevante resta quello di determinare il numero di cluster in cui suddividere le unità statistiche. Tuttavia, quando viene applicato come metodologia il *model-based clustering* questo problema si converte nella scelta del modello più adeguato.

Un approccio per la selezione del modello è basato sul *fattore di Bayes* e le probabilità a posteriori del modello. L'idea alla base è quella che considerando $M_1, ..., M_K$ modelli con probabilità a priori $Pr(M_k), k = 1, ..., K$, che spesso vengono assunte uguali, allora per il teorema di Bayes le probabilità a posteriori per il modello M_k condizionata ai dati D è proporzionale alla probabilità dei dati condizionata al modello M_k moltiplicata per la probabilità a priori del modello, ossia

$$Pr(M_k|D) \propto Pr(D|M_k) \cdot Pr(M_k), k = 1, ..., K. \tag{2.17}$$

Quando i parametri sono ignoti, $Pr(D|M_k)$, chiamata *funzione di verosimiglianza integrata* del modello M_k, si ottiene nel seguente modo:

$$Pr(D|M_k) = \int Pr(D|\boldsymbol{\theta}_k, M_k) \cdot Pr(\boldsymbol{\theta}_k|M_k)\mathrm{d}\boldsymbol{\theta}_k, \tag{2.18}$$

in cui $Pr(\boldsymbol{\theta}_k|M_k)$ è la distribuzioni a priori di $\boldsymbol{\theta}_k$.

Si andrà quindi a scegliere il modello che risulta essere più verosimile a posteriori. Se le probabilità a priori del modello $Pr(M_k)$ sono uguali, questo approccio equivale a scegliere il modello la cui verosimiglianza integrata assume il valore maggiore.

Confrontando due modelli, M_k e M_j con $k, j = 1, ..., K, k \neq j$, il *fattore di Bayes* è definito dal rapporto tra le due verosimiglianze integrate ad essi associate

$$BF(D) = \frac{Pr(D|M_k)}{Pr(D|M_j)}. \tag{2.19}$$

In base al valore assunto da $BF(D)$ vi sarà una differente evidenza a favore di un modello rispetto all'altro. In particolare, se $BF(D) > 1$ sarà da preferire il modello M_k e si avrà una forte evidenza a favore di questo modello qualora $BF(D) > 100$.

Data la difficoltà nel calcolo della funzione di verosimiglianza integrata, spesso viene utilizzata una sua approssimazione, data dal *Bayesian Information Criterion* o *BIC* (Schwarz, 1978). L'approssimazione, ottenuta con il metodo di Laplace, è la seguente:

$$\log Pr(D|M_k) = \log Pr(D|\widehat{\boldsymbol{\theta}}_k, M_k) - \frac{\upsilon_k}{2}\log(N) + o(1), \tag{2.20}$$

dove $\widehat{\boldsymbol{\theta}}_k$ corrisponde alla stima di massima verosimiglianza di $\boldsymbol{\theta}_k$, dunque $Pr(D|\widehat{\boldsymbol{\theta}}_k, M_k)$ è la funzione di verosimiglianza massimizzata del modello M_k, υ_k è il numero di parametri liberi da stimare nel modello M_k e N è la dimensione del campione. Prendendo come funzione di perdita $-2\log Pr(D|M_k)$ si ottiene che questa corrisponde al *BIC*

$$BIC_k = -2\log Pr(D|M_k) = -2\log Pr(D|\widehat{\boldsymbol{\theta}}_k, M_k) + \upsilon_k log(Nn).$$

Quindi scegliere il modello con il minore valore del *BIC* equivale a selezionare il modello con la maggiore probabilità (approssimata) a posteriori.

Nonostante i modelli di mistura non soddisfino le condizioni di regolarità sottostanti la dimostrazione dell'approssimazione sopra riportata, molti studi hanno mostrato che utilizzare il *BIC* per la scelta del modello più adeguato sia ragionevole. In particolare, come criterio di selezione il *BIC* è asintoticamente consistente, dunque data una famiglia di modelli che include il vero modello, la probabilità che il *BIC* selezioni il modello corretto tende a uno quando $N \to \infty$, dove N è la dimensione del campione.

Capitolo 3

Raggruppamento dei Paesi europei

Esistono diversi pacchetti R (R Core Team, 2021) da poter utilizzare per applicare la *cluster analysis*. In questo capitolo verranno presentati ed analizzati due pacchetti che possono essere utilizzati per applicare il *model-based clustering* ai dati completi del dataset presentato nel capitolo 1: *mclust* (Scrucca e altri, 2016) e *MixAll* (Iovleff, 2019). Successivamente verranno illustrati i risultati ottenuti.

3.1 Librerie R

3.1.1 *mclust*

Il pacchetto R *mclust* contiene una serie di funzioni volte all'applicazione del *model-based clustering* nell'ambito dei modelli di mistura Gaussiani.

Comprende funzioni per la stima dei parametri tramite l'utilizzo dell'algoritmo *EM* e una moltitudine di strutture per la matrice di varianze e covarianze. In particolare, il pacchetto *mclust* implementa i seguenti modelli di mistura Gaussiani: *EII*, *VII*, *EEI*, *VEI*, *EVI*, *VVI*, *EEE*, *EEV*, *VEV* e *VVV*. Per maggiori dettagli riguardo alla parametrizzazione della matrice di varianze e covarianze si rimanda alla tabella 2.1.

Il numero ottimale di cluster in cui suddividere le unità statistiche e la parametrizzazione per la matrice di varianze e covarianze vengono determinati utilizzando come criterio per la selezione del modello il *BIC*. Il *BIC* consente quindi il confronto tra modelli con un differente numero di parametri e/o una diversa struttura per la matrice di varianze e covarianze.

In *mclust*, a differenza di quanto precisato nel capitolo 2, viene selezionato il modello che presenta il valore maggiore del *BIC*. Questo perché *mclust* definisce il *BIC* nel seguente modo:

$$BIC \equiv 2 \log Pr(D|\widehat{\boldsymbol{\theta}}_k, M_k) - \upsilon_k log(N), \tag{3.1}$$

dove $\log Pr(D|\widehat{\boldsymbol{\theta}}_k, M_k)$ è il valore della log-verosimiglianza massimizzata dal modello M_k, υ_k è il numero di parametri liberi del modello e N è la dimensione del campione.

mlcust non provvede direttamente a trattare i valori mancanti presenti nel dataset. Sarà quindi necessario, prima di applicare il *model-based clustering*, effettuare un'imputazione. Per completare i dataset con valori mancanti, in *mclust* è stata aggiunta la funzione *imputeData* del pacchetto R *mix* (Schafer, 2022). I valori imputati mediante *imputeData* variano a seconda del *seed* specificato all'interno della funzione.

3.1.2 *MixAll*

Il pacchetto R *MixAll* rientra nella libreria *SKT++* (*The **S**tatistical **T**ool**K**it*) che consiste in una collezione di classi C++ per le analisi in ambito statistico. Grazie a questo pacchetto è possibile implementare una serie di modelli di mistura che consente di applicare il *model-based clustering* in un contesto molto ampio.

In particolare, vengono implementati:

- *Modelli di mistura Gaussiani diagonali* con il comando *clusterDiagGaussian*;
- *Modelli di mistura gamma* con il comando *clusterGamma*;
- *Modelli di mistura categoriali* con il comando *clusterCategorical*;
- *Modelli di mistura Poisson* con il comando *clusterPoisson*;
- Un modello denominato *Mixed Data* con il comando *clusterMixedData*.

Utilizzando *MixAll* è possibile applicare la *cluster analysis* basata su modello anche nel caso di dataset con valori mancanti. Qualora presenti, questi ultimi verranno sostituiti effettuando un'imputazione durante il processo di stima.

Com'è stato descritto nel capitolo precedente, nell'ambito del *model-based clustering* ai dati osservati viene associato un vettore di etichette $\{\mathbf{z}_1, ..., \mathbf{z}_n\}, \mathbf{z}_i = (z_{i1}, ..., z_{iK})$ con $z_{ik} \in \{0, 1\}$ a seconda che $\mathbf{x}_i$ provenga dalla k-esima componente della mistura. Tuttavia, trattandosi di variabili latenti il processo di stima viene effettuato con l'ausilio di un algoritmo. *MixAll* provvede a realizzare questo processo tramite l'algoritmo *EM*, presentato nel capitolo 2, o alternativamente tramite gli algoritmi *SEM*, *CEM* o *SemiSEM*.

Tutti gli algoritmi di stima sopra elencati necessitano di un valore iniziale per il vettore dei parametri $\boldsymbol{\theta}$. Per minimizzare la possibilità di ottenere un valore iniziale per $\boldsymbol{\theta}$ sfortunato, vengono prese in considerazione più inizializzazioni e viene conservata solamente quella che, dal punto di vista della verosimiglianza, risulta essere la migliore.

È possibile poi specificare la *strategia* di stima, tramite il comando *clusterStrategy*. Per strategia si intende un modo per ottenere una buona stima dei parametri del modello di mistura evitando punti di massimo locale della funzione di verosimiglianza. La strategia implementata in *MixAll* è composta da tre step:

- *Ricerca*: costruzione di un metodo di ricerca che consenta di generare un certo numero di posizioni iniziali;
- *Esecuzione*: esecuzione di un algoritmo *breve*, ossia con poche iterazioni, per ogni posizione iniziale;
- *Selezione*: selezionare la soluzione che presenta il valore migliore della funzione di verosimiglianza ed eseguire un algoritmo *lungo*, ovvero con un maggior numero di iterazioni rispetto all'algoritmo *breve*, per questa soluzione.

3.2 Applicazione al dataset

In questo paragrafo verranno riportati i risultati ottenuti dal *model-based clustering* applicato al dataset economico presentato nel capitolo 1 con entrambi i pacchetti R precedentemente descritti.

Le variabili non sono espresse nella stessa unità di misura, motivo per il quale è stato necessario standardizzare i dati. Inoltre, come emerso anche dalla figura 1.1, alcune variabili risultano perfettamente correlate. In questo contesto si possono adottare due soluzioni:

- eliminare abbastanza variabili per eliminare la collinearità;
- applicare un metodo di riduzione della dimensionalità.

Per questo motivo il *model-based clustering* è stato implementato con due diverse modalità: eliminando la variabile *GDP* ed applicando preventivamente l'analisi delle componenti principali tramite il comando *prcomp* del pacchetto R *stats* (R Core Team, 2021). È da sottolineare che utilizzare l'analisi delle componenti principali consente di ridurre la dimensione del dataset con la minima perdita di informazione possibile. La PCA consente di creare nuove variabili chiamate *componenti pricipali* tra loro incorrelate, evitando problemi di collinearità, a scapito di una più complicata interpretazione dei risultati.

3.2.1 *mclust*

Dopo aver imputato i dati tramite il comando *imputeData*, che applica come metodologia l'imputazione multipla, è possibile utilizzare il pacchetto *mclust*. Come già evidenziato, il problema della decisione del numero ottimale di gruppi nell'ambito del *model-based clustering* si traduce in un problema di selezione del modello più adeguato.

Il modello che viene scelto utilizzando come criterio il *BIC* è un modello di mistura Gaussiana *VEV* con $k = 3$ componenti. Come descritto nel capitolo 2, in base alla struttura adottata per la matrice di varianze e covarianze $\mathbf{\Sigma}_k$ cambiano orientamento, dimensione e forma dei cluster formati. I cluster hanno una forma ellittica, con uguale orientamento ma volume differente.

Di seguito vengono riportati nella tabella 3.1 i valori del *BIC* assunti dai tre migliori modelli di mistura Gaussiani e nella figura 3.1 un grafico contenente il valore del *BIC* associato ai modelli considerati in *mclust* al variare del numero di cluster.

Modello	*VEV*	*VVV*	*EEE*
Numero di cluster	3	2	9
BIC	1638.364	1471.821	1443.390

Tabella 3.1: 3 migliori modelli in base al *BIC*

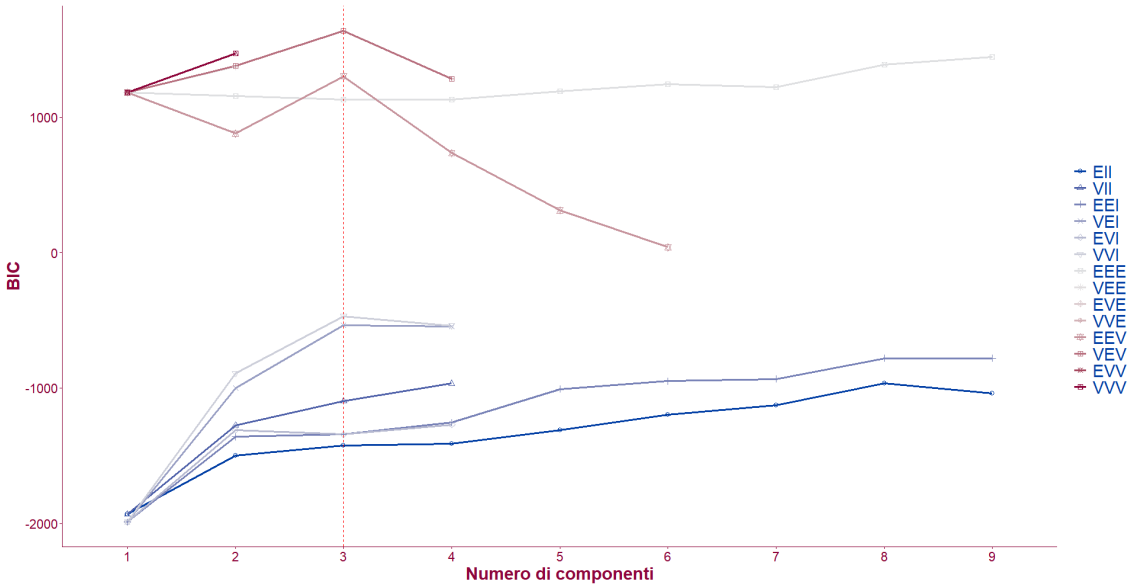


Figura 3.1: Selezione del modello in *mclust*

Dalla figura 3.1 si nota che, per alcuni modelli, al variare del numero di componenti, il valore del *BIC* non è sempre disponibile. Questo si verifica nel caso in cui si riscontri una matrice di varianze e covarianze singolare. Come anticipato nel capitolo 2, in presenza di matrici di varianze e covarianze singolari o quasi singolari, l'algoritmo *EM* può arrestarsi non permettendo la stima del modello di mistura e conseguentemente il calcolo del valore del *BIC* associato a tale modello.

Una volta stimato il modello, per visualizzare la suddivisione effettuata è possibile utilizzare il comando *mclustDR* che consente di stimare le basi di una superficie ridotta in grado di catturare la maggior parte della struttura contenuta all'interno dei dati. Di seguito vengono riportati due grafici: la figura 3.2 è relativa alla suddivisione dei Paesi effettuata con il modello di mistura Gaussiana implementato mentre la figura 3.3 mostra la densità della distribuzione di mistura Gaussiana con $k = 3$ componenti.

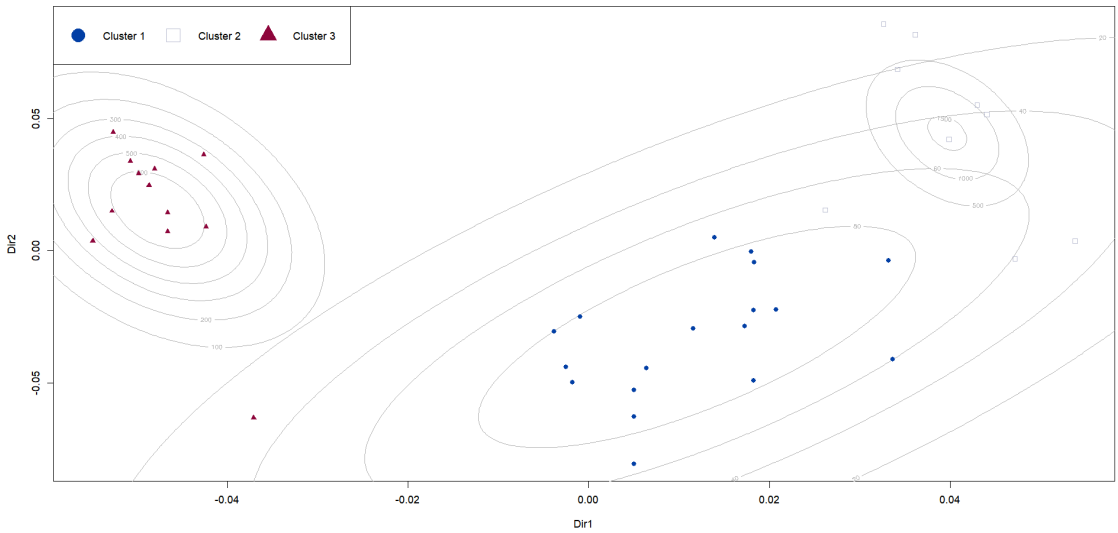


Figura 3.2: Cluster ottenuti con *mclust*

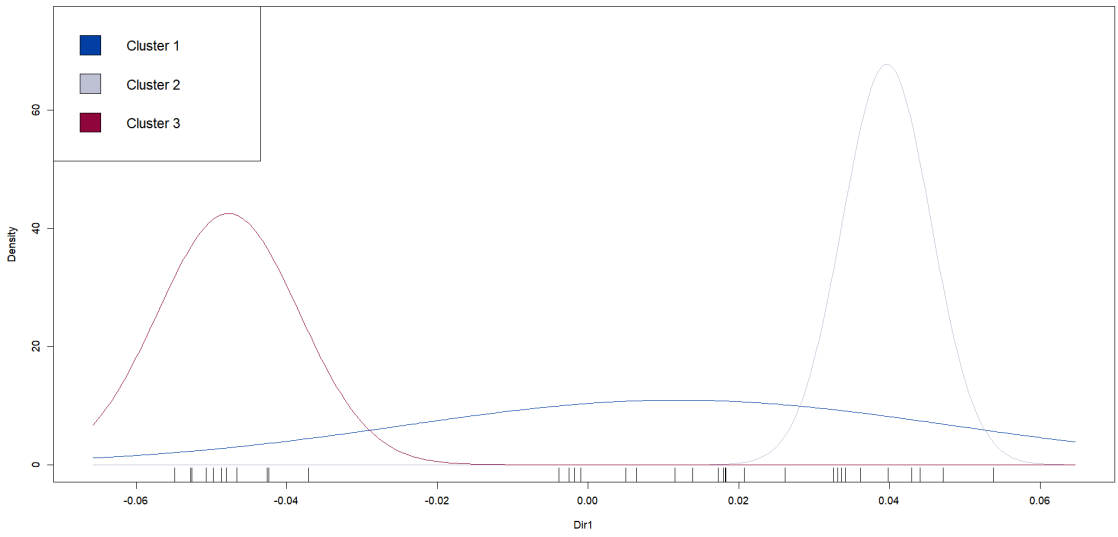


Figura 3.3: Densità della mistura Gaussiana VEV con $k = 3$

Dalla figura 3.2 si nota una sovrapposizione tra i cluster 1 e 2, mentre il cluster 3 risulta essere più separato dagli altri. Dal grafico in figura 3.3 è possibile evidenziare alcune caratteristiche delle singole componenti della mistura. In particolare, la prima componente, corrispondente al cluster 1, ha un campo di variabilità molto ampio rispetto alle altre due. Questo aspetto, assieme al fatto che le medie della prima e della seconda componente siano più vicine rispetto a quanto lo siano quelle della prima e della terza, si riflette nella sovrapposizione tra i cluster 1 e 2 evidenziata per l'interpretazione della figura 3.2. Questa sovrapposizione influisce sicuramente sul grado di incertezza che si ha nel classificare un Paese in un gruppo piuttosto che in un altro. Ritornando a quanto descritto a livello teorico nel paragrafo 2.3 per alcune unità la misura di incertezza della classificazione non è ridotta.

Viene riportata nella figura 3.4 la suddivisione geografica ottenuta applicando il pacchetto *mclust* al dataset meno la variabile *GDP*.

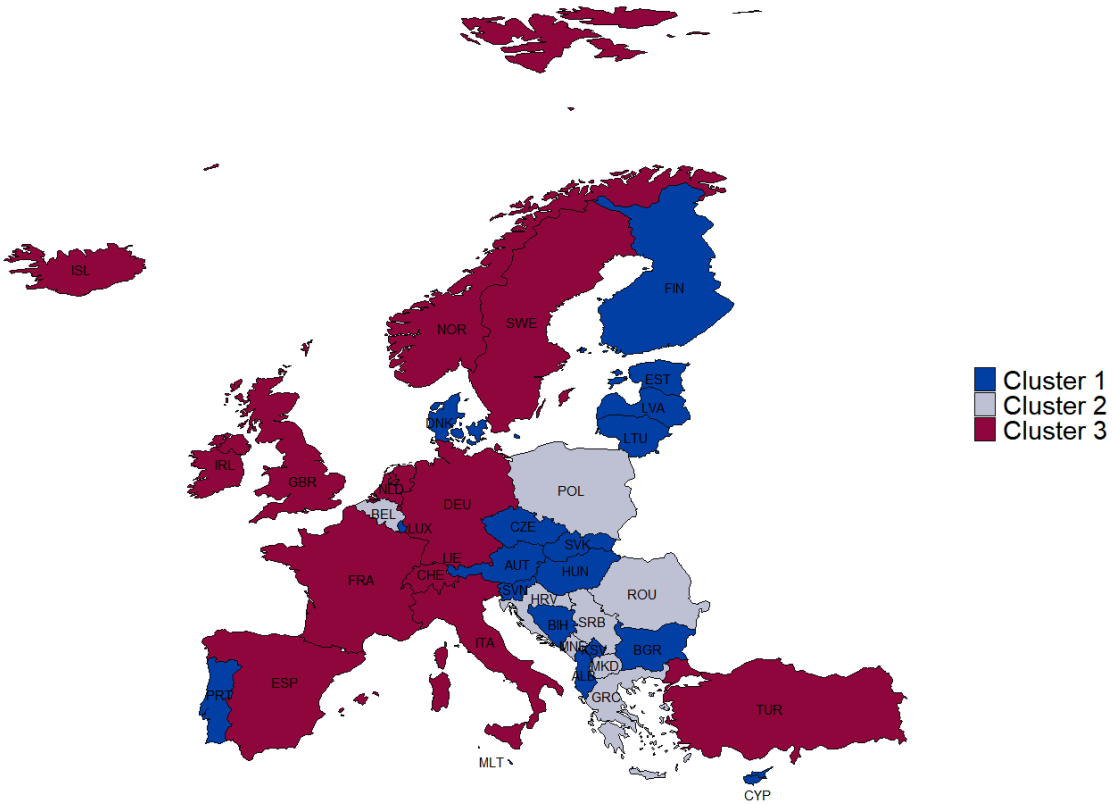


Figura 3.4: Suddivisione geografica con *mclust*

Vediamo ora cosa si ottiene applicando le stesse funzioni del pacchetto *mclust* ma successivamente all'applicazione dell'analisi delle componenti principali.

Le prime due componenti principali spiegano all'incirca l'85% della variabilità totale. Lo *screeplot*, che riporta la varianza spiegata cumulata all'aumentare del numero delle componenti principali, viene riportato nella figura 3.5. Nel grafico si nota la presenza del gomito in corrispondenza della seconda componente principale.

La prima componente riassume il PIL, tutte le sue componenti e, in misura minore, il bilancio dei pagamenti mentre la seconda rappresenta i tassi del mercato del lavoro e, con un peso inferiore, l'indice dei prezzi al consumo armonizzato. Si nota che *HICP* e *UNEMPL* entrano con il segno negativo nella seconda componente principale.

	PC1	**PC2**
GDP	0.280	-0.027
GVA	0.280	-0.026
FCE_GG	0.280	-0.015
FCE_NPISH_H	0.276	-0.047
AIC	0.277	-0.037
GFCF	0.277	-0.019
CI_A.DV	0.236	0.069
E_GS	0.269	0.049
I_GS	0.272	0.037
CE	0.280	0.008
GOS_GMI	0.275	-0.066
T.S_PI	0.273	-0.050
T.S_P	0.278	-0.040
BOP	0.163	0.094
HICP	-0.015	-0.226
LFA	0.010	0.566
EMPL	0.024	0.594
UNEMPL	-0.021	-0.494

Tabella 3.2: Matrice di rotazione: coefficienti delle componenti principali

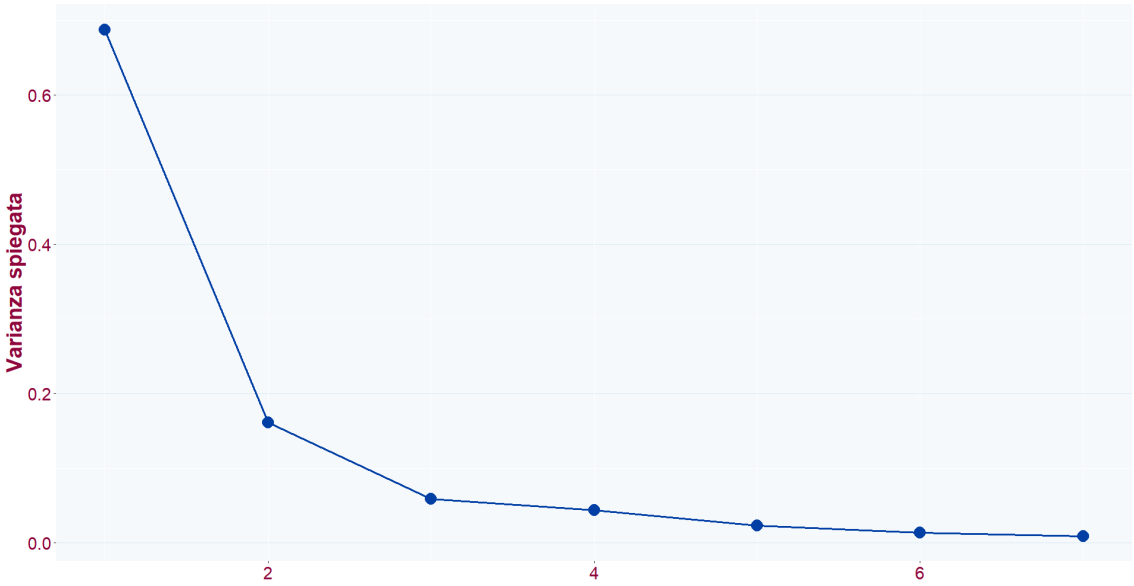


Figura 3.5: Varianza spiegata dalle componenti principali

Utilizzando le prime due componenti principali vengono formati, come nel caso precedente, tre gruppi. Tuttavia, il modello selezionato dal *BIC* (figura 3.6) è in questo caso il modello di mistura Gaussiana *VVI*, ossia i cluster formati sono caratterizzati da forma e volume differenti e orientamento allineato con gli assi.

Il modello selezionato in questo caso è più semplice rispetto al modello *VEV* selezionato nel caso precedente in cui è stata eliminata la variabile *GDP* prima di effettuare l'analisi dei gruppi. Per più semplice si intende che richiede la stima di un numero minore di parametri. L'applicazione dell'analisi delle componenti principali prima di utilizzare il pacchetto *mclust* consente di diminuire ulteriormente il numero di parametri da stimare e quindi di ridurre il costo computazionale richiesto per la stima e il confronto tra i modelli di mistura.

Dalla figura 3.7 si evidenzia una minore sovrapposizione delle componenti della mistura rispetto a quanto emerso nel grafico 3.2. Tuttavia, anche in questo frangente si nota la presenza di alcune unità sulle code delle Gaussiane che compongono la mistura. Queste unità presentano quindi una minore probabilità di appartenere al gruppo in cui sono state collocate rispetto ad unità localizzate vicino alla media delle corrispettive componenti.

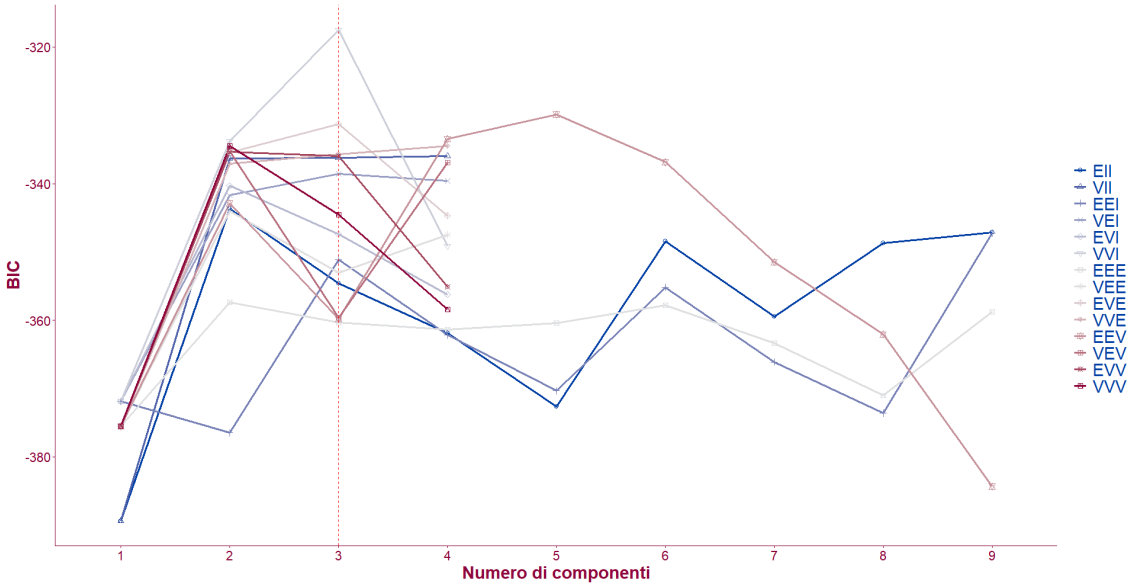


Figura 3.6: Selezione del modello con *mclust* dopo la PCA

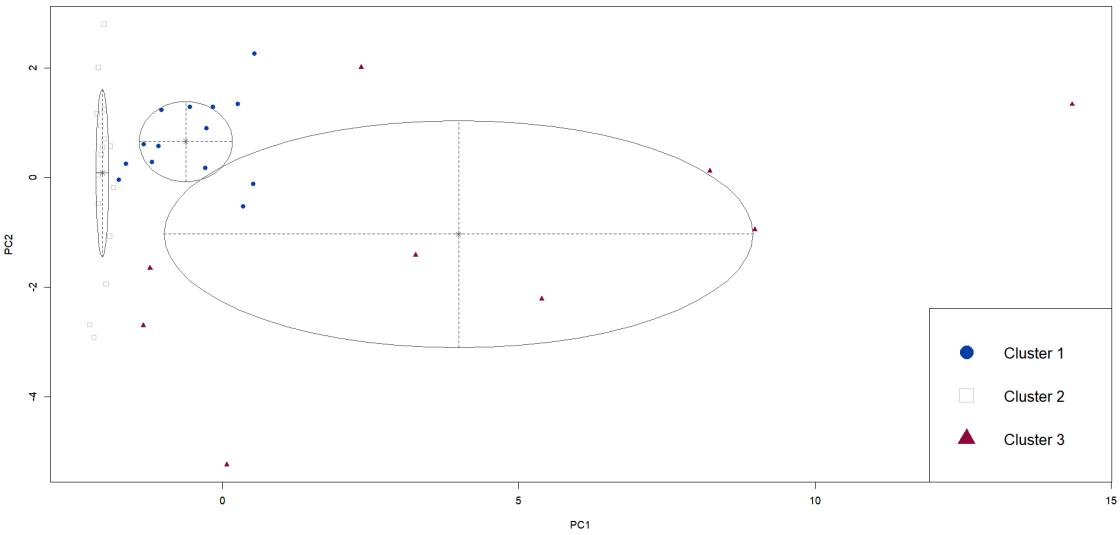


Figura 3.7: Cluster ottenuti con *mclust* dopo la PCA

Viene riportata nella figura 3.8 la suddivisione geografica ottenuta applicando *mclust* dopo aver effettuato l'analisi delle componenti principali al dataset completo.

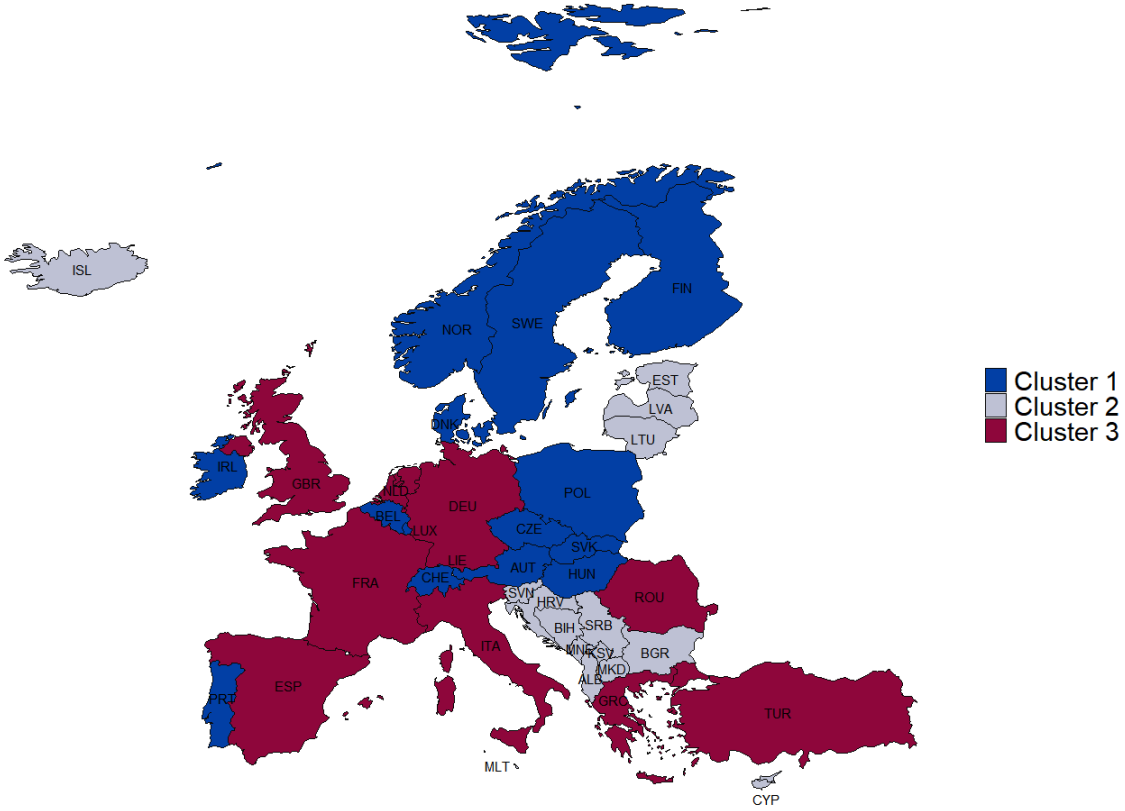


Figura 3.8: Suddivisione geografica con *mclust* dopo la PCA

3.2.2 *MixAll*

Nel caso in esame è stato applicato il comando *clusterDiagGaussian* che permette di implementare i modelli di mistura Gaussiani diagonali. Rispetto ad *mclust* vi è quindi un'assunzione aggiuntiva: utilizzando la funzione *clusterDiagGaussian* si assume che la matrice di varianze e covarianze $\mathbf{\Sigma}_k$ per ogni componente k sia una matrice diagonale.

Assunzioni sulle proporzioni della mistura e sulle deviazioni standard forniscono 8 modelli che vengono confrontati in fase di implementazione. In particolare:

- le proporzioni possono essere uguali o variare;
- le deviazioni standard possono essere uguali o differenti per tutte le variabili;
- le deviazioni standard possono essere uguali o differenti per tutti i cluster.

I modelli sono riassunti nella tabella 3.3.

Modello	**Proporzioni**	**s.d. nelle variabili**	**s.d. nei cluster**
gaussian_p_sjk	Uguali	Libere	Libere
gaussian_p_sj	Uguali	Libere	Uguali
gaussian_p_sk	Uguali	Uguali	Libere
gaussian_p_s	Uguali	Uguali	Uguali
gaussian_pk_sjk	Libere	Libere	Libere
gaussian_pk_sj	Libere	Libere	Uguali
gaussian_pk_sk	Libere	Uguali	Libere
gaussian_pk_s	Libere	Uguali	Uguali

Tabella 3.3: Modelli di mistura Gaussiani implementati con *MixAll*: assunzioni sulle proporzioni e sulle deviazioni standard nelle variabili e nei cluster

Anche in questo caso applichiamo i comandi di *MixAll* prima al dataset completo a meno della variabile *GDP* e successivamente dopo aver effettuato l'analisi delle componenti principali.

Il miglior modello in entrambi i casi è stato selezionato sulla base del valore assunto dal criterio *BIC*.

Il modello selezionato utilizzando il comando *clusterDiagGaussian* al dataset senza la variabile relativa al PIL è il modello *gaussian_pk_sj* che presenta differenti proporzioni e deviazioni standard tra variabili ma uguali deviazioni standard tra componenti della mistura. Il *BIC* inoltre seleziona come miglior modello quello con $k = 5$ componenti.

A differenza di quanto è stato possibile effettuare con *mclust*, *MixAll* non provvede ad una rappresentazione grafica su una superficie ridotta ma consente solamente di visualizzare la suddivisione delle unità per ogni variabile del dataset.

Considerando la presenza di 17 variabili nel dataset per permettere la visualizzazione della suddivisione effettuata, dopo aver stimato il modello, e di conseguenza imputato i valori mancanti, è stata applicata l'analisi delle componenti principali e sono state utilizzate le prime due componenti principali per rappresentare i cluster formati. Per l'interpretazione delle componenti principali al dataset si rimanda alla tabella 3.2.

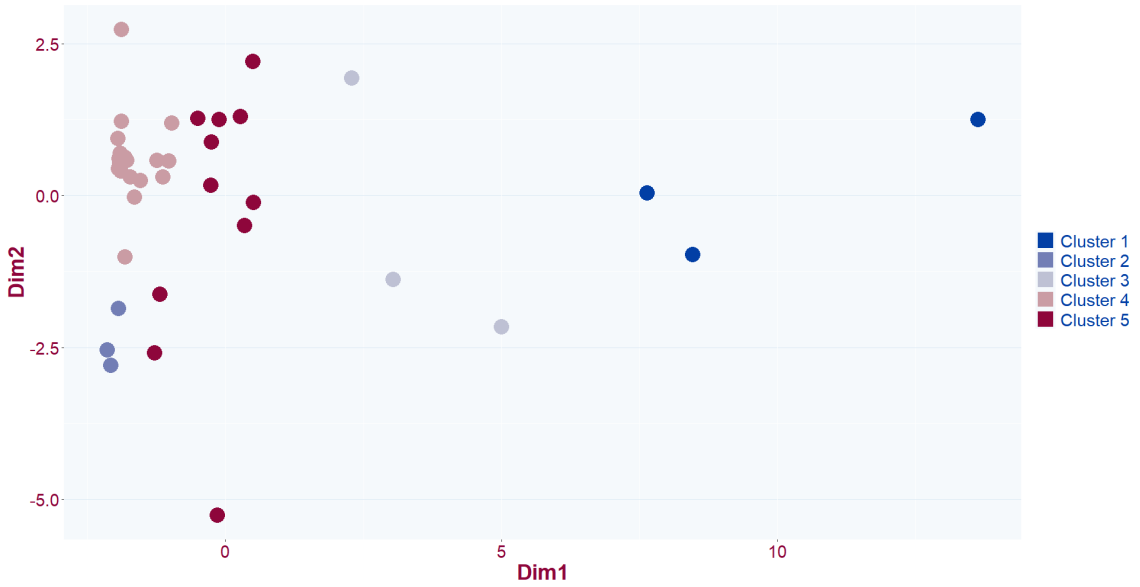


Figura 3.9: Cluster ottenuti con *MixAll*

A livello geografico si ottiene la suddivisione rappresentata nella figura 3.10.

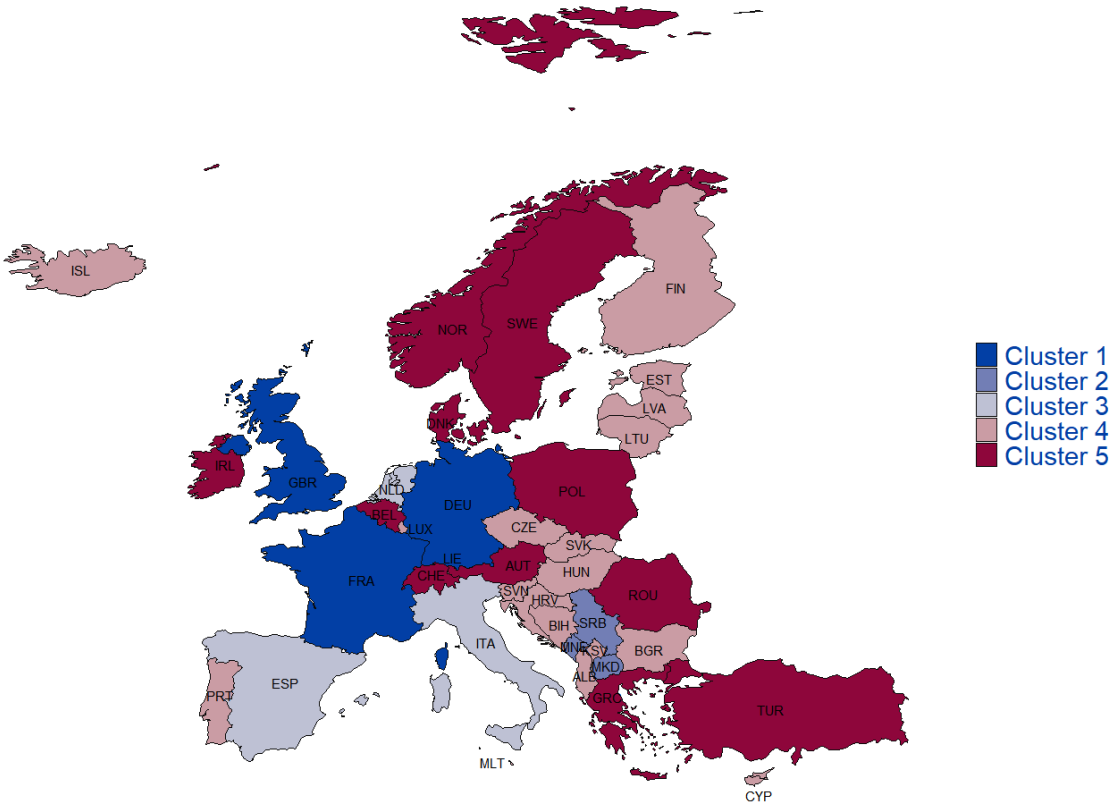


Figura 3.10: Suddivisione geografica con *MixAll*

Applicando invece le componenti principali prima dell'implementazione del modello di mistura, il modello selezionato dal *BIC* continua ad avere $k = 5$ componenti ma in questo caso con differenti proporzioni, uguali deviazioni standard tra variabili ma diverse tra componenti della mistura.

La suddivisione ottenuta è riportata nelle figure 3.11 e 3.12.

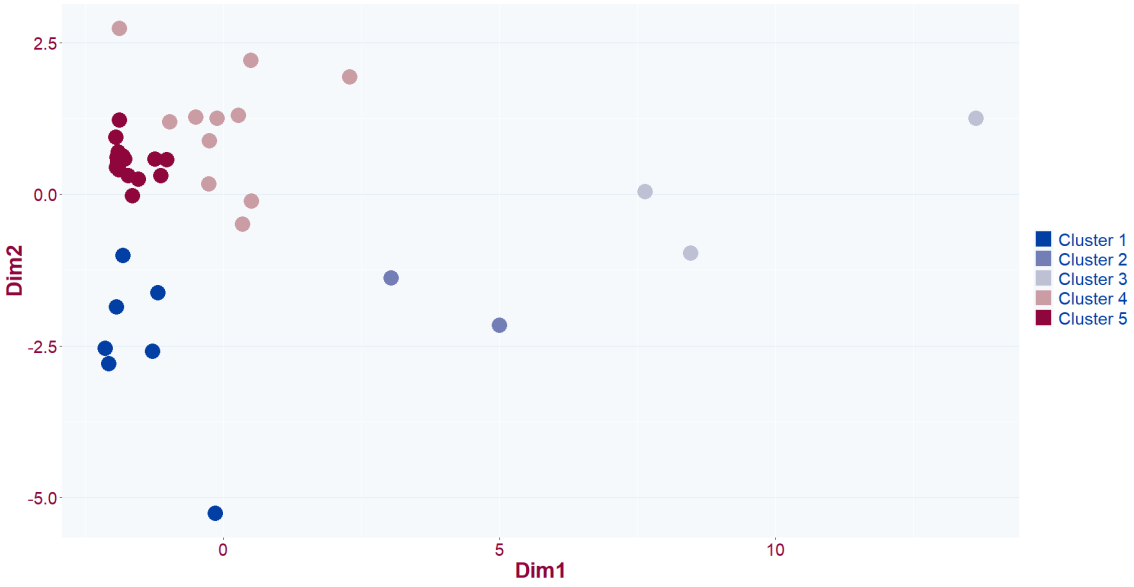


Figura 3.11: Cluster ottenuti con *MixAll* dopo la PCA

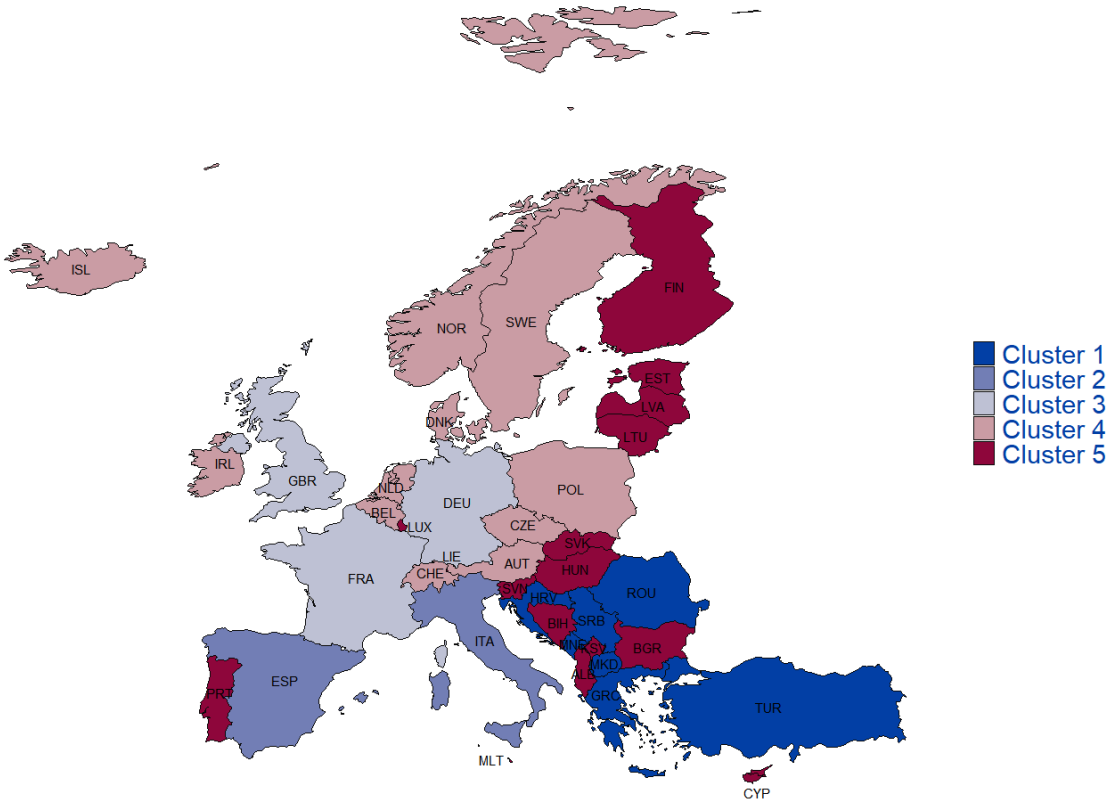


Figura 3.12: Suddivisione geografica con *MixAll* dopo la PCA

Applicando il pacchetto *MixAll* il raggruppamento realizzato al dataset completo meno la variabile *GDP* e quello effettuato dopo l'analisi delle componenti principali porta a modelli di mistura differenti ma con un numero analogo di componenti. La suddivisione delle unità tuttavia non risulta completamente differente, in entrambi i casi infatti si noti come, ad esempio, le principali economie in Europa, quali Germania, Regno Unito e Francia, vengano poste sempre in un medesimo cluster.

Si sottolinea inoltre che non sono state rappresentate graficamente le densità delle Gaussiane che compongono la mistura poiché alcune componenti non contengono osservazioni sufficienti per consentire il disegno di ellissi. Ad ogni modo, come nell'applicazione del pacchetto *mclust*, si può immaginare anche in questo caso una sovrapposizione delle componenti.

3.3 Conclusioni

Entrambi i pacchetti utilizzati per l'applicazione del *model-based clustering* presentano alcune limitazioni. Il pacchetto *mclust* richiede che non siano presenti valori mancanti, rendendo necessaria una preventiva imputazione qualora apparissero valori mancanti nel dataset. *MixAll* supera questo limite ma, assumendo che la matrice di varianze e covarianze sia diagonale, ha un campo di applicazione ancora più ristretto. Tuttavia, sia la presenza di dati mancanti sia la correlazione nei dati, è usuale. Com'è emerso dall'analisi esplorativa entrambi questi problemi sono presenti nel dataset in esame.

La difficoltà nel cogliere la struttura sottostante i dati applicando *MixAll* è giustificata dalla forte correlazione presente per la maggior parte delle variabili.

Per quanto riguarda l'utilizzo dell'analisi delle componenti principali prima di effettuare il *model-based clustering* è possibile notare che in *MixAll* i modelli di mistura Gaussiani implementati siano differenti ma con lo stesso numero di componenti. Nonostante la differenza nelle caratteristiche dei due modelli, la suddivisione delle unità, eccetto per pochi casi, è essenzialmente molto simile. L'applicazione a priori dell'analisi delle componenti principali consente tuttavia di facilitare il processo computazionale per l'implementazione del modello, richiedendo la stima di un numero minore di parametri.

È possibile fare un discorso analogo riguardo ai modelli implementati con *mclust*. In entrambi i casi vengono implementati modelli di mistura Gaussiani con $k = 3$ componenti e l'applicazione della PCA prima di effettuare il *clustering* contribuisce a diminuire notevolmente il numero di parametri da stimare. Tuttavia la suddivisione ottenuta con i due metodi non è la medesima.

Non essendoci una reale suddivisione dei Paesi a livello economico, non è possibile valutare la performance del *model-based clustering* implementato tramite i due pacchetti R *mclust* e *MixAll*. L'utilizzo di indici come il *Rand Index* che consente di confrontare due suddivisioni, come quella reale e quella ottenuta sul dataset, non è in questo caso possibile. Nonostante siano presenti indici di *cluster validation* di tipo interno che richiedono solamente i dati disponibili e si basano sul concetto di distanza, è stato dimostrato che questi non siano ottimali per valutare i raggruppamenti ottenuti in un contesto in cui i cluster non abbiano una forma convessa (Grün, 2018).

Lo scopo di questa tesi rimane quello di illustrare e implementare una metodologia di *cluster analysis* più robusta, quale il *model-based clustering*, ricordando, ad ogni modo, che nella maggior parte dei casi la *cluster analysis* viene utilizzata come una tecnica di analisi esplorativa.

Capitolo 4

Studio di simulazione

In questo capitolo vengono illustrati i risultati di uno studio di simulazione, realizzato allo scopo di fornire una valutazione più precisa sulla performance dei due pacchetti R, *mclust* e *MixAll*, utilizzati per l'applicazione dell'analisi dei gruppi basata su modello.

In questo studio di simulazione verranno generati 100 dataset provenienti da una mistura di Gaussiane. A differenza di quanto visto nel capitolo 3 in questo caso si ha una conoscenza completa sulla struttura dei dati ed in particolare del numero di gruppi in cui le osservazioni sono suddivise.

Vengono inoltre considerati differenti scenari che coinvolgono diverse percentuali di valori mancanti, specificatamente del 5 e 10%, con uguali e differenti proporzioni di essi tra variabili del dataset, e diverse strutture della matrice di varianze e covarianze per le componenti della mistura.

Per quanto riguarda la dimensione dei dataset, sono state generate tante unità statistiche quanti sono i Paesi nel dataset economico utilizzato per l'analisi dei gruppi basata su modello, i cui risultati sono riportati nel capitolo 3. Per semplicità di rappresentazione grafica, sono state simulate solamente due variabili. I gruppi in cui le unità statistiche sono suddivise, e quindi le componenti della mistura, sono $k = 3$.

Avendo a disposizione la reale classificazione delle unità statistiche è possibile utilizzare misure di *cluster validation*. In particolare, per questo studio di simulazione verrà utilizzato il *Rand Index*.

Si consideri un insieme di n unità statistiche e due partizioni $C = \{C_1, ..., C_m\}$ e $P = \{P_1, ..., P_s\}$ di esse da confrontare. Siano (x_v, x_u), $u, v = 1, ..., n, u \neq v$ coppie di osservazioni, definiamo:

- a: se entrambe le osservazioni appartengono allo stesso gruppo in C e in P;
- b: se le osservazioni fanno parte dello stesso gruppo in C ma gruppi differenti in P;
- c: se le osservazioni appartengono a gruppi differenti in C ma negli stessi gruppi in P;
- d: se le osservazioni fanno parte di differenti gruppi sia in C sia in P.

Il *Rand Index* allora risulta essere:

$$R = \frac{a + d}{a + b + c + d}, \tag{4.1}$$

dove $a+b+c+d = M$ è il numero massimo di coppie di osservazioni, ossia $M = \binom{n}{2} = \frac{n(n-1)}{2}$. Il *Rand Index* assume valori in $(0, 1)$, in particolare è pari a 0 quando le due partizioni non concordano per nessuna coppia ed è pari a 1 quando le partizioni coincidono. L'indice misura quindi il grado di somiglianza tra le due partizioni C e P.

Per questo studio di simulazione, il *Rand Index* viene calcolato utilizzando il comando *RI* del pacchetto *aricode* (Chiquet J. e altri, 2022).

4.1 Dati mancanti

In primo luogo consideriamo gli scenari legati alla presenza di dati mancanti all'interno del dataset.

Per cercare di isolare il problema della presenza dei dati mancanti dalla complessità legata alla struttura dei cluster, il dataset è stato generato con uguali matrici di varianze e covarianze tra componenti della mistura, cambieranno invece le proporzioni e i vettori delle medie. In questo modo sarà possibile valutare quanto la presenza di valori mancanti influisce sui risultati ottenuti.

Vengono riportati, nell'ordine, le proporzioni della mistura (formula 4.2), i vettori delle medie (formula 4.3) e le matrici di varianze e covarianze (formula 4.4) per le componenti della mistura utilizzati per la simulazione dei dataset:

$$p = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.5 \end{bmatrix}, \tag{4.2}$$

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \mu_3 = \begin{bmatrix} 0 \\ 10 \end{bmatrix}, \tag{4.3}$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{4.4}$$

Il quadro di riferimento è il seguente: una presenza di valori mancanti generati casualmente nel dataset del 5% e del 10% con uguali proporzioni tra variabili e del 10% ma con differenti proporzioni.

Per la generazione di dati mancanti in modo casuale è stato utilizzato il comando *delete_MCAR* del pacchetto *missMethods* (Rockel, 2022) impostando la probabilità che il dato sia mancante a 0.05 e 0.1; in questo modo è possibile ottenere una presenza casuale di valori mancanti del 5% e del 10% rispettivamente.

Per quanto riguarda invece la generazione di dati mancanti in modo casuale ma con differenti proporzioni tra variabili, è stata effettuata un'estrazione casuale degli indici per le due variabili del dataset; in particolare, una selezione casuale di 6 indici per la prima variabile e di 2 indici per la seconda. Anche in questo caso si otterrà una presenza di valori mancanti del 10% ma maggiormente concentrati nella prima variabile piuttosto che nella seconda.

Nella figura 4.1 vengono rappresentati i pattern dei dati mancanti all'interno del primo dataset simulato per i tre scenari legati alla presenza di valori mancanti considerati.

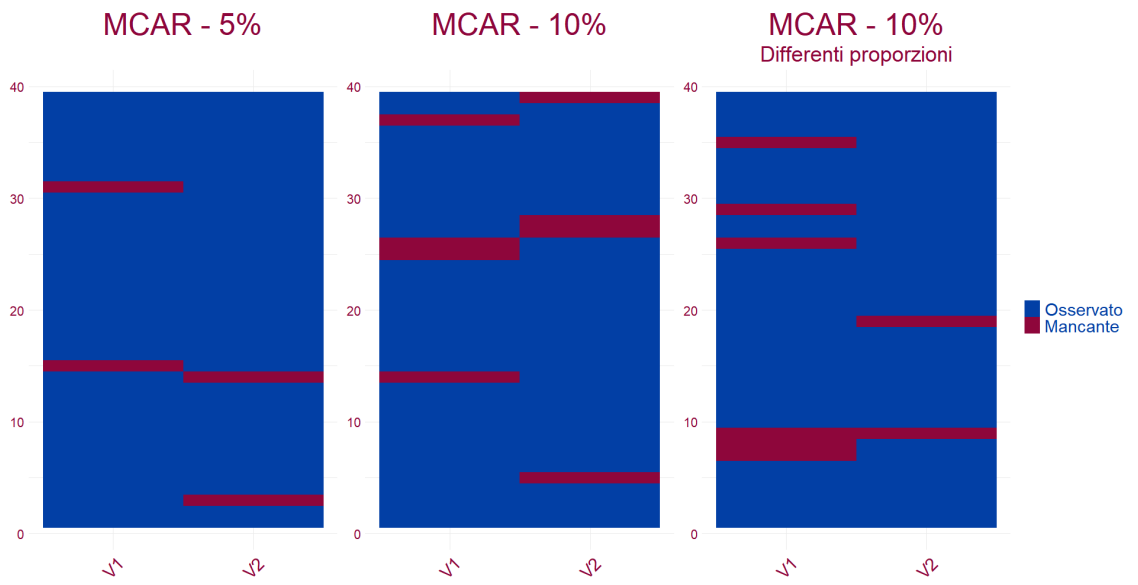


Figura 4.1: Scenari legati alla presenza di dati mancanti: presenza del 5% e del 10% mancanti in modo casuale (MCAR) e del 10% mancanti in modo casuale (MCAR) suddivisi in differenti proporzioni

4.1.1 *mclust*

Si procede ora con l'applicazione del *model-based clustering* tramite l'utilizzo del pacchetto *mclust* ai tre scenari legati alla presenza dei valori mancanti all'interno del dataset descritti nel paragrafo precedente. Verranno illustrati i risultati ottenuti per il primo dataset per poi passare alla valutazione del *Rand Index* ottenuto per i 100 dataset simulati.

Si ricorda che forma, dimensione e orientamento dei cluster sono definiti in fase di generazione dei dataset. In questo contesto si considera il modello di mistura Gaussiana più semplice nella classe dei modelli in cui la forma dei cluster è ellittica. I gruppi hanno quindi un'uguale forma, dimensione e orientamento rispetto agli assi.

Il primo quadro di riferimento è quello relativo ad una presenza del 5% di dati mancanti in modo casuale con uguali percentuali di questi ultimi tra variabili all'interno del dataset. Si sottolinea, come già descritto nel capitolo 3, che *mclust* richiede che l'imputazione dei dati mancanti avvenga prima della stima del modello di mistura; come effettuato nel dataset economico, anche in questo ambito, viene utilizzato il comando *imputeData* inserito all'interno del pacchetto *mclust*.

Come si nota dalla figura 4.2 per il primo dataset simulato viene selezionato, correttamente, il modello *EEI* con $k = 3$ componenti. Nella figura 4.3 viene riportato un grafico che rappresenta la classificazione ottenuta. Ci si trova in una situazione più definita rispetto al dataset economico reale analizzato nel capitolo 3, si evidenzia una minore sovrapposizione dei cluster. Inoltre, dalla tabella 4.1 si può notare come solo un'unità sia classificata erroneamente.

	Cluster da *mclust*		
Cluster reali	1	2	3
1	12	0	1
2	0	0	9
3	0	17	0

Tabella 4.1: Confronto tra classificazione ottenuta con *mclust* e classificazione del primo dataset simulato con il 5 % di dati mancanti in modo casuale

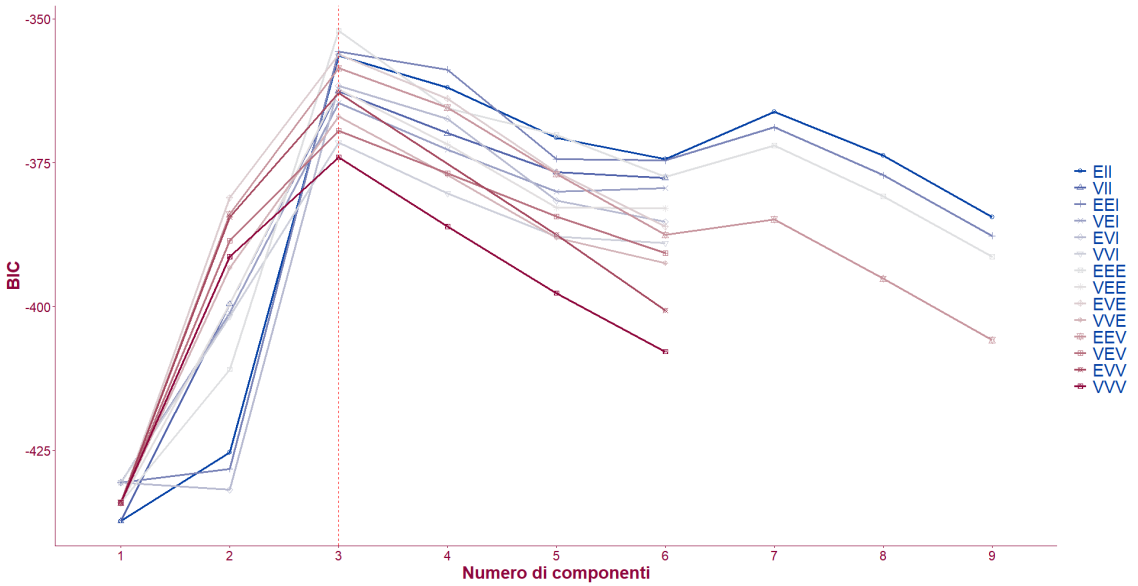


Figura 4.2: Selezione del modello in *mclust* nel primo dataset simulato con la presenza del 5% di dati mancanti in modo casuale

Il *Rand Index* associato alla classificazione riportata nella figura 4.3 è pari a 0.97, si ottiene quindi un ottimo raggruppamento. Vengono rappresentati in figura 4.4 l'istogramma e la densità del *Rand Index* delle classificazioni ottenute per tutti e 100 i dataset simulati.

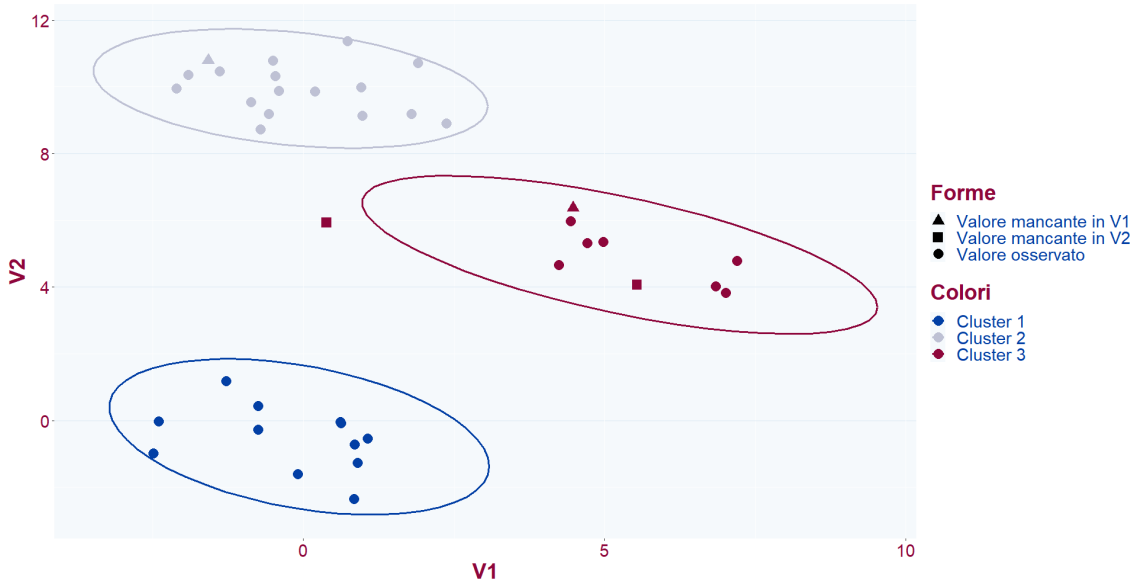


Figura 4.3: Cluster ottenuti con *mclust* nel primo dataset simulato con la presenza del 5% di dati mancanti in modo casuale

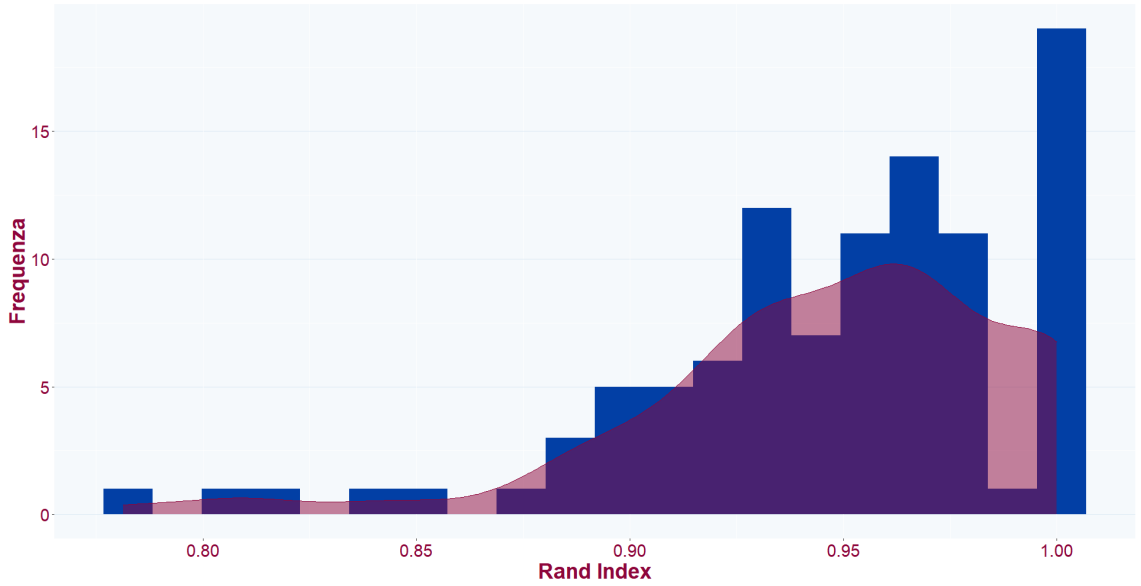


Figura 4.4: Istogramma e densità del *Rand Index* per le classificazioni ottenute con *mclust* a partire dai dataset simulati con la presenza del 5% di dati mancanti in modo casuale

Dalla figura 4.4 si può notare come vi sia un'alta frequenza di valori prossimi a 1 del *Rand Index*. Il minimo valore assunto da questo indice rimane comunque alto, pari a 0.78.

Consideriamo il secondo scenario, ossia una presenza del 10% di dati mancanti in modo casuale all'interno del dataset con uguali percentuali di questi ultimi tra variabili.

In questo caso per il primo dataset simulato, il modello selezionato è il modello *EII*, con $k = 4$ componenti, che prevede che i cluster abbiano forma sferica con uguale volume. Viene quindi formato un cluster in più rispetto al numero di gruppi presenti realmente nei dati. Oltre a questo si nota una maggiore sovrapposizione e incertezza nella classificazione, come emerge dalla figura 4.6 e dalla tabella 4.2.

	Cluster da *mclust*			
Cluster reali	1	2	3	4
1	12	1	0	0
2	0	0	9	0
3	1	14	0	2

Tabella 4.2: Confronto tra classificazione ottenuta con *mclust* e classificazione del primo dataset simulato con il 10% di dati mancanti in modo casuale

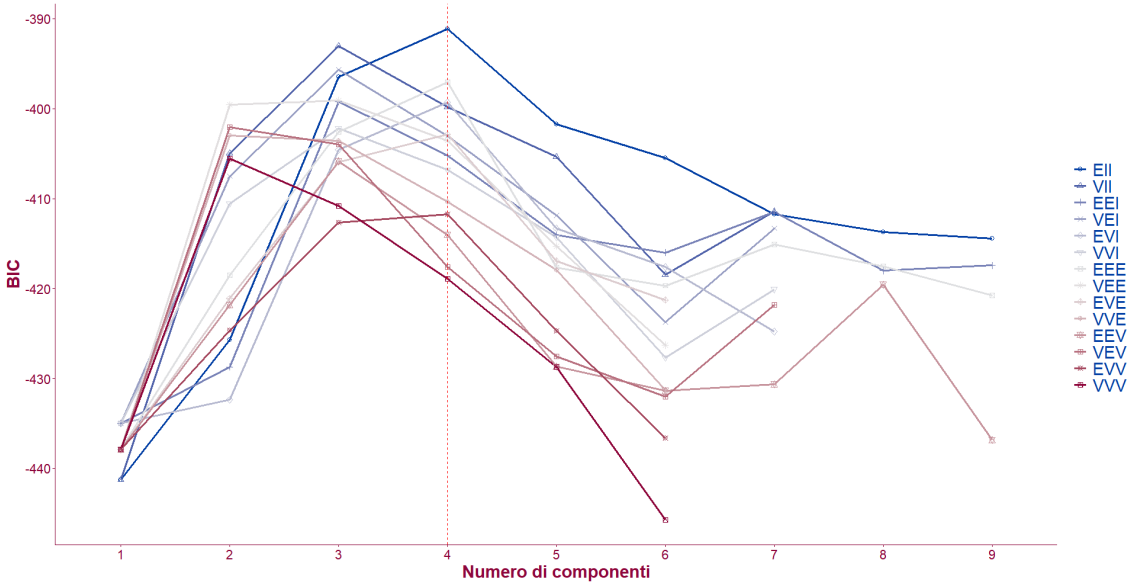


Figura 4.5: Selezione del modello in *mclust* nel primo dataset simulato con la presenza del 10% di dati mancanti in modo casuale

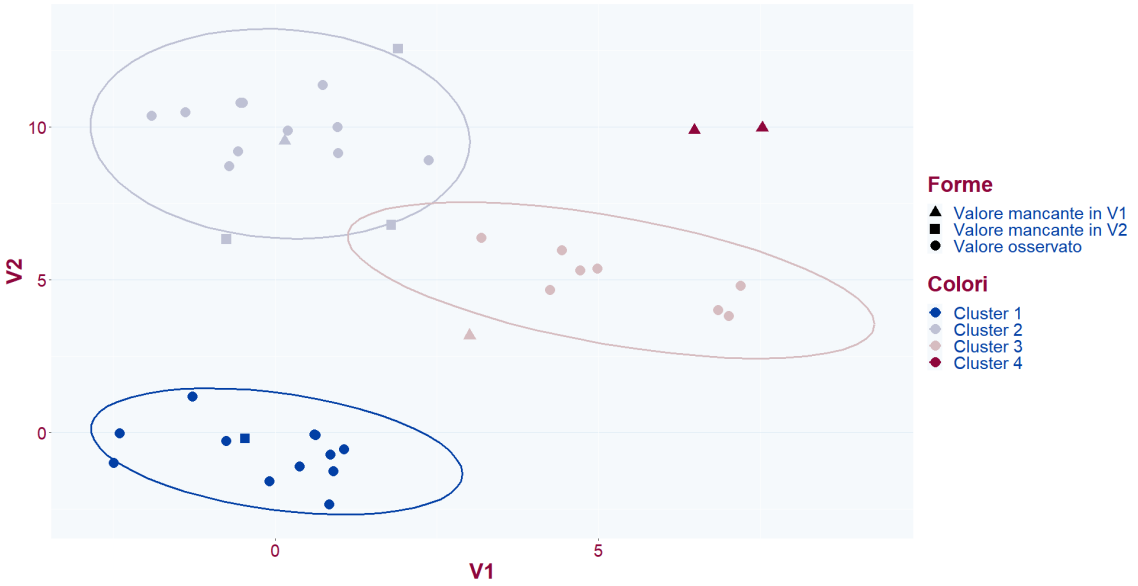


Figura 4.6: Cluster ottenuti con *mclust* nel primo dataset simulato con la presenza del 10% di dati mancanti in modo casuale

Il valore del *Rand Index* associato alla classificazione in figura 4.6 è pari a 0.89. Vengono riportati nella figura 4.7 l'istogramma e la densità del *Rand Index* delle classificazioni ottenute per tutti i dataset simulati con una presenza del 10% di valori mancanti in modo casuale.

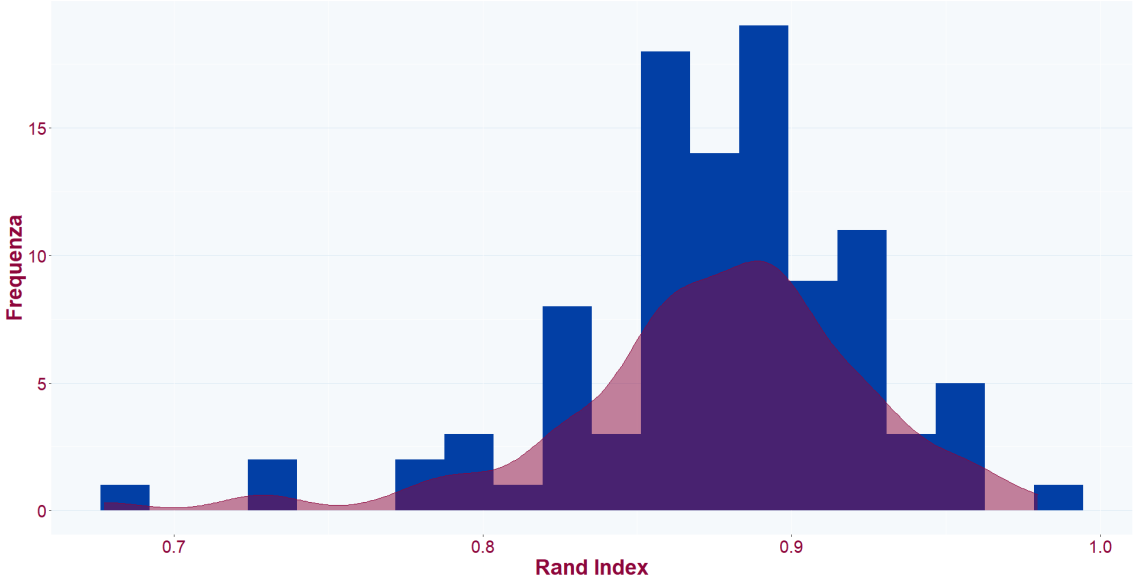


Figura 4.7: Istogramma e densità del *Rand Index* per le classificazioni ottenute con *mclust* a partire dai dataset simulati con la presenza del 10% di dati mancanti in modo casuale

In questo caso la distribuzione del *Rand Index* si concentra maggiormente vicino al valore 0.9. In generale si nota che, ad una maggiore presenza di valori mancanti, corrisponde una maggiore difficoltà nella classificazione. Rispetto alla figura 4.4 emerge infatti una diminuzione della frequenza di valori del *Rand Index* superiore allo 0.9.

Infine, si considera il caso in cui vi sia il 10% di dati mancanti ma in proporzioni diverse per le due variabili.

Il modello che viene selezionato per il primo dataset simulato è il modello *EEE* (uguale forma, dimensione e orientamento) con $k = 5$ componenti.

Dalla tabella 4.3 emerge che il numero di unità classificate in modo errato non cambia rispetto al raggruppamento ottenuto applicando il pacchetto *mclust* ai dati in cui la presenza di valori mancanti ha la stessa percentuale del caso analizzato, ossia del 10% rispetto alla totalità dei dati, ma in cui i valori mancanti sono suddivisi con uguali proporzioni tra le due variabili all'interno del dataset.

	Cluster da *mclust*				
Cluster reali	1	2	3	4	5
1	11	1	1	0	0
2	0	0	8	1	0
3	0	0	0	0	17

Tabella 4.3: Confronto tra classificazione ottenuta con *mclust* e classificazione del primo dataset simulato con il 10% di dati mancanti in modo casuale suddivisi in diverse proporzioni tra variabili

Di seguito vengono riportati in figura 4.8 la selezione del modello effettuata tramite il valore del *BIC* e nella figura 4.9 la classificazione ottenuta dal modello selezionato per il primo dataset simulato.

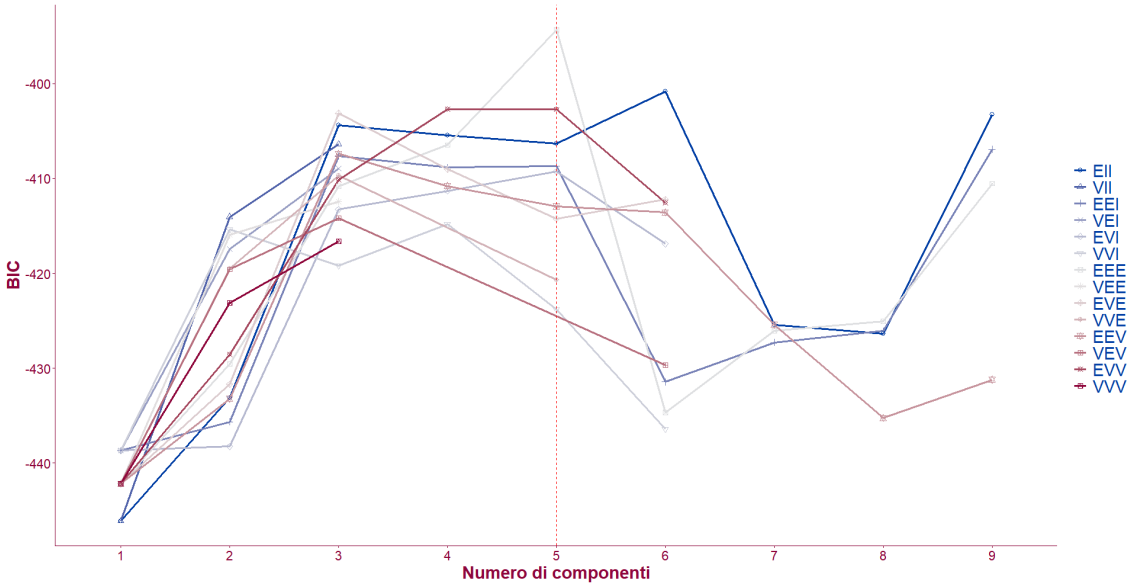


Figura 4.8: Selezione del modello in *mclust* nel primo dataset simulato con la presenza del 10% di dati mancanti in modo casuale suddivisi in diverse proporzioni tra variabili

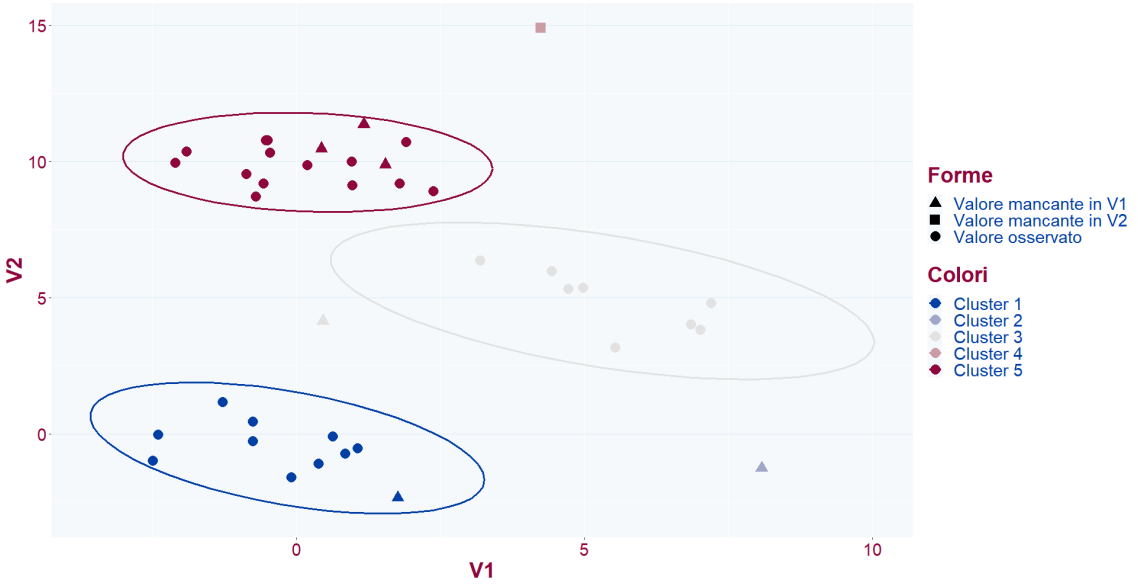


Figura 4.9: Cluster ottenuti con *mclust* nel primo dataset simulato con la presenza del 10% di dati mancanti in modo casuale suddivisi in diverse proporzioni tra variabili

Il valore del *Rand Index* per la classificazione rappresentata in figura 4.9 è pari a 0.95. Interessante notare che questo indice assuma un valore maggiore nel caso in cui la presenza di dati mancanti è del 10% ma ci si trova in una situazione in cui i valori mancanti sono suddivisi con differenti proporzioni per le due variabili che costituiscono il dataset simulato. Ma potendosi trattare solamente di un caso fortunato, anche in questo caso si riportano in figura 4.10 l'istogramma e la densità del *Rand Index* per lo scenario che prevede la presenza del 10% di dati mancanti suddivisi in diverse proporzioni per le due variabili.

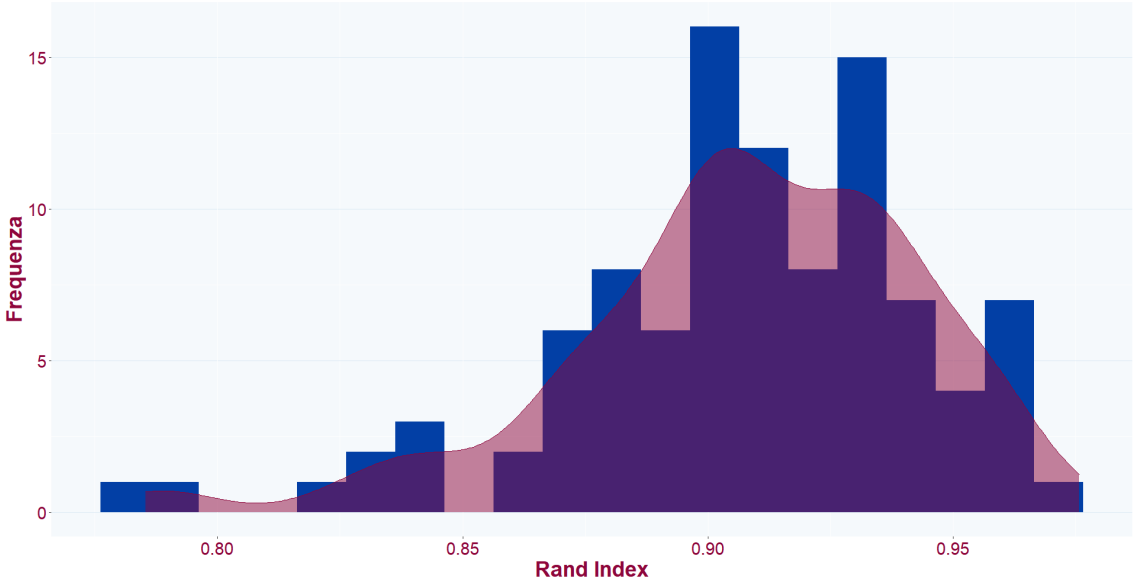


Figura 4.10: Istogramma e densità del *Rand Index* per le classificazioni ottenute con *mclust* a partire dai dataset simulati con la presenza del 10% di dati mancanti in modo casuale suddivisi in diverse proporzioni tra variabili

Dalla figura 4.10 emerge che, anche in questo caso, il valore medio del *Rand Index* è vicino allo 0.9.

Vengono di seguito riportati nella tabella 4.4 le medie dei valori del *Rand Index* per i raggruppamenti ottenuti utilizzando *mclust* per tutti e 100 i dataset simulati per i tre scenari legati alla presenza dei dati mancanti considerati in questo paragrafo.

Alla luce dei valori nella tabella 4.4, si può affermare che tramite l'utilizzo di *mclust* si ottengono buoni raggruppamenti anche con una maggior presenza di dati mancanti distribuiti non uniformemente.

	Media dei *Rand Index*
MCAR (5%)	0.95
MCAR (10%)	0.88
MCAR (10%) con differenti proporzioni	0.91

Tabella 4.4: Valore medio dei *Rand Index* per le partizioni ottenute con *mclust* negli scenari legati alla presenza di dati mancanti

4.1.2 *MixAll*

In questo paragrafo verrà esaminato quanto ottenuto dall'applicazione del pacchetto *MixAll* nei tre scenari legati alla presenza di dati mancanti. Procedendo analogamente al paragrafo 4.1.2, verrano presentati i risultati considerando il primo dataset generato con, nell'ordine, la presenza di dati mancanti in modo casuale del 5%, del 10% suddivisi in uguali e differenti proporzioni tra variabili.

Per il primo dataset simulato con il 5% di dati mancanti in modo casuale viene selezionato il modello *gaussian_pk_sj* (differenti proporzioni della mistura e deviazioni standard delle variabili ma uguali deviazioni standard nei cluster) con $k = 5$ componenti. Incrociando la reale classificazione con quella ottenuta dal modello stimato da *MixAll* si ottiene la tabella 4.5.

	Cluster da *MixAll*				
Cluster reali	1	2	3	4	5
1	0	12	0	1	0
2	0	0	0	1	8
3	6	0	11	0	17

Tabella 4.5: Confronto tra classificazione ottenuta con *MixAll* e classificazione del primo dataset simulato con il 5% di dati mancanti in modo casuale

Anche senza il calcolo del *Rand Index* si può vedere come vi sia un numero maggiore di unità non classificate in modo corretto, rispetto a quanto ottenuto dall'applicazione di *mclust* per lo stesso dataset.

Il raggruppamento, ottenuto per il primo dataset simulato, è riportato nella figura 4.11. Si evidenzia una sovrapposizione tra componenti della mistura Gaussiana e ne consegue una maggiore incertezza nella classificazione delle unità.

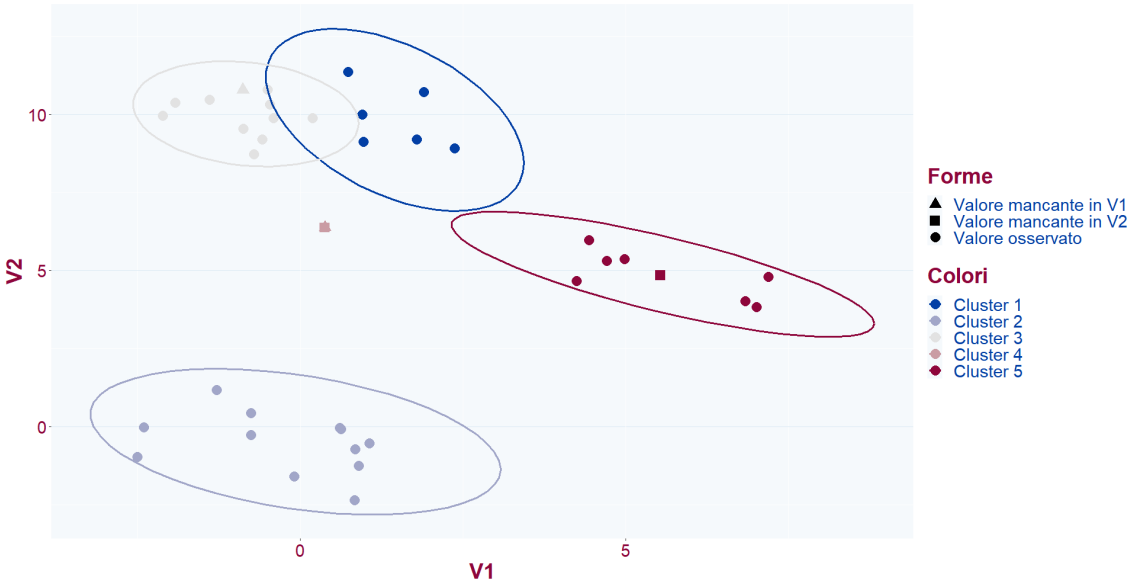


Figura 4.11: Cluster ottenuti con *MixAll* nel primo dataset simulato con la presenza del 5% di dati mancanti in modo casuale

L'istogramma e la densità del *Rand Index* per i raggruppamenti ottenuti mediante l'utilizzo di *MixAll* vengono riportati nella figura 4.12.

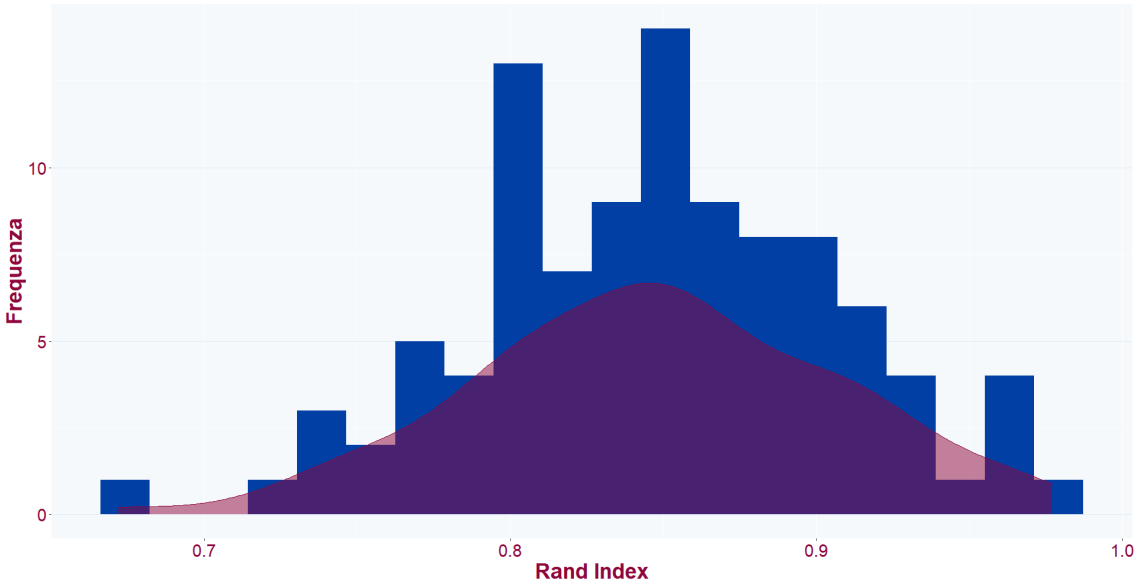


Figura 4.12: Istogramma e densità del *Rand Index* per le classificazioni ottenute con *MixAll* a partire dai dataset simulati con la presenza del 5% di dati mancanti in modo casuale

In generale si ottengono buoni raggruppamenti considerando il valore medio dei *Rand Index* per le varie classificazioni pari a 0.85. Il valore minimo in questo caso è di circa 0.67, un valore inferiore rispetto a quanto ottenuto nello stesso scenario ma utilizzando, per l'applicazione del *model-based clustering*, il pacchetto *mclust*.

Nel secondo scenario preso in esame, ossia una presenza del 10% di dati mancanti in modo casuale il modello selezionato è $gaussian\_pk\_s$ con $k = 5$ componenti. A differenza del modello stimato da *MixAll* sul dataset con la presenza del 5% di dati mancanti casualmente, si assume quindi l'uguaglianza delle deviazioni standard nelle variabili.

Viene riportato nella tabella 4.6 il confronto tra la reale classificazione e quella ottenuta applicando il pacchetto *MixAll*.

	Cluster da *MixAll*				
Cluster reali	1	2	3	4	5
1	0	4	8	1	0
2	8	0	0	0	1
3	0	0	0	16	1

Tabella 4.6: Confronto tra classificazione ottenuta con *MixAll* e classificazione del primo dataset simulato con il 10% di dati mancanti in modo casuale

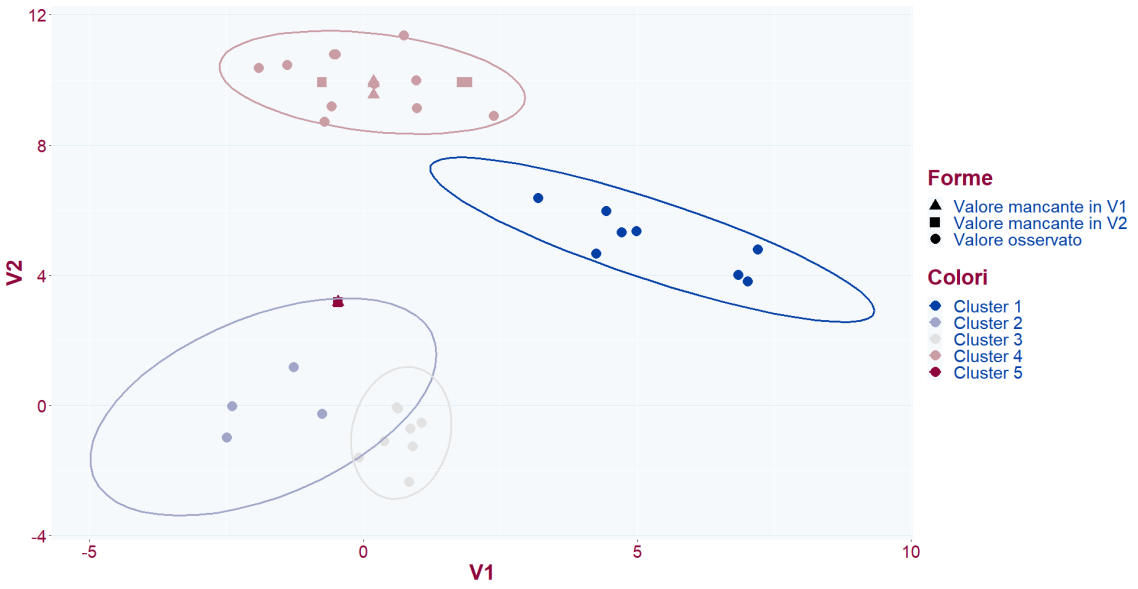


Figura 4.13: Cluster ottenuti con *MixAll* nel primo dataset simulato con la presenza del 10% di dati mancanti in modo casuale

Anche in questo caso, come si può vedere dalla figura 4.13, vi è una sovrapposizione in particolare tra due delle componenti della mistura. A differenza del raggruppamento in figura 4.11 vi è infatti una maggiore incertezza nella classificazione delle unità legate alle componenti nella parte inferiore del grafico (Cluster 1 e Cluster 2 nella figura 4.13).

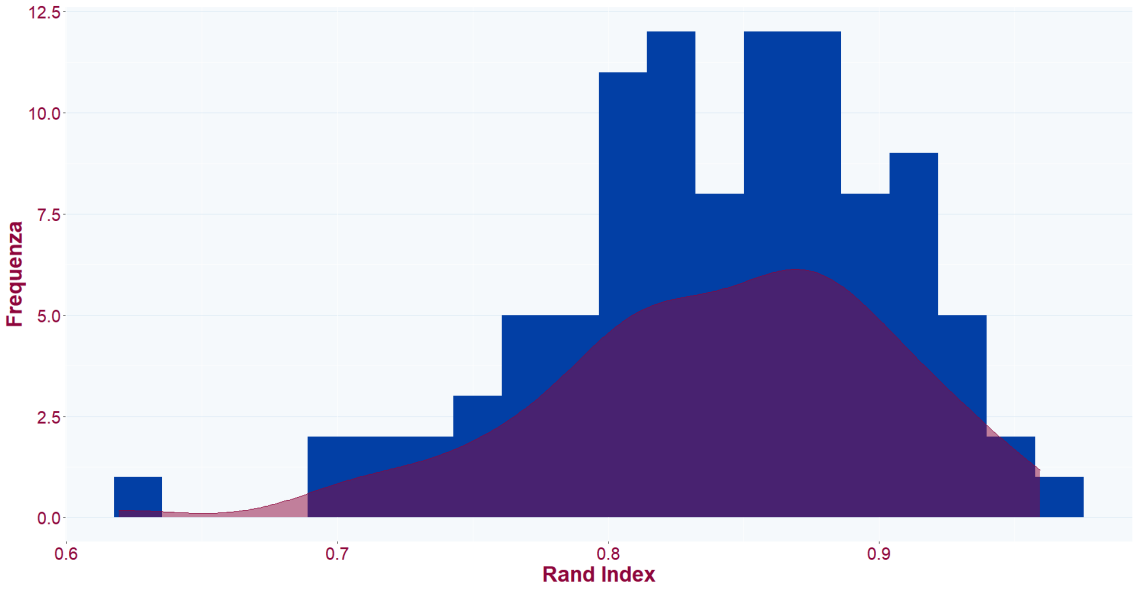


Figura 4.14: Istogramma e densità del *Rand Index* per le classificazioni ottenute con *MixAll* a partire dai dataset simulati con la presenza del 10% di dati mancanti in modo casuale

I raggruppamenti formati per i dataset simulati con il 10% di dati mancanti in modo casuale assumono buoni valori del *Rand Index*. Dalla figura 4.14 si può vedere come il valore medio dell'indice rimane attorno allo 0.85, analogamente al primo scenario analizzato.

Si considera infine l'ultimo scenario: presenza di dati mancanti non equiripartiti con una percentuale del 10%.

Il modello selezionato per il primo dataset simulato è *gaussian_pk_sjk*, che assume differenti proporzioni di mistura, deviazioni standard nelle variabili e deviazioni standard nei cluster, con $k = 5$ componenti.

Come negli scenari precedenti viene riportata la tabella che consente di confrontare la classificazione ottenuta con quella reale (tabella 4.7) e il grafico con la rappresentazione del raggruppamento effettuato (figura 4.15).

	Cluster da *MixAll*				
Cluster reali	1	2	3	4	5
1	1	0	0	12	0
2	2	7	0	0	0
3	0	0	14	0	0

Tabella 4.7: Confronto tra classificazione ottenuta con *MixAll* e classificazione del primo dataset simulato con il 10% di dati mancanti in modo casuale suddivisi in diverse proporzioni tra variabili

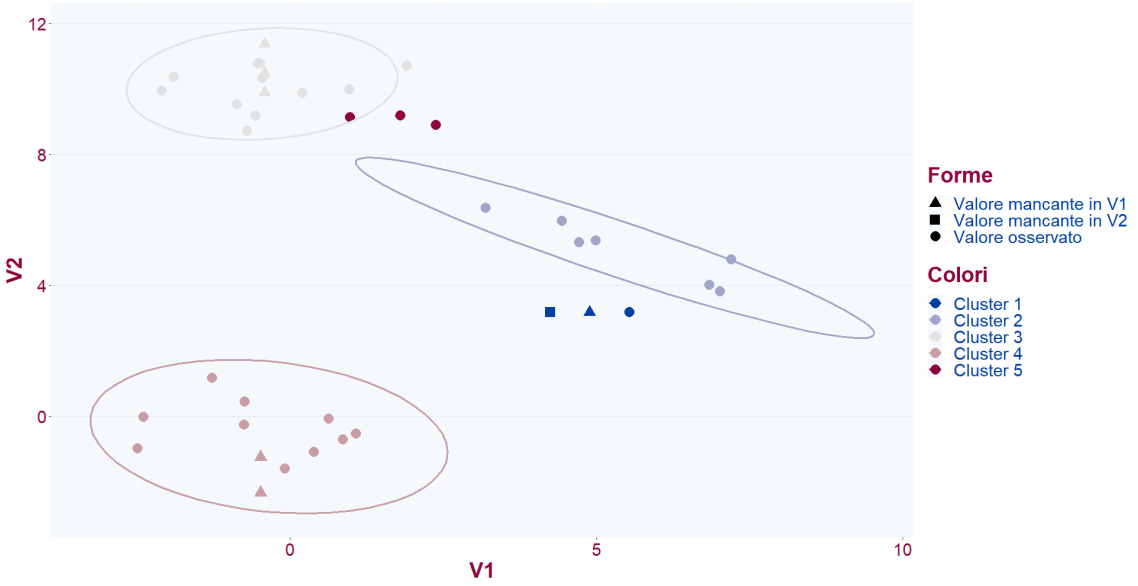


Figura 4.15: Cluster ottenuti con *MixAll* al dataset simulato con la presenza del 10% di dati mancanti in modo casuale suddivisi in diverse proporzioni tra variabili

Nonostante i dati mancanti siano stati generati non in modo equiripartito emerge una minore sovrapposizione dei cluster e quindi una minore incertezza in fase di classificazione delle unità statistiche.

L'istogramma e la densità del *Rand Index* sono riportati nella figura 4.16 e sottolineano che, nonostante una presenza di valori mancanti non distribuiti uniformemente tra le due variabili, anche in questo scenario sono state ottenute ottime classificazioni per i dataset simulati.

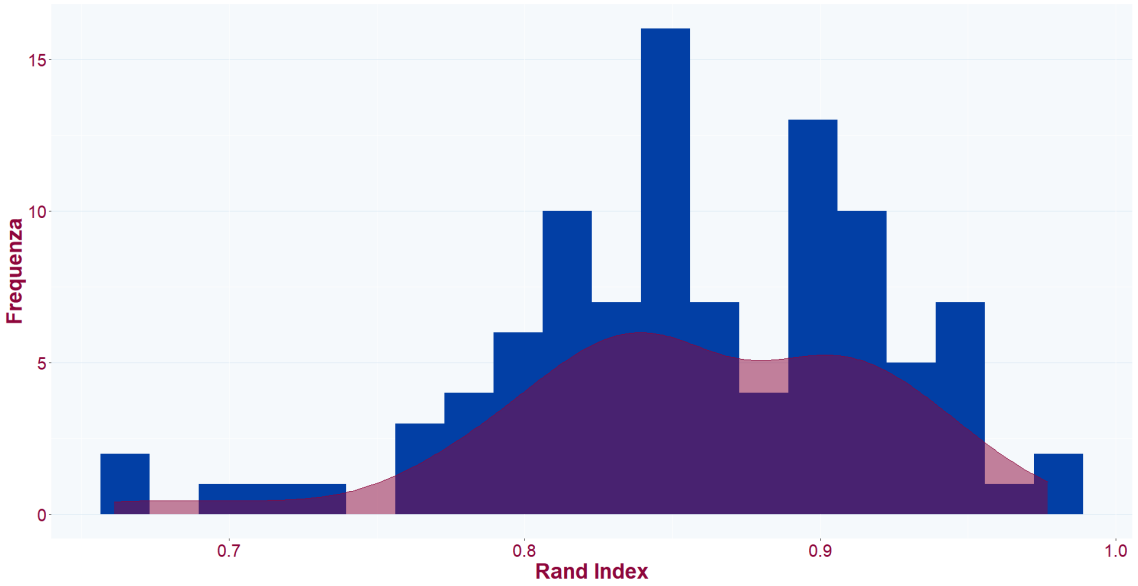


Figura 4.16: Istogramma e densità del *Rand Index* per le classificazioni ottenute con *MixAll* a partire dai dataset simulati con la presenza del 10% di dati mancanti in modo casuale suddivisi in diverse proporzioni tra variabili

Vengono riportati, nella tabella 4.8, i valori medi assunti dal *Rand Index* ottenuti utilizzando *MixAll* per l'applicazione del *model-based clustering* ai dataset simulati nei tre scenari legati alla presenza di valori mancanti.

	Media del *Rand Index*
MCAR (5%)	0.85
MCAR (10%)	0.84
MCAR (10%) con differenti proporzioni	0.86

Tabella 4.8: Rand Index per le partizioni ottenute negli scenari legati alla presenza di dati mancanti con *MixAll*

Nei tre casi considerati i valori medi dei *Rand Index* sono essenzialmente molto simili. Si può notare che a differenza di quanto emerso nella tabella 4.4, in questo caso non vi è un netto miglioramento del *Rand Index* tra la situazione più semplice e quella più complicata tra quelle considerate.

Per riassumere quanto emerso vengono riportati nella figura 4.17 i boxplot dei tassi di errata classificazione per i tre casi legati alla presenza dei dati mancanti suddivisi in base al pacchetto utilizzato per l'applicazione del *model-based clustering* ai dataset simulati.

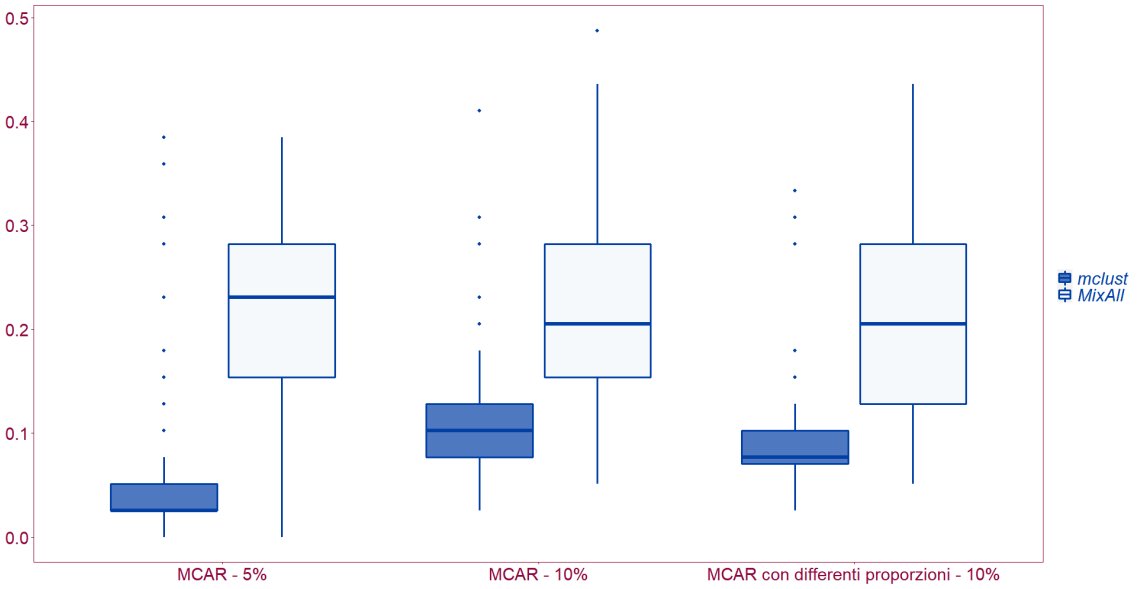


Figura 4.17: Boxplot dei tassi di errata classificazione ottenuti utilizzando *mclust* e *MixAll* per l'applicazione del *model-based clustering* ai dataset simulati con presenza di dati mancanti

Si può notare come in tutti e tre i casi si ottengono dei valori del tasso di errata classificazione superiori utilizzando il pacchetto *MixAll*.

Emerge inoltre un grande divario tra i valori assunti dal tasso di errata classificazione tra i raggruppamenti ottenuti utilizzando *mclust* e *MixAll* nella situazione più semplice tra quelle considerate, ossia la presenza del 5% di dati mancanti in modo casuale, a favore di *mclust*. Negli altri due casi questo divario diminuisce, ottenendo in ogni caso valori minori del tasso di errata classificazione utilizzando il pacchetto *mclust*.

Dai risultati ottenuti è quindi chiaro che diverse percentuali di valori mancanti, la modalità con cui i dati sono mancanti ma anche differenti valori imputati possano influire in modo più o meno importante sul raggruppamento ottenuto dall'applicazione del *model-based clustering*.

4.2 Aspetti geometrici dei cluster

Verrà preso ora in considerazione un caso più complicato, in termini di forma, dimensione e orientamento dei cluster.

In sede di generazione dei dati vengono quindi definite delle matrici di varianze e covarianze per le $k = 3$ componenti della mistura in modo tale da ottenere gruppi, sempre con forma ellittica, ma con differenti dimensioni e orientamenti. I gruppi saranno inoltre caratterizzati da un maggior avvicinamento.

Vengono riportati, nell'ordine, le proporzioni della mistura (formula 4.5), i vettori delle medie (formula 4.6) e le matrici di varianze e covarianze (formula 4.7) per le componenti della mistura utilizzati per la simulazione dei dataset:

$$p = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.5 \end{bmatrix}, \tag{4.5}$$

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \mu_3 = \begin{bmatrix} 0 \\ 6 \end{bmatrix}, \tag{4.6}$$

$$\Sigma_1 = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1.1 & -0.2 \\ -0.2 & 1.1 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{4.7}$$

Si ricorda che i 100 dataset simulati hanno lo stesso numero di unità statistiche del dataset economico analizzato nel capitolo 3 ma con solamente due variabili per semplificare la fase di rappresentazione dei risultati; la dimensione dei dataset generati è quindi 39×2.

Come nello studio di simulazione con differenti percentuali di dati mancanti si vuole anche in questo ambito fornire una valutazione dei due pacchetti R, *mclust* e *MixAll*, utilizzati per l'applicazione dell'analisi dei gruppi basata su modello al dataset economico.

La rappresentazione grafica del primo dataset generato è riportata nella figura 4.18.

4.2.1 *mclust*

Si procede in questo paragrafo con l'applicazione del *model-based clustering* ai dataset simulati tramite l'utilizzo di *mclust*.

Il modello selezionato per il primo dataset simulato è il modello *VVI* che prevede differenti forme e dimensioni ed orientamento allineato con

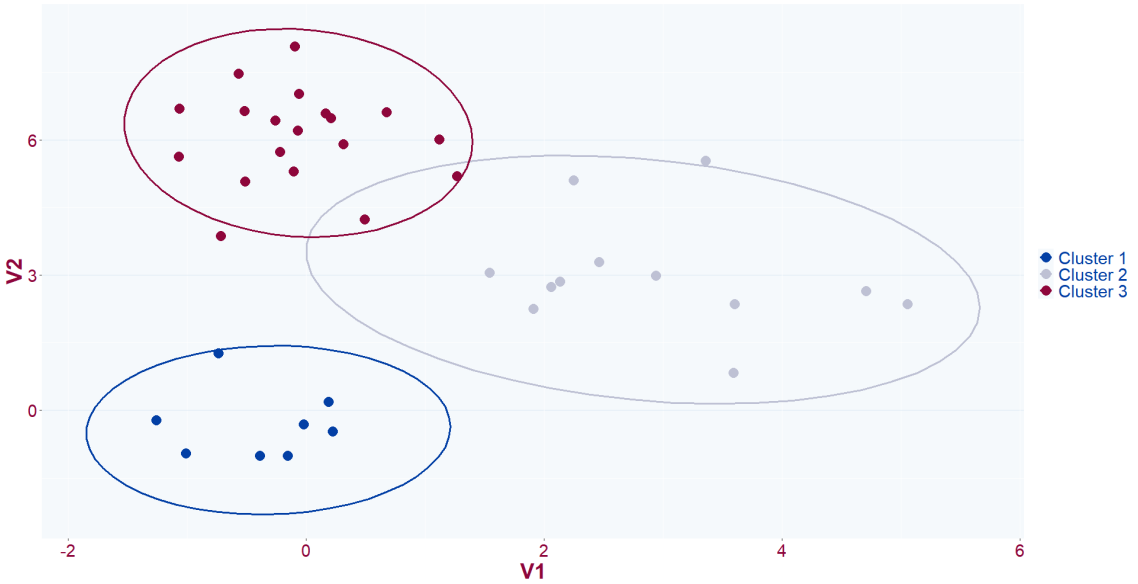


Figura 4.18: Cluster nel primo dataset simulato: dimensione e orientamento differenti

gli assi. Inoltre il numero di componenti del modello selezionato dal *BIC* è $k = 3$.

Osservando la tabella 4.9 che incrocia la classificazione ottenuta con quella realmente presente nel dataset generato si può notare come il numero di unità classificate nel cluster errato non sia così elevato.

	Cluster da *mclust*		
Cluster reali	1	2	3
1	7	1	0
2	0	12	0
3	0	3	16

Tabella 4.9: Confronto tra classificazione ottenuta con *mclust* e classificazione del primo dataset simulato i cui cluster hanno dimensione e orientamento differenti

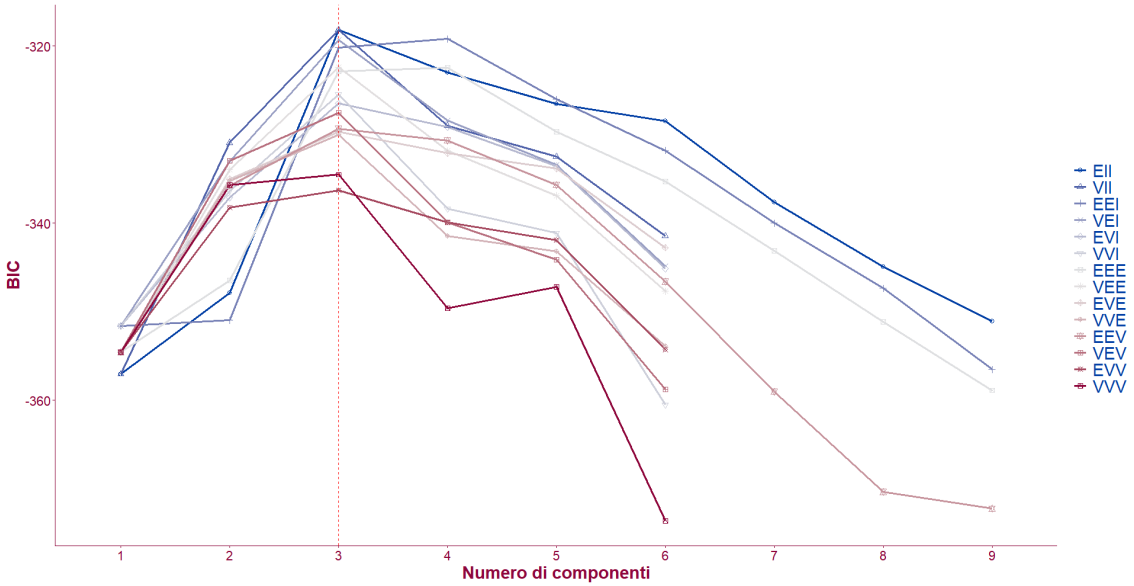


Figura 4.19: Selezione del modello in *mclust* nel primo dataset simulato i cui cluster hanno dimensione e orientamento differenti

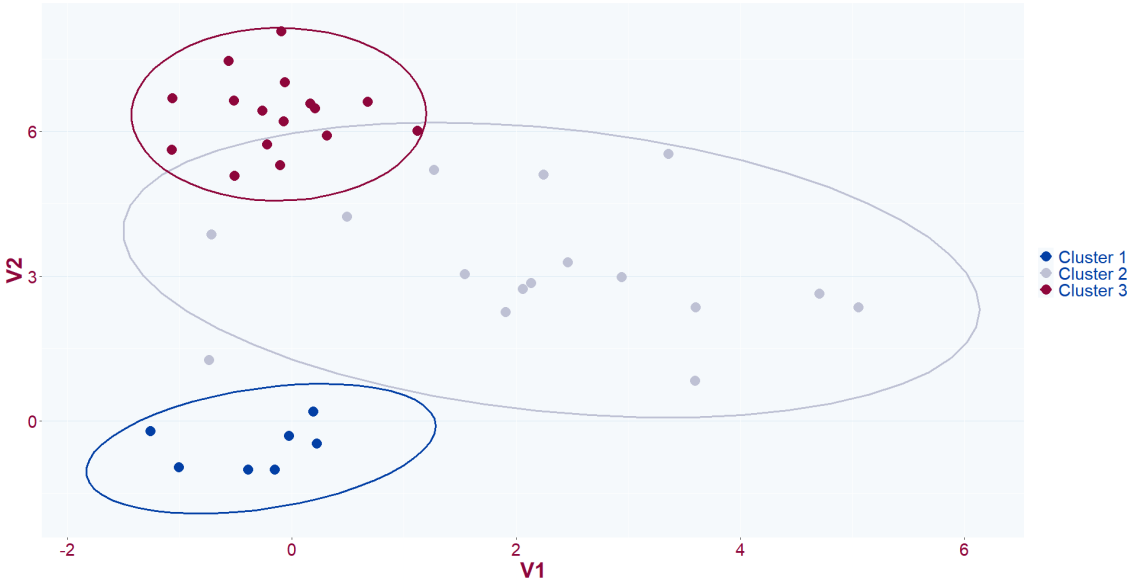


Figura 4.20: Cluster ottenuti con *mclust* nel primo dataset simulato i cui gruppi hanno dimensione e orientamento differenti

Confrontando le figure 4.18 e 4.20, si può affermare che utilizzando *mclust* al primo dataset generato si ottiene un buon raggruppamento delle osservazioni. Il *Rand Index* per questa classificazione è pari a $R = 0.86$.

Per valutare i raggruppamenti ottenuti per tutti e 100 i dataset simulati, nella figura 4.21 vengono rappresentati l'istogramma e la distribuzione del *Rand Index* nella situazione considerata.

In generale, considerando che il valore minimo del *Rand Index* è pari a 0.74, si può affermare che si ottengono ottime classificazioni delle unità statistiche.

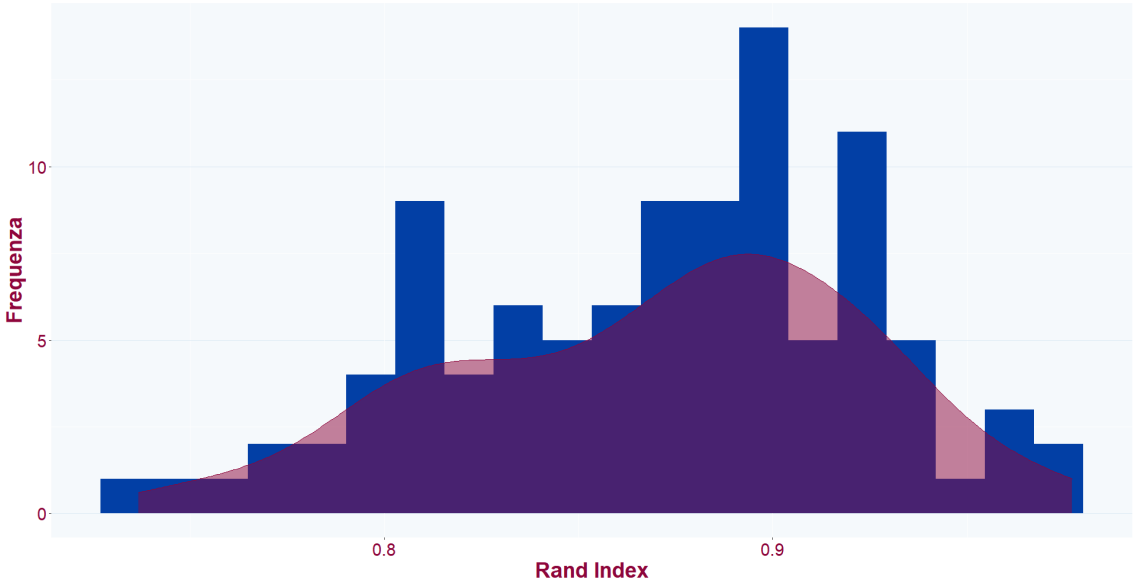


Figura 4.21: Istogramma e densità del *Rand Index* per le classificazioni ottenute con *mclust* a partire dai dataset simulati i cui cluster hanno dimensione e orientamento differenti

Modello selezionato	**Frequenza**
VII	62
EEI	18
VEE	5
VEI	5
EEE	3
VEV	2
VVE	2
EEV	2
EEI	1

Tabella 4.10: Frequenza dei modelli selezionati da *mclust* per i dataset simulati i cui cluster hanno dimensione e orientamento differenti

Nella tabella 4.10 vengono riportati i nomi dei modelli e la frequenza con cui essi vengono selezionati da *mclust* per i dataset simulati. Si ricordi che le $k = 3$ componenti della mistura Gaussiana presentano differenti matrici di varianze e covarianze che definiscono cluster con uguale forma, ellittica, ma dimensione e orientamento differenti. Si possono notare le caratteristiche dei cluster nel grafico in figura 4.18. Facendo riferimento alla tabella 2.1 si potrebbe definire come corretto il modello *VEV*.

Dalla tabella 4.10 si può notare come per la maggior parte dei dataset simulati venga selezionato il modello *VVI* che prevede per i cluster un'uguale forma sferica, volume differente e orientamento non definito. Il secondo modello più selezionato è il modello *EEI* che prevede che i gruppi abbiano un'uguale forma ellittica, orientamento allineato con gli assi e uguale volume.

Per la gran parte dei dataset quindi *mclust*, nonostante riesca a definire buoni raggruppamenti delle unità statistiche, non riesce a cogliere perfettamente le caratteristiche delle componenti della mistura. In particolare, non riesce ad identificare in modo esatto l'orientamento e la forma a differenza di quanto accade per il volume.

4.2.2 *MixAll*

L'applicazione del pacchetto *MixAll* nel contesto sopra descritto porta a selezionare per il primo dataset simulato il modello *gaussian_pk_sk* in cui le proporzioni e le deviazioni standard tra componenti della mistura sono differenti mentre le deviazioni standard nelle variabili sono uguali. Il numero di componenti selezionate è $k = 5$, quindi vengono formati due gruppi in più rispetto a quanto effettivamente presente nel dataset.

La classificazione ottenuta viene riportata nella figura 4.22. Si nota una maggiore difficoltà a cogliere la struttura sottostante il dataset generato.

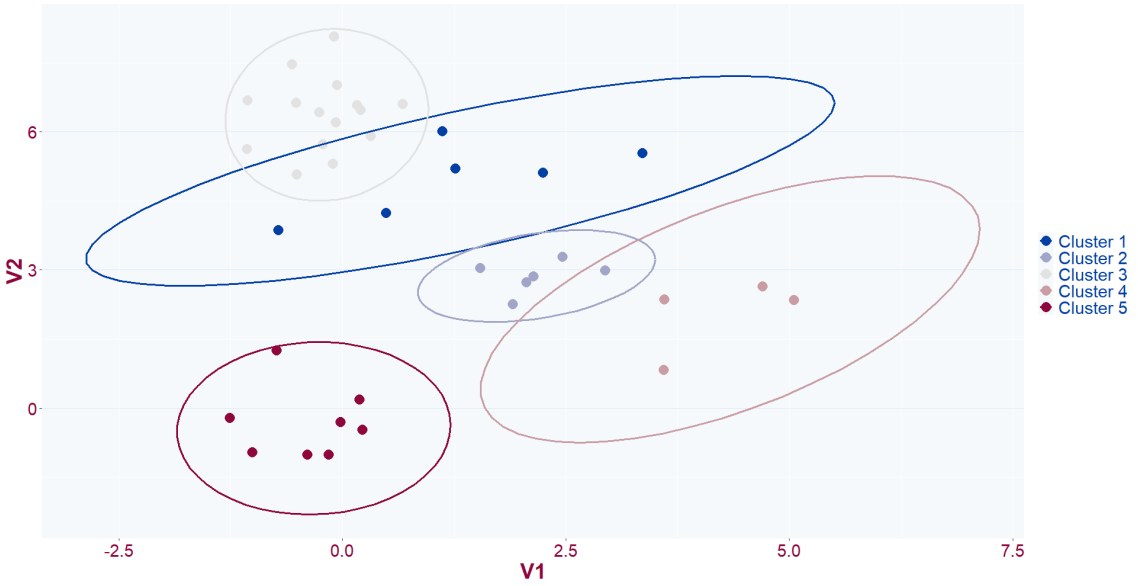


Figura 4.22: Cluster ottenuti con *MixAll* nel primo dataset simulato i cui cluster hanno dimensione e orientamento differenti

	Cluster da *MixAll*				
Cluster reali	1	2	3	4	5
1	0	0	0	0	8
2	2	6	0	4	0
3	4	0	15	0	0

Tabella 4.11: Confronto tra classificazione ottenuta con *MixAll* e classificazione del primo dataset simulato i cui cluster hanno dimensione e orientamento differenti

Il *Rand Index* per il raggruppamento ottenuto applicando *MixAll* assume un valore pari a 0.84.

Anche in questo caso vengono riportati nella figura 4.23 l'istogramma e la densità del *Rand Index* utilizzando *MixAll* per l'applicazione del *model-based clustering* nel contesto in cui i cluster reali presentano differenti dimensioni e orientamenti.

Per la maggior parte dei dataset simulati si ottengono buoni raggruppamenti, dato che il valore medio dei *Rand Index* si aggira attorno a 0.82. Si può però notare che il valore minimo assunto da questo indice, utilizzato per valutare i raggruppamenti, è inferiore rispetto a quanto ottenuto utilizzando il pacchetto *mclust*.

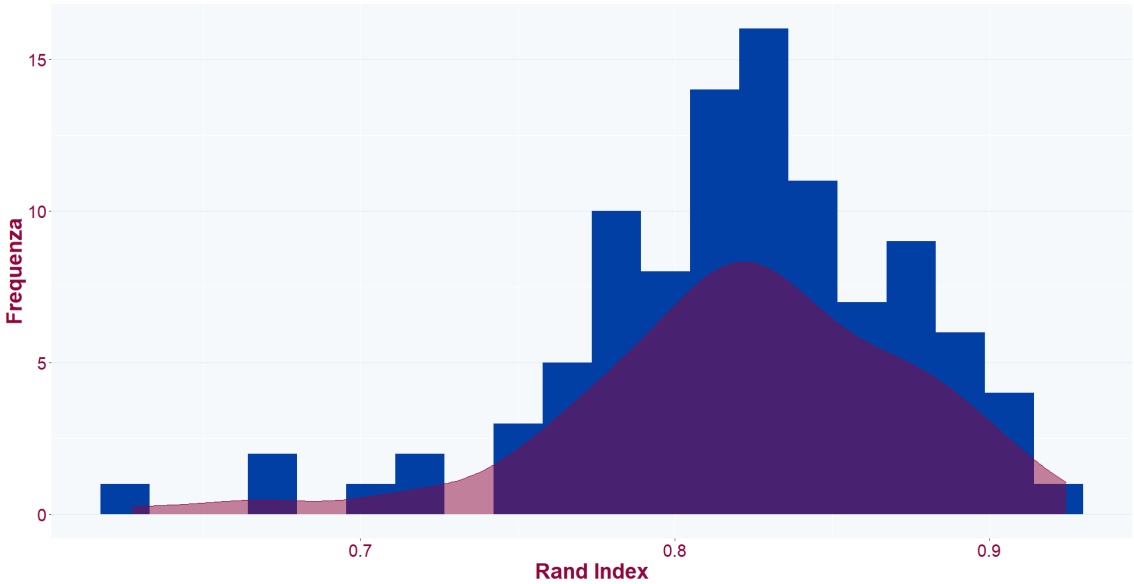


Figura 4.23: Istogramma e densità del *Rand Index* per le classificazioni ottenute con *mclust* a partire dai dataset simulati i cui cluster hanno dimensione e orientamento differenti

Come si può vedere dalla tabella 4.12, a differenza di quanto emerso negli scenari legati alla presenza di valori mancanti, i valori medi del *Rand Index* ottenuti utilizzando i due pacchetti sono molto simili.

	Media del *Rand Index*
mclust	0.87
MixAll	0.82

Tabella 4.12: Rand Index per le partizioni ottenute negli scenari legati a differenti caratteristiche geometriche dei cluster utilizzando *mclust* e *MixAll*

Modello selezionato	**Frequenza**
gaussian_pk_sjk	38
gaussian_pk_sj	28
gaussian_pk_sk	14
gaussian_pk_s	12
gaussian_p_sjk	3
gaussian_p_sk	3
gaussian_p_sj	1
gaussian_p_s	1

Tabella 4.13: Frequenza dei modelli selezionati da *MixAll* per i dataset simulati i cui cluster hanno dimensione e orientamento differenti

Tra i modelli implementati da *MixAll* quello maggiormente selezionato risulta essere il modello *gaussian_pk_sjk* che prevede che le proporzioni della mistura, le deviazioni standard nei cluster e tra i cluster siano differenti. Nel caso in esame, tuttavia, le deviazioni standard all'interno dei gruppi sono uguali.

Si può notare però come vi sia una quota abbastanza importante di dataset simulati per cui il modello selezionato da *MixAll* risulta essere il modello *gaussian_pk_sj* in cui le proporzioni della mistura e le deviazioni standard tra i cluster variano mentre rimangono uguali le deviazioni standard all'interno dei cluster. In 28 casi viene quindi selezionato il modello corretto.

Viene infine riportata nella figura 4.24 una rappresentazione dei boxplot dei tassi di errata classificazione ottenuti utilizzando *mclust* e *MixAll* per i 100 dataset simulati in cui i cluster hanno dimensione e orientamento differenti. Si evidenzia una differenza importante tra i tassi di errata classificazione ottenuti utilizzando i due pacchetti. Mediante *mclust*, per la maggior parte dei dataset simulati, si riescono ad ottenere raggruppamenti migliori di quelli realizzati con *MixAll*.

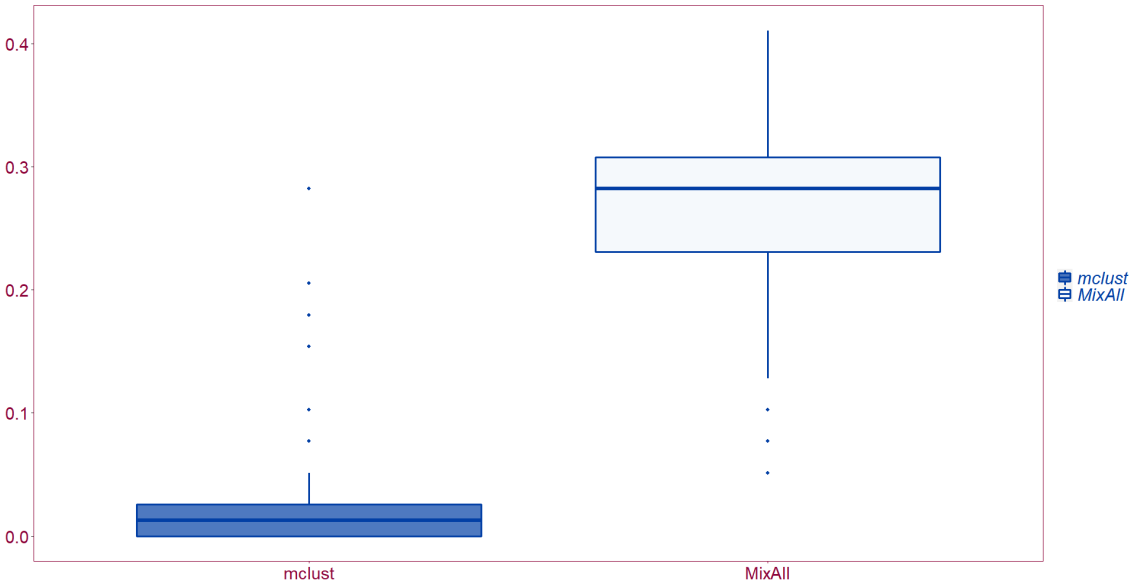


Figura 4.24: Boxplot dei tassi di errata classificazione ottenuti utilizzando *mclust* e *MixAll* per l'applicazione del *model-based clustering* ai dataset simulati i cui cluster hanno dimensione e orientamento differenti

4.3 Commenti

I risultati ottenuti in questo studio di simulazione mostrano una migliore capacità dei modelli stimati da *mclust* di cogliere la reale suddivisione delle osservazioni e il numero delle componenti della mistura. È da sottolineare infatti che per tutti e 100 i dataset simulati con la presenza del 5% di dati mancanti, *MixAll* selezioni, tramite il valore del *BIC*, un modello di mistura con $k = 5$ componenti. Questo avviene anche negli altri due scenari legati alla presenza di dati mancanti ma, per alcuni dataset, *MixAll* seleziona un modello di mistura con $k = 4$ componenti.

Considerando i risultati ottenuti è quindi possibile affermare che, in generale, nel contesto di presenza di dati mancanti, i modelli di mistura implementati con *MixAll* sovrastimino il numero di cluster. A differenza di questo, i modelli selezionati da *mclust* riescono maggiormente a cogliere la struttura sottostante i dati. Questo aspetto viene riscontrato anche considerando i dataset simulati i cui cluster hanno dimensione e orientamento differenti.

Si sottolinea tuttavia che, nonostante la quasi totalità dei modelli stimati da *MixAll* abbia un numero di componenti superiore a quanto effettivamente presente nei dataset simulati, da quanto emerso nello scenario in cui i cluster presentano orientamento e volume differenti, si può affermare che *MixAll* riesca a cogliere in modo migliore le caratteristiche geometriche dei cluster rispetto a *mclust*.

Infine, è importante evidenziare come vi sia un maggior costo computazionale utilizzando *MixAll* in particolar modo nei contesti di presenza di valori mancanti. A differenza di quanto avviene con *mclust*, *MixAll* provvede con la funzione *clusterDiagGaussian* al processo di imputazione di dati mancanti.

Conclusione

In questa tesi è stata esposta una delle procedure dell'analisi dei gruppi, la *cluster analysis* basata su modello, la quale è stata applicata sia ad un reale dataset di natura economica, sia a dataset simulati allo scopo di confrontare i modelli di mistura implementati tramite i due pacchetti R *mclust* e *MixAll*.

Dopo aver presentato le variabili macroeconomiche del dataset reale, relativo alla maggior parte dei Paesi europei, sono state rimarcate alcune caratteristiche del dataset stesso legate alla presenza di dati mancanti, concentrate soprattutto su alcune variabili e determinati Paesi, e alla forte correlazione tra alcune delle variabili considerate, quali PIL e le sue componenti.

L'obiettivo di applicare l'analisi dei gruppi, nello specifico il *model-based clustering*, al dataset economico è stato perseguito mediante l'utilizzo di due pacchetti R, *mclust* e *MixAll*. Con entrambi i pacchetti è possibile stimare modelli di mistura Gaussiani. Il problema della presenza di valori mancanti viene trattata direttamente tramite l'utilizzo di *MixAll* a differenza di *mclust* che richiede una preventiva imputazione dei valori mancanti. Inoltre, la problematica legata alla forte correlazione tra variabili all'interno del dataset ha reso necessario percorrere due strade: eliminare una delle variabili che riportano una perfetta correlazione, ossia il PIL, e applicare l'analisi delle componenti principali prima di applicare il *model-based clustering*. Si è visto come i risultati ottenuti nei due casi, a livello di classificazione dei Paesi europei, siano molto simili. Tuttavia, l'utilizzo delle componenti principali ha contribuito a diminuire il costo computazionale richiesto per la stima ed il confronto dei modelli di mistura.

È da sottolineare che il dataset economico considera solamente Paesi appartenenti al continente europeo. È ragionevole ritenere che i Paesi considerati, o la maggior parte di essi, presentino già delle similitudini, anche a livello economico. Verosimilmente è possibile che estendendo l'analisi alla totalità degli Stati mondiali, i Paesi europei, fatta qualche

eccezione, vengano classificati in un unico cluster. Questo può in parte giustificare il fatto che le unità statistiche considerate nel dataset economico non formino dei cluster ben separati, sia mediante l'utilizzo di *mclust* sia tramite l'uso di *MixAll*.

Non avendo una reale classificazione dei Paesi europei considerati all'interno del dataset economico, non è stato possibile valutare puntualmente le classificazioni ottenute. Per questo motivo è stato condotto uno studio di simulazione cercando di prendere in considerazione sia la presenza di valori mancanti all'interno del dataset sia una situazione in cui le caratteristiche geometriche delle componenti della mistura fossero più complesse. In particolare, sono stati considerati tre scenari legati alla presenza di valori mancanti con differenti percentuali e una diversa suddivisione di essi tra le variabili generate. L'ulteriore scenario analizzato prevede cluster con differenti dimensioni e orientamenti.

Nello studio di simulazione è emersa una migliore capacità del pacchetto *mclust* a cogliere la struttura dei dati, e quindi il numero reale di cluster presenti all'interno dei dataset simulati, sia in una situazione caratterizzata da una maggiore presenza di valori mancanti, sia in una circostanza in cui il pattern dei dati mancanti, ma anche le caratteristiche geometriche dei cluster, non definiscono la situazione più semplice. È comunque necessario considerare il fatto che nello studio di simulazione condotto sono state generate solamente due variabili e che, negli scenari legati alla presenza di dati mancanti, i cluster, a differenza di quanto riscontrato nell'applicazione al dataset economico, sono ben distinti.

Ad ogni modo, la grande differenza dei cluster formati nello studio di simulazione dai due pacchetti utilizzati si riflette anche nei risultati ottenuti sul dataset a cui è stata applicata l'analisi dei gruppi basata su modello con lo scopo di indagare sulla struttura sottostante i Paesi europei in base alle variabili macroeconomiche.

Alla luce di quanto emerso dallo studio di simulazione si potrebbe quindi ritenere migliore la classificazione ottenuta dall'applicazione del *model-based clustering* al dataset economico tramite l'utilizzo di *mclust*. Si può infatti notare come, in questo caso, le principali potenze europee vengano classificate in uno stesso gruppo, così come la maggior parte dei Paesi dell'est Europa e dei Paesi balcanici.

Appendice

Codice R

Viene di seguito riportata la funzione creata in R per la generazione dei dati provenienti da una distribuzione di mistura Gaussiana:

```
  gaussian_mixture = function(p,n,mu,sigma){
  cluster = sample(c(1:length(p)),size = n,replace = T,prob = p) %>%
    sort()
  dim_cluster = table(cluster)
  data = vector(mode = "list",length = length(p))
  for(i in 1:length(p)){
    data.mixture = rmvnorm(dim_cluster[i],mu[[i]],sigma[[i]])
    data[[i]] = data.mixture
  }
  data = list.rbind(data) %>%
    cbind(cluster)
}
```

La funzione richiede come input le proporzioni della mistura, il numero di osservazioni da generare, una lista contenente i vettori delle medie delle componenti e una lista con le matrici di varianze e covarianze per le diverse componenti.

Vengono inizialmente generati i gruppi di appartenenza per le n unità statistiche in base alle proporzioni della mistura. Queste ultime definiscono quindi anche il peso delle gaussiane che compongono la mistura. Successivamente, per ogni gruppo, viene generata una normale multivariata la cui dimensione è data dal numero di osservazioni che appartengono all'i-esimo gruppo, il vettore delle medie dall'i-esimo elemento della lista contenente le medie delle componenti e la matrice di varianza e covarianza dall'i-esimo elemento della lista che include le matrici di varianze e covarianze delle componenti, dove i assume valori da 1 fino alla lunghezza del vettore che definisce le proporzioni della mistura. La lunghezza del vettore delle proporzioni della mistura definisce il numero di componenti o, analogamente, il numero di gruppi.

Il comando utilizzato per la generazione delle normali multivariate è *rmvnorm* del pacchetto *mvtnorm* (Alan Genz e altri, 2021). Come ultimo passaggio viene creato un unico dataset contenente le i distribuzioni combinando per riga le gaussiane generate.

Bibliografia

Blashfield, R. K., Aldenderfer, M. S. (1988), *The Methods and Problems of Cluster Analysis. In: Nesselroade, J.R., Cattell, R.B. Handbook of Multivariate Experimental Psychology. Perspectives on Individual Differences*, Springer, Boston.

Celeux G., Govaert G. (1993), *Gaussian parsimonious clustering models*, RR-2028, INRIA, pp. 3-4: pp. 5-7.

Chiquet J., Rigaill G., Sundqvist M. (2022), *aricode: Efficient Computations of Standard Clustering Comparison Measures. R package version 1.0.1.*

Fraley C., Raftery E.A. (1998), *How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis*, The Computer Journal, Vol. 41, No. 8, Department of Statistics, University of Washington, USA, pp. 579-582.

Fraley C., Raftery E. A. (2002), *Model-Based Clustering, Discriminant Analysis, and Density Estimation*, Journal of the American Statistical Association, Vol. 97, No. 458, pp. 611–615.

Fraley C., Raftery E. A., Murphy T. B., Scrucca L. (2019), *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*, University of Washington.

Genz A., Bretz F., Miwa T., Mi X., Leisch F., Scheipl F., Hothorn T.(2021), *mvtnorm: Multivariate Normal and t Distributions. R package version 1.1-3.*

Grün B. (2018), *Model-based Clustering. In: Fruhwirth-Schnatter S., Celeux G, Robert P.C. Handbook of Mixture Analysis*, Chapman and Hall/CRC, pp. 19-20.

Halkidi M., Batistakis Y., Vazirgiannis M. (2001), *On Clustering Validation Techniques*, Journal of Intelligent Information Systems 17, pp. 107–145.

Hastie T., Tibshirani R., Triedman J. (2008), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, pp. 233-235: pp. 234-235.

Iovleff S. (2019), *MixAll: Clustering and Classification using Model-Based Mixture Models*, R package version 1.5.1.

Iovleff S. (2019), *MixAll: Clustering Mixed data with Missing Values*, University of Lille, INRIA.

Johnson R. A., Dean W. W. (2007), *Applied Multivariate Statistical Analysis*, Pearson Education, New Jersey, pp. 703-705.

MacQueen, J. (1967), *Some Methods for Classification and Analysis of Multivariate Observations* in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281-297.

R Core Team (2021), *R: A language and environment for statistical computing.*, R Foundation for Statistical Computing, Vienna, Austria.

Rockel T. (2022), *missMethods: Methods for Missing Data.*, R package version 0.4.0.

Schafer J. L.(2022), *mix: Estimation/Multiple Imputation for Mixed Categorical and Continuous Data*, R package version 1.0-11.

Scrucca L., Fop M., Murphy T. B., Raftery A. E. (2016), *mclust 5: clustering, classification and density estimation using Gaussian finite mixture models*, The R Journal , Vol. 8, pp. 289-317.

Tryon, R. C. (1939), *Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*, Ann Arbor, Mich: Edwards brothers, inc., lithoprinters and publishers.

Sitografia

Eurostat, https://ec.europa.eu/eurostat/data/database.

Eurostat, https://ec.europa.eu/eurostat/cache/metadata/en/nama10_esms.htm.

Istat (2011), https://www.istat.it/it/files/2011/04/glossario1.pdf.

Istat (2015), https://www.istat.it/it/files/2015/05/Glossario1.pdf.

Morbieu S.(2018), *Generate datasets to understand some clustering algorithms behavior*, https://smorbieu.gitlab.io/generate-datasets-to-understand-some-clustering-algorithms-behavior/

Office for Nation Statistics, https://www.ons.gov.uk/employmentandlabourmarket.

Ringraziamenti

Ritengo doveroso, alla fine di questo elaborato, ringraziare alcune persone senza le quali la realizzazione di questa tesi non sarebbe stata possibile.

Un ringraziamento particolare è rivolto alla mia relatrice, Prof.ssa Manuela Cattelan, che ha sempre dimostrato una grande disponibilità e professionalità. La ringrazio per avermi trasmetto, durante le Sue lezioni, la passione per la Statistica.

Ringrazio di cuore i miei genitori che mi sostengono e mi incoraggiano ogni giorno nelle mie scelte. Ringrazio anche mio fratello Giacomo per il suo prezioso aiuto durante questo mio percorso universitario.

Più in generale ringrazio tutta la mia famiglia che ha saputo comprendere la mia frequente assenza dovuta allo studio.

Vorrei poi ringraziare le mie care amiche Camilla, Eleonora e Beatrice, per essere come un braccio destro ed esserci sempre nei momenti di bisogno.

Infine ringrazio i miei colleghi universitari con cui ho condiviso questo importante percorso.