Università degli Studi di Napoli
Federico II

La divergenza di Jensen-Shannon nell'algoritmo di clustering dinamico per dati descritti da distribuzioni multivariate

Francesca Condino

Tesi di Dottorato di Ricerca in
Statistica

XXII Ciclo

La divergenza di Jensen-Shannon nell'algoritmo di clustering dinamico per dati descritti da distribuzioni multivariate

Napoli, 30 novembre 2009

A chi, consapevolmente o meno,
mi ha più incoraggiato:
a mio marito e al nostro piccolo Valentino

Ringraziamenti

Giunta al termine di questo percorso, desidero ringraziare coloro i quali, in un modo o nell'altro, hanno fatto sì che questo progetto si concretizzasse.

I miei più sentiti ringraziamenti vanno alla Prof.ssa Rosanna Verde, non solo per avermi indirizzato e guidato in questo percorso di studio e di ricerca, ma anche e soprattutto per quella costante disponibilità e attenzione, peculiare solo di chi è caratterialmente empatico, che ben lungi dall'essere meramente assolvimento di doveri istituzionali, si è tradotta per me in sostegno concreto.

Insieme a lei, il Dott. Antonio Irpino, grazie alla sua supervisione e agli importanti spunti di riflessione, è stato essenziale per lo sviluppo dei temi affrontati. Il suo modo diretto e pratico mi ha spesso messo di fronte alle questioni irrisolte, spronandomi affinchè cercassi soluzioni appropriate.

Un ringraziamento particolare al Prof. Carlo Lauro, coordinatore del dottorato, che è stato fondamentale per l'inizio e la prosecuzione di questo cammino. Senza di lui adesso non starei scrivendo queste righe!

Grazie al Prof. Filippo Domma, non soltanto per aver contribuito, con i suoi preziosi suggerimenti, a rendere migliore il presente lavoro, ma anche per essere, da diversi anni ormai, un punto di riferimento per la mia crescita culturale e personale. L'impegno e la passione che

mette nel suo lavoro sono per me valori fondamentali, che solo un *maestro* sa trasmettere.

Se ho avuto la possibilità di intraprendere questo triennio è anche grazie al Prof. Aldo Quattrone, responsabile della mia attività presso l'Istituto di Scienze Neurologiche del CNR. Ringrazio lui e i tanti ricercatori con i quali ho il piacere di confrontarmi ogni giorno.

Un grazie ai miei amici più cari, i quali oltre a sostenermi hanno condiviso con me le ansie e il peso della fatica, supportandomi oltre che sopportandomi.

Ringrazio i miei colleghi di dottorato, che mi hanno sempre fatto sentire a casa e tutti quelli che, almeno una volta, mi hanno detto *'dai che ce la fai!'*: alcuni di loro hanno avuto un tempismo davvero perfetto, incoraggiandomi quando ne avevo più bisogno.

Ovviamente un grazie pieno e sentito alla mia famiglia, per il sostegno non solo morale, ma anche fattivo nella gestione funambolica della quotidianità. Senza il loro affetto e le loro attenzioni non avrei avuto la serenità necessaria per arrivare fin qui.

Indice

Introduzione

Il presente lavoro di tesi si inserisce nel contesto del trattamento di dati complessi e più nello specifico, del problema di classificazione di dati non puntuali. Questi dati, si caratterizzano per la molteplicità dei valori assunti in corrispondenza di ciascun oggetto considerato e come tali necessitano non solo di ridefinizioni circa le modalità di raccolta e di descrizione degli stessi, ma anche di nuove metodologie di analisi costruite ad hoc.

In tale ambito, alcuni autori hanno proposto di recente nuove tecniche di classificazione, basate su misure di distanza applicabili ai descrittori a disposizione. Contestualmente, la necessità di considerare anche le relazioni tra i suddetti descrittori ha spinto verso nuove proposte, incentrate sullo studio di misure di distanza tra gli oggetti attraverso le quali si possa considerare anche la presenza di dipendenza nei dati.

Nel caso specifico di dati descritti da distribuzioni, l'utilizzo della funzione copula, nasce come naturale strumento per la descrizione di tali dati. La modellizzazione attraverso la funzione copula consente infatti di considerare separatamente i descrittori, ovvero le distribuzioni marginali, e la struttura di dipendenza, e quindi le informazioni aggiuntive circa le relazioni tra i descrittori.

Simultaneamente alla proposta di modellizzazione attraverso lo strumento della copula, si propone di utilizzare la divergenza di Jensen-

Shannon, per valutare la discrepanza esistente tra le distribuzioni considerate e dunque il grado di dissimilarità tra gli oggetti. Tale misura, essendo basata sulle funzioni di probabilità o densità multivariate, consente di tenere in conto proprio la dipendenza tra i dati, fermo restando l'ipotesi di indipendenza tra gli oggetti considerati. Inoltre gode di proprietà, verificate nel contesto del presente lavoro, che la rendono idonea ad essere impiegata nell'ambito degli algoritmi di classificazione dinamica.

Dopo una introduzione sulla particolare tipologia dei dati trattati, ovvero i dati non puntuali, il secondo capitolo è dedicato alla descrizione della metodologia inerente le funzioni copula. Oltre alle definizioni principali, sono affrontate le tematiche delle misure per lo studio della dipendenza e della stima dei parametri. Inoltre sono descritti i modelli più diffusi e utilizzati e le relative proprietà.

Il terzo capitolo è invece dedicato alla descrizione e alla proposta di misure di dissimilarità o distanza applicabili a dati descritti da distribuzioni multivariate, sempre con particolare attenzione alla problematica della valutazione della struttura di dipendenza tra essi. Molte di queste misure risultano strettamente connesse a concetti propri della cosiddetta Teoria dell'Informazione, della quale sono brevemente richiamate alcune nozioni fondamentali.

I due principali approcci di classificazione per dati non convenzionali, quello di classificazione dinamica (DCA) e l'algoritmo basato su tabelle di distanza, sono introdotti nel quarto capitolo. In questa sede viene rivisitato l'algoritmo di classificazione dinamica al fine di proporre una procedura atta alla classificazione di dati a cui siano associati descrittori multivariati. In questo contesto alcune soluzioni originali relative alla divergenza di Jensen-Shannon giustificano e avvalorano l'utilizzo di tale distanza in questo ambito. L'approccio attraverso la funzione copula e i risultati ottenuti nel contesto del DCA concludono il capitolo.

Infine, a supporto e a completamento della parte di descrizione della

metodologia proposta, sono presentati i risultati ottenuti applicando l'algoritmo su dati simulati e reali.

Capitolo 1

I dati non puntuali

1.1 Introduzione

Molti ambiti di studio sono oggi caratterizzati da una sempre crescente complessità di informazioni, che si rendono fruibili generalmente attraverso architetture complesse di archiviazione e gestione. Questa ingente mole di dati ha reso necessario negli ultimi tempi, lo sviluppo di nuovi approcci finalizzati alla ristrutturazione dell'informazione stessa sotto nuove forme.

Una prima formulazione in tal senso è stata proposta da Diday [1987], il quale introduce il concetto di *oggetto simbolico* e propone nuovi strumenti di analisi. Interesse prioritario dunque è quello di sintetizzare quanto più possibile le informazioni in termini di nuovi concetti al fine di estrarre da essi nuove conoscenze.

Contestualmente all'introduzione di questi nuovi *oggetti* sono fiorite una serie di tecniche di analisi statistica che hanno il vantaggio di poter analizzare dati di un livello superiore, consentendo di trattare direttamente con concetti piuttosto che con dati elementari.

Questo capitolo è dedicato alla definizione e alla descrizione di dati

non convenzionali, e più in particolare di dati descritti da distribuzioni. Dopo alcune definizioni preliminari e alcuni esempi relativi ai descrittori univariati, si tratterà il tema della dipendenza e quindi dei descrittori multivariati. Alcune considerazioni sui vantaggi nell'utilizzo di questa particolare tipologia di dati concluderanno il capitolo.

1.2 Aspetti storici e filosofici

Sebbene l'idea di considerare formalmente nell'ambito delle scienze matematiche e statistiche oggetti di livello superiore sia abbastanza recente, tale idea ha radici antiche. Già Aristotele nel IV secolo A.C. distingue nettamente tra quello che è rappresentato dal singolo individuo, quale unità elementare, e il concetto o la categoria alla quale il singolo è ascrivibile. Introduce cioè quello che, da alcuni autori, più di recente, viene detto 'paradigma dei due livelli' [Bock and Diday, 2000].

Tra le opere di Aristotele sulla Logica raccolte nell'Organon, vi sono in particolare due libri, *'Sulle Categorie'* e *'Sull'interpretazione'* nei quali viene descritto in dettaglio cosa debba intendersi per *Concetto*.

Ogni concetto viene definito da un certo numero di caratteristiche; tanto maggiori saranno le proprietà che caratterizzano un concetto (intensione) tanto minori saranno gli oggetti (estensione) a cui quel dato concetto si riferirà. Intensione ed estensione sono tra loro inversamente proporzionali, ossia aumentando il numero di caratteristiche considerate diminuisce il numero di individui a cui il concetto si riferisce e, di contro, diminuendo le caratteristiche di un concetto aumenta il numero di individui a cui questo si riferisce. Dunque i concetti saranno ordinabili gerarchicamente in modo che si avranno ai livelli più bassi i concetti con massima intensione e minima estensione e a quelli più alti i concetti con massima estensione e minima intensione.

Il dualismo tra l'idea dell' *Intensione* e quella dell'*Estensione* riferibile ai concetti è stato successivamente ripreso nel XVII secolo da Arnault and Nicole [1662].

Nel campo delle scienze cognitive invece l'idea del concetto si affianca a quella del prototipo. Infatti Rosh [1978] sostiene che i concetti devono essere rappresentati da classi che tendono ad essere definite in termini di prototipi. Gli stessi saranno tali se conterranno gli attributi maggiormente rappresentativi degli individui appartenenti alla classe.

Cercare di riportare questo tipo di logica nell'ambito del *Data Mining* significa ridefinire i concetti a cui ci si vuole riferire e formalizzare gli attributi e le descrizioni per ogni singolo concetto. Anche la rappresentazione del dato stesso non potrà più essere basata su database di tipo *individual-oriented*, ma piuttosto su matrici *object-oriented*.

1.3 I Data-set non convenzionali

La peculiarità principale di questo approccio risiede innanzitutto nella particolare ristrutturazione delle informazioni che trova la sua rappresentazione più adatta in una nuova forma di data-set, in cui ad ogni colonna corrisponde una variabile non convenzionale, meglio definita nel prosieguo, mentre ad ogni riga corrisponde un oggetto o un concetto e contiene i descrittori utilizzati per la rappresentazione di ciascun individuo. Quindi, a differenza delle strutture standard, nelle quali ogni cella contiene il valore assunto dalla singola variabile in corrispondenza di un singolo individuo, la cella di questo tipo di data-set può contenere una molteplicità di informazioni, di natura diversa a seconda delle variabili e degli oggetti considerati.

Al fine di chiarire alcuni aspetti del problema, si consideri preliminarmente il seguente esempio. Si supponga di dover descrivere una varietà di fiore prodotto da una certa pianta. Si consideri un'espressione del tipo *'questo fiore può assumere colorazioni dal bianco al giallo*

e può avere un diametro della corolla da 2 a 5 cm, ma se il colore è bianco allora la corolla non sarà di diamentro superiore a 3 cm'. Questo tipo di asserzione non può essere rappresentata direttamente in un database classico, dato che le celle al suo interno dovranno contenere necessariamente un solo valore. Una soluzione a tale problema è l'utilizzo di variabili definite in maniera non convenzionale, ad esempio nell'ottica dell'approccio simbolico, che verranno associate direttamente ad un concetto di livello superiore, ovvero al concetto di fiore, e non alla singola unità statistica (ovvero il singolo fiore nato dalla specifica pianta). Di conseguenza si potranno definire due variabili, *Colore* e *Diametro*, le quali assumeranno valori non puntuali, quali ad esempio liste di categorie o intervalli di $\mathbb{R}$, nella seguente forma:

'[Colore={bianco, giallo}], [Diametro=[2,5]] e [se {Colore=bianco} allora {Diametro $\leqslant$ 3}'

La formalizzazione appena proposta viene detta descrizione associata al concetto di 'fiore'.

Da qui si intuisce come sia necessario distinguere tra le diverse tipologie di variabili, a seconda delle realizzazioni di ciacuna e formalizzare le possibili relazioni tra esse esistenti. Diverse sono le proposte di variabili (definite quindi in senso ampio) atte a descrivere oggetti complessi e di seguito verranno brevemente richiamate le principali definizioni.

Dato che la proposta metodologica oggetto del presente lavoro di tesi nasce dall'esigenza di comparare e classificare oggetti descritti da distribuzioni, si tratterà con maggiore dettaglio il caso di descrittori intesi come distribuzioni di probabilità e delle relazioni tra essi intercorrenti.

1.4 Descrittori per dati non puntuali

Si consideri un insieme di m unità statistiche $u_s \in \Omega$ con $s = 1, \ldots, m$ e sia $E = \{\omega_1, \ldots, \omega_n\}$ l'insieme formato dagli n oggetti ottenuti considerando una qualche descrizione delle unità elementari, in modo che il generico elemento ω_i rappresenterà un insieme di elementi di Ω che soddisfano la descrizione suddetta.

Sia Y una variabile casuale definita per tutti gli elementi di Ω, con dominio $\mathcal{Y}$. Se la variabile Y è rilevata per l'intera categoria ω_i, allora assumerà valori $Y(\omega_i) = \xi_i$, che non saranno generalmente singoli. Quindi, a seconda dei valori ξ_i potranno essere identificate diverse tipologie di descrittori, alcune delle quali verranno brevemente illustrate di seguito.

Variabili di tipo Set-valued
Una variabile Y avente dominio $\mathcal{Y}$, definita per tutti gli elementi che compongono l'insieme E, è detta di tipo *Set-valued* se i valori $Y(\omega_i)$ da essa assunti appartengono all'insieme $\{U|U \subseteq \mathcal{Y}\}$ costituito da tutti i sottoinsiemi non vuoti di $\mathcal{Y}$. [1]

Alla classe delle variabili *Set-valued* appartengono le *Variabili ad Intervallo* e le *Variabili Multi-valore*:

- *Variabili ad Intervallo*
  Una variabile set-valued Y è detta *Variabile ad Intervallo* se per ogni valore $\omega_i \in E$ il sottoinsieme $U := Y(\omega_i)$ è un intervallo di $\mathbb{R}$ oppure un intervallo con riferimento ad un dato ordine su $\mathcal{Y} : Y(\omega_i) = [\alpha, \beta]$ per qualche $\alpha, \beta \in \mathcal{Y}$ (rispettivamente con $\alpha \leqslant \beta$ e $\alpha \preceq \beta$).

[1] Il caso di dati standard, in cui ad ogni singola unità statistica si associa un unico valore, si può ottenere come caso particolare dalla definizione di variabile *Set-valued* quando si richiede che $Y(\omega_i)$ abbia cardinalità unitaria per tutti i valori $\omega_i \in E$.

- *Variabili Multi-valore*
  Una variabile Y definita per tutti gli elementi di un insieme E è detta *Multi-valore* se i valori da essa assunti sono sottoinsiemi finiti del sottostante dominio $\mathcal{Y} : |Y(\omega_i)| < \infty$.

  In particolare, una variabile multi-valore Y è detta *categoriale* se assume valori in un insieme finito di categorie, mentre è detta *quantitativa* se i valori $Y(\omega_i)$ sono insiemi finiti di numeri reali: $Y(\omega_i) \subset \mathbb{R}$

A differenza delle variabili ad intervallo, le variabili multi-valore assumono perciò valori che non sono necessariamente ordinabili e contigui tra loro.

Variabili Modali
Una variabile Y è detta *Variabile Modale* con dominio $\mathcal{Y}$ se per ogni oggetto $\omega_i \in E$ si ha $Y(\omega_i) \subseteq \mathcal{Y}$ e contestualmente, ad ogni valore di $y \in Y(\omega_i)$, viene associata una frequenza, una probabilità o un peso $\pi(y)$.

Più in particolare, se la variabile Y ha come dominio di definizione un dominio sovrapponibile a quello di una variabile multi-valore si avranno le cosiddette variabili *modali multi-valore*, mentre se il suo dominio è quello di una variabile ad intervallo, allora si avranno le cosiddette *variabili a istogramma*.

In senso più ampio, la definizione di variabili modali dunque comprende tutte quelle variabili i cui valori sono distribuzioni di probabilità o di densità e quindi anche quelli che nel prosieguo verranno indicati come descrittori univariati.

1.4.1 Descrittori univariati: le distribuzioni marginali

Si considerino $\omega_1, \ldots, \omega_n$ oggetti costituenti l'insieme E. Sia T un data-set contenente n righe e p colonne. Si supponga che l'*i-esima* riga $(i = 1, \ldots, n)$ corrisponda ad un oggetto ω_i e si consideri un vettore di variabili casuali $X^{(1)}, \ldots, X^{(p)}$ di interesse per la descrizione di tale oggetto.

Visto che ω_i generalmente rappresenta un insieme di individui elementari che soddisfano alcune caratteristiche prestabilite, ci si riferirà alle realizzazioni delle variabili suddette in corrispondenza degli individui ascrivibili al concetto ω_i. In particolare si denoterà on x_s^j il valore assunto dalla generica variabile casuale $X^{(j)}$ in corrispondenza dell'individuo *s-esimo*. Dunque, all'oggetto ω_i corrisponderanno tante realizzazioni di $X^{(j)}$ quanti sono gli individui compresi nella classe/categoria ω_i.

Se si suppone che la variabile casuale segua una legge di probabilità nota, descritta da un certo modello parametrico, allora si potrà associare all'$i - esimo$ oggetto una funzione di densità o probabilità e dunque la cella di T, individuata per l'*i-esima* riga e per la *j-esima* colonna, potrà contenere una funzione di ripartizione $F_i^{(j)}$. I descrittori del singolo oggetto saranno allora distribuzioni, note oppure stimate attraverso le realizzazioni campionarie di $X^{(j)}$.

Il data-set considerato assumerà dunque la struttura indicata in tabella 1.1. Ovviamente, insieme alle funzioni di ripartizione, saranno note anche le funzioni di probabilità o densità associate all'*i-esimo* oggetto $f_i^{(1)}, \ldots, f_i^{(p)}$.

Oggetto	$Y^{(1)}$	...	$Y^{(j)}$	...	$Y^{(p)}$
ω_1	$F_1^{(1)}$	...	$F_1^{(j)}$	...	$F_1^{(p)}$
ω_2	$F_2^{(1)}$	...	$F_2^{(j)}$	...	$F_2^{(p)}$
...	...		...		...
ω_i	$F_i^{(1)}$	...	$F_i^{(j)}$	...	$F_i^{(p)}$
...	...		...		...
ω_n	$F_n^{(1)}$	...	$F_n^{(j)}$	...	$F_n^{(p)}$

Tabella 1.1: Oggetti descritti da distribuzioni

Alcuni esempi

Esempio 1[2]**.**
Si supponga di essere interessati alla distribuzione dei consumi di differenti tipi di combustibili, nei diversi stati USA e si supponga che i consumi seguano una distribuzione normale. Allora, ad ogni singolo combustibile verrà associata una distribuzione normale con parametri che potranno essere stimati dai dati, ottenendo la seguente tabella:

ω	Tipo	$Y(\omega)$
ω_1	Petrolio	$Norm(76.7, 92.5^2)$
ω_2	Gas naturale	$Norm(45.9, 69.5^2)$
ω_3	Carbone	$Norm(47.0, 43.3^2)$
ω_4	Energia idroelettrica	$Norm(6.4, 14.3^2)$
ω_5	Energia Nucleare	$Norm(25.5, 20.4^2)$

Tabella 1.2: Il consumo di combustibili in USA

La tabella 1.2 fa riferimento ad una singola variabile $Y(\omega)$, ma si può supporre di essere interessati a più caratteristiche riguardanti gli

[2]Questo esempio è riportato in Billard and Diday [2007] nel contesto dei dati simbolici.

oggetti $\omega_1, \ldots, \omega_5$. Ad esempio, potrebbe essere utile valutare la distribuzione dei prezzi praticati nei diversi stati, per ogni singola fonte energetica. In questo caso avremmo due variabili, corrispondenti alle due distribuzioni marginali di consumi e costi.

Esempio 2.
Una fonte di dati strutturabili come i precedenti potrebbe essere ad esempio rappresentata dall'andamento nel tempo di titoli finanziari. Se si considera quale oggetto di interesse il singolo titolo e come descrittore un modello parametrico per la valutazione dei rendimenti nel tempo, allora si potrà ottenere una tabella simile alla precedente.

Va comunque sottolineato che per il trattamento di questa tipologia di dati è necessaria una particolare cautela in quanto viene meno l'ipotesi di indipendenza tra gli oggetti. L'andamento di un titolo infatti è spesso fortemente legato all'andamento degli altri titoli e dell'intero mercato finanziario.

Esempio 3.
Una fonte comune di dati non puntuali è rappresentata dalle applicazioni nell'ambito delle scienze biomediche. In tale contesto infatti è comune la prassi di classificare i soggetti secondo le patologie di cui risultano affetti e studiare le caratteristiche demografiche e cliniche degli stessi al fine di ottenere un profilo che possa in qualche modo descrivere la patologia stessa. Si supponga ad esempio di considerare le diverse patologie di origine neurologica e alcune variabili demografiche e cliniche quali ad esempio l'età anagrafica, l'età di esordio della malattia, le scale cliniche per la disabilità cognitiva. Se, per la descrizione di queste variabili casuali, è noto o può essere stimato un modello parametrico, allora le informazioni potranno essere contenute in un data-set avente la struttura precedentemente illustrata.

1.4.2 La dipendenza e i descrittori multivariati

Nel considerare il data-set T sorge il problema di verificare l'ipotesi di indipendenza tra i descrittori. Infatti non è generalmente verosimile che ciascuna variabile $Y^{(j)}$, considerata per la descrizione degli oggetti in questione, assuma determinati valori indipendentemente dalle realizzazioni delle altre variabili.

Verde and Irpino [2008] hanno proposto metodi di analisi, inerenti nuove misure di distanza per dati ad istogramma, che permettono di tenere in considerazione anche la dipendenza tra i descrittori, con particolare riferimento alla dipendenza di tipo lineare. La necessità di considerare strutture di dipendenza differenti e di estendere la metodologia a dati di diversa natura, ha portato a valutare approcci alternativi, che nel presente lavoro si traducono nello studio del legame specifico e particolare esistente tra i descrittori mediante l'utilizzo delle funzioni copula.

Nel contesto in cui ogni oggetto sia descritto da una serie di distribuzioni, valutare l'ipotesi di dipendenza tra i descrittori equivale formalmente a studiare la distribuzione multivariata delle variabili casuali sottostanti. Infatti, se tali variabili sono di natura continua, ciascuna funzione di ripartizione marginale $F_i^{(j)}$ può essere essa stessa considerata una variabile casuale, la cui distribuzione sarà di tipo uniforme definita nell'intervallo $[0, 1]$. Allora studiare la particolare struttura di dipendenza esistente tra le variabili casuali $X^{(1)}, \ldots, X^{(p)}$ equivale a studiare la distribuzione congiunta dei descrittori $Y^{(1)}, \ldots, Y^{(p)}$. A seconda del tipo di dipendenza esistente tra le variabili casuali $X^{(1)}, \ldots, X^{(p)}$, la forma funzionale che lega le corrispondenti funzioni di ripartizione assumerà una particolare struttura, identificando una data distribuzione di probabilità, che sarà perciò una distribuzione di distribuzioni. Quindi, conoscere la distribuzione di queste funzioni di ripartizione, equivale a conoscere la corrispondente distribuzione multivariata delle variabili $X^{(1)}, \ldots, X^{(p)}$. Questo risultato è quello che è

stato ampiamente formalizzato e sviluppato nella teoria delle funzioni copule, che saranno descritte dettagliatamente nel prossimo capitolo.

1.5 Vantaggi nell'utilizzo di dati non puntuali

Uno dei paradigmi statistici di base è quello della sintesi delle informazioni. In ogni ambito della teoria statistica uno degli obiettivi prioritari è quello di estrarre, dalla molteplicità delle informazioni, solo le caratteristiche maggiormante rilevanti al fine di predisporre valutazioni, confronti, decisioni o altro. Sebbene la sintesi si accompagni generalmente ad una certa perdita di informazioni è comunque consolidata l'idea che essa sia assolutamente necessaria per la descrizione di qualsivoglia realtà.

L'utilizzo di strutture di dati non puntuali ha in sè lo stesso paradigma. E' ovvio che considerare quale dato di partenza un dato che è già una sintesi, in quanto descrizione di un oggetto comprendente più unità singole, comporti nella maggior parte dei casi una perdita di informazione, ma tale svantaggio è certamente bilanciato dalla possibilità di utilizzare data-set molto più sintetici quali input per l'utilizzo di tecniche statistiche proprie.

La predisposizione di strumenti atti al trattamento di dati più complessi, quali ad esempio dati ad intervallo o dati multi-valore, consente di ritenere solo l'informazione già sintetizzata secondo tali forme e non anche l'informazione circa il singolo individuo. Quando poi è addirittura noto il processo che ha generato il dato singolo, ad esempio perchè è nota la distribuzione da cui esso è estratto, poter fare affidamento su tecniche per dati descritti dalle distribuzioni è certamente preferibile.

Si pensi poi al caso in cui per ciascuna unità statistica si hanno a disposizione più valori (ad esempio misure effettuate in istanti di

tempo diversi). Alcune procedure classiche di classificazione, dovendo gestire in input dati singoli, si baseranno su valori di sintesi, quali ad esempio le medie aritmetiche. Considerare l'intera distribuzione consente di tener presente non solo la centralità, ma anche altri aspetti, quale la dispersione o l'asimmetria presente nei dati.

Ecco perchè, accanto alla proposta di archiviazione dei dati in forme già strutturate è importante la predisposizione di tecniche per il trattamento degli stessi.

Capitolo 2

La funzione copula

2.1 Definizioni e principali proprietà

Lo studio della dipendenza tra due o più variabili casuali è uno degli ambiti di maggiore interesse della teoria statistica, il cui scopo, tra gli altri, è anche quello di valutare se e come il comportamento di una certa variabile possa essere messo in relazione al comportamento di una seconda variabile, di una terza e così via. A tale aspetto sono ascrivibili numerose metodologie dirette a studiare il significato, l'intensità, e il verso della dipendenza. In tale ambito è possibile inquadrare la funzione copula, in quanto, come si vedrà, essa risulta un utile strumento per la descrizione e lo studio delle relazioni di dipendenza tra più variabili casuali.

Le origini di questa metodologia sono relativamente recenti, anche se l'interesse verso la stessa è rapidamente cresciuto negli ultimi anni. Come si evince dall'etimologia del termine, la copula è una funzione che mette insieme, ossia associa distribuzioni marginali di due o più variabili ad una distribuzione multivariata. Alternativamente, la copula può anche essere vista come la distribuzione multivariata di variabili

casuali aventi distribuzione marginale di tipo uniforme nell'intervallo (0,1).

L'importanza di questo strumento risiede nella capacità di poter separare la struttura di dipendenza dal comportamento delle singole marginali, ossia, come si vedrà meglio più avanti, è possibile individuare due parti, una relativa alle singole variabili casuali, l'altra riferibile al solo legame che unisce le stesse. La copula perciò conterrà tutte le informazioni sulla natura della dipendenza tra le variabili casuali, indipendentemente dalle espressioni delle distribuzioni marginali. Un secondo motivo di interesse risiede nella possibilità di considerare la funzione copula quale punto di partenza per la costruzione di nuove distribuzioni multivariate, avendo a disposizione le espressioni di due distribuzioni marginali.

Nel prosieguo verranno innanzitutto fornite le definizioni di base e successivamente i principali risultati.

2.1.1 Definizione

La copula è una distribuzione congiunta, definita sul cubo unitario n-dimensionale $\boldsymbol{I}^n = [0,1]^n$, tale che ogni distribuzione marginale è di tipo uniforme nell'intervallo $[0,1]$.
Formalmente se

$$C : [0,1]^n \to [0,1]$$

è una copula n-dimensionale (per brevità n-copula), allora:

- $C(\mathbf{u}) = 0$ se il vettore $\mathbf{u} \in [0,1]^n$ ha almeno una componente pari a 0;
- $C(\mathbf{u}) = u_i$ se il vettore $\mathbf{u} \in [0,1]^n$ ha tutte le componenti pari ad 1 eccetto l'i-esima, la quale è pari ad u_i;

- per ogni $\mathbf{a}$ e $\mathbf{b}$ in $\mathbf{I}^n$ tali che $\mathbf{a} \leqslant \mathbf{b}$, ovvero tali che $a_k \leqslant b_k \forall k$

$$V_C([\mathbf{a}, \mathbf{b}]) \geqslant 0$$

dove

$$V_C([\boldsymbol{a}, \boldsymbol{b}]) = \boldsymbol{\Delta}_{\boldsymbol{a}}^{\boldsymbol{b}} C(\boldsymbol{t}) = \Delta_{a_n}^{b_n} \Delta_{a_{n-1}}^{b_{n-1}} \dots \Delta_{a_1}^{b_1} C(\boldsymbol{t})$$

è il cosiddetto *C-volume*, calcolato come la differenza di ordine n di C sul rettangolo n-dimensionale $[\boldsymbol{a}, \boldsymbol{b}]$.[1]

2.1.2 Il teorema di Sklar

Il teorema centrale nella teoria delle funzioni copula è il teorema di Sklar [1959] per la prima volta apparso in relazione alla teoria delle funzioni di ripartizione multivariate. Esso definisce il ruolo della funzione copula quale anello di congiunzione tra le distribuzioni marginali e la distribuzione multivariata e garantisce l'esistenza e, sotto alcune condizioni, l'unicità della funzione stessa, avendo fissate le marginali. Di seguito viene riportato l'enunciato del teorema.

Sia H una funzione di ripartizione congiunta n-dimensionale con funzioni marginali F_1, $F_2, \dots, F_n$.
Allora esiste una copula C tale che per ogni x in $\mathbb{R}^n$,

$$H(x_1, x_2, ..., x_n) = C[F_1(x_1), F_2(x_2), ..., F_n(x_n)] \tag{2.1}$$

[1] La funzione differenza Δ è definita secondo la seguente espressione:

$$\Delta_{a_k}^{b_k} C(\boldsymbol{t}) = C(t_1, \dots, t_{k-1}, b_k, t_{k+1}, \dots, t_n) - C(t_1, \dots, t_{k-1}, a_k, t_{k+1}, \dots, t_n)$$

Se $F_1, F_2, \ldots, F_n$ sono tutte funzioni continue allora C è unica; altrimenti C è univocamente determinata sul $RanF_1 \times RanF_2 \times \ldots \times RanF_n$. Viceversa, se C è una copula ed $F_1, F_2, \ldots, F_n$ sono funzioni di ripartizione, allora la funzione H definita dalla (2.1) è una funzione di ripartizione congiunta n-dimensionale con marginali $F_1, F_2, \ldots, F_n$.

Dall'enunciato del teorema si evince quindi l'essenza stessa della copula, quale funzione che assegna a ciascun insieme di valori delle funzioni di ripartizione marginali uno e un solo valore della distribuzione congiunta. Inoltre si sottolinea come condizione necessaria affinché la funzione C sia unica è l'assoluta continuità delle funzioni di ripartizione marginali.

Come si è detto, questo teorema risulta fondamentale non soltanto perché chiarisce il senso della funzione copula, ma anche perché, come vedremo, risulta il punto di partenza per la costruzione delle copule stesse (metodo di inversione) e per l'individuazione di nuove funzioni di ripartizione congiunte.

Come corollario importante del teorema appena enunciato si ha il seguente:

Siano H, C, $F_1, F_2, \ldots, F_n$ definite come nel teorema precedente, e siano $F_1^{[-1]}, F_2^{[-1]}, \ldots, F_n^{[-1]}$ le funzioni quasi-inverse di $F_1, F_2, \ldots, F_n$. Allora per ogni u in $\boldsymbol{I}^n$ si ha:

$$C(u_1, ..., u_n) = H(F_1^{[-1]}(u_1), ..., F_n^{[-1]}(u_n)) \tag{2.2}$$

Tale risultato è alla base del cosiddetto metodo di inversione per la costruzione delle funzioni copula in quanto grazie alla (2.2), avendo fissate le funzioni di ripartizione marginali e l'espressione della distribu-

zione congiunta è possibile ottenere l'espressione della funzione copula che caratterizza la dipendenza tra le variabili in questione. Inoltre, proprio grazie alla (2.2) si può verificare la mancanza di unicità della funzione copula quando le variabili casuali sono discrete. Infatti, considerando per semplicità il caso bivariato, l'uguaglianza precedente si ottiene sfruttando sostanzialmente due proprietà:

1. se esiste l'inversa di una funzione di ripartizione F allora

$$F^{-1}(u) \sim F$$

   dove $U \sim U(0,1)$;

2. se F è una funzione continua tale che $X \sim F \;\Rightarrow\; F(x) \sim U(0,1)$.

Quindi, posto $X \sim F$ e $Y \sim G$, aventi distribuzione congiunta H, se F e G sono funzioni continue, per la (2.) si ha:

$$U = F(x) \sim U(0,1) \text{ e } V = G(y) \sim U(0,1).$$

Il vettore casuale bidimensionale $\boldsymbol{U} = [F(x); G(y)]$ è quindi un vettore di variabili casuali uniformi e dunque $[F(x), G(y)] \sim C$, in quanto per definizione la funzione C non è altro che la distribuzione di un vettore casuale uniforme. Inoltre per la (1.) $F^{-1}(u) \sim F$ e $G^{-1}(v) \sim G$ da cui si ha che $[F^{-1}(u), G^{-1}(v)] \sim H$ per cui vale la (2.2).

Al contrario, se le variabili coinvolte non fossero di tipo continuo la funzione copula che garantisce l'uguaglianza (2.1) non è più unica e ciò è dovuto al fatto che non vale più la (2.) [Joe, 1997]. Il teorema di

Sklar garantisce comunque l'esistenza di un'unica sub-copula[2] C' tale che

$$Dom\ C' = RanF \times RanG$$

$$H(x, y) = C'[F(x), G(y)]\ \forall x, y \in R \tag{2.3}$$

Inoltre, dato che ogni sub-copula può essere estesa ad una copula, ma tale estensione è generalmente non unica [Nelsen, 2006], esiste più di una funzione che soddisfa l'uguaglianza (2.1), ovvero esiste un'intera classe C_H di funzioni [Carley, 2002] costituita da tutte le possibili estensioni dell'unica sub-copula definita sul $RanF \times RanG$. Tutte queste funzioni ovviamente coincidono per i punti individuati dal prodotto cartesiano dei condomini delle funzioni di ripartizione marginali.

Una possibile soluzione dell'equazione (2.1), e dunque una funzione copula appartenente alla classe C_H, è ottenuta definendo la funzione copula sulla griglia di punti individuati dai condomini di F e G e assumendo poi che la funzione stessa sia uniforme tra un punto e l'altro. Ciò porta ad ottenere una funzione di densità multivariata con distribuzioni marginali uniformi.

Nel prosieguo, anche se non esplicitamente indicato, si intenderà fornire i risultati per variabili casuali di tipo continuo.

2.1.3 I limiti di Frechet-Hoeffding

L'estensione delle distribuzioni multivariate e quindi anche delle funzioni copula è un risultato noto che va sotto il nome di limiti di Frechet-Hoeffding.

[2]La sub-copula è definita similmente alla copula con la differenza che il suo dominio è un sottoinsieme dello spazio $\boldsymbol{I}^n$. Per approfondimenti si veda Nelsen [2006].

Si consideri una funzione di ripartizione congiunta n-dimensionale $H(x_1, \ldots, x_n)$ con distribuzioni marginali F_1, F_2,..., F_n. Allora si può dimostrare che la funzione H è limitata inferiormente e superiormente rispettivamente dalle funzioni F_L e F_U così definite:

$$F_L(x_1, x_2, \ldots, x_n) = max[\sum_{i=1}^{n} F_i - n + 1, 0]$$

$$F_U(x_1, x_2, \ldots, x_n) = min[F_1, F_2, ..., F_n]$$

Di conseguenza anche la funzione copula è sempre compresa tra un limite inferiore e uno superiore essendo:

$$W(u_1, u_2, \ldots, u_n) \leqslant C(u_1, u_2, \ldots, u_n) \leqslant M(u_1, u_2, \ldots, u_n)$$

con

$$W(u_1, u_2, \ldots, u_n) = max[\sum_{i=1}^{n} u_i - n + 1, 0]$$

e

$$M(u_1, u_2, \ldots, u_n) = min(u_1, u_2, \ldots, u_n)$$

Il limite inferiore viene generalmente indicato con C^- e detto *minimum copula* oppure limite inferiore di Frechet. Analogamente, il limite superiore, generalmente indicato con C^+, viene detto *maximum copula* oppure limite superiore di Frechet. Va sottolineato che, mentre

il limite superiore C^+ è esso stesso una copula, il limite inferiore C^- lo è solo nel caso in cui $n = 2$, oppure nel caso in cui $n > 2$ solo se sono soddisfatte alcune condizioni [Joe, 1997, pag.61]

2.2 Le misure di dipendenza

In questa sezione si vedrà come, all'atto pratico, la funzione copula possa essere utilizzata per valutare la dipendenza tra le variabili casuali, indipendentemente dal comportamento delle singole marginali. A tal fine si rivedranno alcune delle misure di dipendenza note, riscritte in termini di copula. Infatti, dato che si può dimostrare [Nelsen, 2006] che la funzione copula di un vettore casuale (X, Y) è invariante rispetto a trasformazioni strettamente crescenti delle componenti del vettore stesso, esisteranno delle misure di dipendenza a loro volta invarianti rispetto a dette trasformazioni. Tali misure saranno quelle calcolate a partire dalla funzione copula e dipenderanno perciò solo dai parametri di tale funzione e non dai parametri che caratterizzano le distribuzioni marginali.

2.2.1 La concordanza

Due tra le misure di tipo scala-invarianti, molto diffuse sono gli indici *tau di Kendall* e *rho di Spearman*. Si vedrà dunque come possano essere espressi in termini di funzione copula e si verificherà la relazione tra loro intercorrente. Questi due indici misurano quella che viene detta *concordanza*, ovvero la propensione dei valori alti di una variabile ad associarsi a valori alti della seconda variabile e viceversa, dei valori bassi ad associarsi a valori bassi. Più precisamente si parlerà di concordanza se, date due osservazioni (x_i, y_i) e (x_j, y_j), estratte dal vettore casuale (X, Y), si ha $x_i < x_j$ e $y_i < y_j$ oppure $x_i > x_j$ e

$y_i > y_j$,viceversa di discordanza se $x_i < x_j$ e $y_i > y_j$ oppure $x_i > x_j$ e $y_i < y_j$.

Ne consegue che le due coppie (x_i, y_i) e (x_j, y_j) saranno concordanti se $(x_i - x_j)(y_i - y_j) > 0$ e discordanti se $(x_i - x_j)(y_i - y_j) < 0$.

Si può perciò definire l'indice di concordanza *tau* di Kendall tra due vettori casuali (X_1, Y_1) e (X_2, Y_2) come la differenza tra la probabilità di concordanza e quella di discordanza:

$$\begin{aligned} &\tau_{X,Y} = \\ &= P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0] \end{aligned} \tag{2.4}$$

Quindi, potendo esprimere le probabilità contenute nella (2.4) in termini di funzione copula si ottiene, dopo alcuni passaggi:

$$\tau_{X,Y} = 4E[C(U, V)] - 1 \tag{2.5}$$

Dall'espressione (2.5) è immediato verificare che l'indice $\tau_{X,Y}$ non dipende in nessun caso dai parametri delle distrubuzioni marginali, ma solo dai parametri della funzione C.

L'indice *rho di Spearman*, anch'esso basato sulle probabilità di concordanza e discordanza, viene costruito a partire da tre vettori (X_1, Y_1), (X_2, Y_2) e (X_3, Y_3) estratti dalla coppia di variabili casuali (X, Y) avente distrubuzione congiunta H.
L'indice $\rho_{X,Y}$ ha la seguente espressione:

$$\begin{aligned} &\rho_{X,Y} = \\ &= 3\left(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]\right) \end{aligned} \tag{2.6}$$

quindi tiene conto di una coppia di variabili (X_1, Y_1) con distribuzione

congiunta H e di una coppia (X_2, Y_3) di variabili indipendenti. Da queste considerazioni, dopo alcuni passaggi, si ottiene l'espressione dell'indice in termini di funzione C come segue:

$$\rho_{X,Y} = 12 \int\int_{\boldsymbol{I^2}} C(u,v)dudv - 3 \tag{2.7}$$

I due indici sono strettamente legati, e si può dimostrare [Nelsen] che tra $\tau_{X,Y}$ e $\rho_{X,Y}$ può essere stabilita la seguente doppia disuguaglianza:

$$-1 \leqslant 3\tau - 2\rho \leqslant 1$$

Inoltre possono essere derivate altre disuguaglianze che non verranno qui riportate .

2.2.2 Alcune nozioni di dipendenza

Insieme ai due indici appena presentati è interessante andare a valutare alcune propietà di dipendenza di cui le copule possono godere. Si daranno soltanto le definizioni più salienti.

Positively Quadrant Dependent (PQD). Due variabili casuali X e Y che godono di questa proprietà sono due variabili la cui probabilità di presentare valori simultaneamente alti (o simultaneamente bassi) è almeno pari alla probabilità che indipendentemente una dall'altra presentino valori alti (o bassi). Formalmente, X e Y si dicono PQD se per ogni x,y in $\mathbb{R}^2$ si ha:

$$P[X \leqslant x, Y \leqslant y] \geqslant P[X \leqslant x]P[Y \leqslant y] \tag{2.8}$$

o equivalentemente

$$P[X > x, Y > y] \geqslant P[X > x]P[Y > y]$$

Analogamente, si dicono *Negatively Quadrant Dependent (NQD)* se:

$$P[X \leqslant x, Y \leqslant y] \leqslant P[X \leqslant x]P[Y \leqslant y]$$

o anche

$$P[X > x, Y > y] \leqslant P[X > x]P[Y > y]$$

Considerando la (2.8) in termini di funzione copula è immediato verificare che X e Y saranno PQD se $C(u, v) \geqslant uv$ e, di converso, saranno NQD se $C(u, v) \leqslant uv$. La (2.8) può essere letta anche in termini di distribuzioni condizionate, dato che essa può essere riscritta come segue:

$$P[Y \leqslant y | X \leqslant x] \geqslant P[Y \leqslant y]$$

Inoltre, si dimostra che se vale la proprietà di PQD si ha:

$$3\tau_{X,Y} \geqslant \rho_{X,Y} \geqslant 0$$

Left Tail Decreasing (LTD). Sempre considerando la distribuzione condizionata di Y data la variabile X si dice che Y è LTD in X $[LTD(Y|X)]$ se $P[Y \leqslant y | X \leqslant x]$ è una funzione non crescente in x per tutti i valori di y e dunque se $C(u, v)/u$ è non crescente in u $\forall v \in \boldsymbol{I}$. Analogamente Y è *Right Tail Increasing* in X $[RTI(Y|X)]$ se $P[Y > y | X > x]$ è una

funzione non descrescente di x per tutti i valori di y e conseguentemente se $[1 - u - v + C(u, v)]/(1 - u)$ è non decrescente in u $\forall v \in \boldsymbol{I}$. In generale queste due proprietà appena enunciate vengono dette di *Tail Monotonicity* e implicano la precedente proprietà di PQD.

Stocastichally Increasing (SI). Considerando ancora le distribuzioni condizionate, questa volta ponendo un'uguaglianza per la variabile casuale condizionante, si dice che Y è *Stochastically Increaising* di x $[SI(Y|X)]$ se la funzione $P[Y > y|X = x]$ è una funzione non decrescente di x per tutti i valori di y. Al contrario, si dice che Y è *Stochastically Decreasing* in x $[SD(Y|X)]$ se la funzione $P[Y > y|X = x]$ è una funzione non crescente di x per tutti i valori di y. In termini di funzione copula si può verificare che $SI(Y|X)$ se, $\forall v \in I$, $\partial C(u, v)/\partial u$ è non crescente in u e, viceversa, $SD(Y|X)$ se, $\forall v \in I$, $\partial C(u, v)/\partial u$ è non decrescente in u
Questa proprietà implica la proprietà di *Tail Monotonicity* e di conseguenza la proprietà PQD. Infatti si dimostra che se $SI(Y|X)$ allora $LTD(Y|X)$ e $RTI(Y|X)$.

2.3 I metodi di stima

La stima del modello ipotizzato può avvenire secondo diverse metodologie. Innanzitutto vanno distinte le procedure non parametriche da quelle semi-parametriche e parametriche.

L'approccio parametrico è forse quello maggiormente impiegato e largamente basato sull'utilizzo della funzione di verosimiglianza. Di seguito verrà quindi fornita l'espressione della funzione di log-verosimiglianza e i due metodi basati su questa: il metodo FML (Full Maximum Likelihood) e il metodo IFM (Inference For Margins). Per gli stimatori ottenuti verranno dettagliate le proprietà fondamentali.

Si farà poi un accenno agli altri metodi di stima, considerando il metodo dei momenti e i metodi non parametrici.

2.3.1 La funzione di densità e la funzione di verosimiglianza

Sia $C(\boldsymbol{u}; \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\theta})$ una funzione copula definita nello spazio $\mathbf{I}^n$, dove $\boldsymbol{\alpha}_i$ è il vettore dei parametri associato all'i-esima funzione di ripartizione marginale e $\boldsymbol{\theta}$ il vettore che caratterizza la funzione copula C.

Le metodologie seguenti verranno presentate ipotizzando che i vettori suddetti siano tra loro differenti e dunque non vi siano parametri comuni a più funzioni.[3].

Visto che la funzione copula non è altro che una distribuzione congiunta di un vettore di variabili casuali uniformi, si può definire la funzione di densità $c(\boldsymbol{u}; \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\theta})$ associata a C nel modo seguente:

$$c(\boldsymbol{u}; \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\theta}) = \frac{\partial^n C(\boldsymbol{u}; \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\theta})}{\partial u_1 \ldots \partial u_n}$$

Inoltre dalla (2.1) si ha:

$$\frac{\partial^n H(\boldsymbol{x}; \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\theta})}{\partial x_1 \ldots \partial x_n} = \\ = \frac{\partial^n C[F_1(x_1; \boldsymbol{\alpha}_1), \ldots, F_n(x_n; \boldsymbol{\alpha}_n), \boldsymbol{\theta})]}{\partial F_1(x_1; \boldsymbol{\alpha}_1) \ldots \partial F_n(x_n; \boldsymbol{\alpha}_n)} \cdot \prod_{i=1}^{n} \frac{\partial F_i(x_i; \boldsymbol{\alpha}_i)}{\partial x_i}$$

e quindi

[3]Procedure particolari andrebbero adottate in caso di parametri comuni a più funzioni.

$$h(\boldsymbol{x};\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n,\boldsymbol{\theta}) = c[F_1(x_1;\boldsymbol{\alpha}_1),\ldots,F_n(x_n;\boldsymbol{\alpha}_n)]\cdot\prod_{i=1}^{n} f_i(x_i;\boldsymbol{\alpha}_i)$$

dove $f_i(x_i;\boldsymbol{\alpha}_i)$ è la funzione di densità associata all'i-esima funzione di ripartizione marginale $F_i(x_i;\boldsymbol{\alpha}_i)$.

Dunque la funzione di log-verosimiglianza, per un campione di ampiezza m è data da:

$$\begin{aligned} &l(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_m,\boldsymbol{\theta}) = \\ &= \sum_{j=i}^{m} \log c[F_1(x_{1j};\boldsymbol{\alpha}_1),\ldots,F_n(x_{nj};\boldsymbol{\alpha}_n)] + \sum_{j=1}^{m}\sum_{i=1}^{n} \log f_i(x_{ij},\boldsymbol{\alpha}_i) \end{aligned} \tag{2.9}$$

I metodi per il calcolo delle stime di massima verosimiglianza saranno perciò basati sull'espressione (2.9), ma a seconda dell'approccio utilizzato per la ricerca del punto di massimo, si avranno diversi risultati, esaminati in dettaglio nel prosieguo. Si sottolinea inoltre che, per tali metodi, generalmente non esistono soluzioni in forma chiusa ed è necessario associare procedimenti di calcolo numerico per ottenere gli stimatori. Andrà perciò posta particolare attenzione alla scelta del metodo numerico e di conseguenza ai valori iniziali da utilizzare per implementarlo.

2.3.2 Il metodo FML

Il metodo FML prevede l'utilizzo classico del metodo di massima verosimiglianza. Ovvero, data la funzione di densità congiunta, e la

relativa funzione di log-verosimiglianza, il vettore di stimatori di massima verosimiglianza sarà dato dalle coordinate del punto di massimo della funzione.

Analiticamente, data la (2.9), lo stimatore di massima verosimiglianza del vettore $\boldsymbol{\xi} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\theta}]$ sarà dato da:

$$\hat{\boldsymbol{\xi}}_{FML} = \arg \max_{\boldsymbol{\xi} \in \Xi} l(\boldsymbol{\xi})$$

dove Ξ indica lo spazio parametrico.

Se si assumono le usuali condizioni di regolarità, gli stimatori che si ottengono sono consistenti, asintoticamente efficienti e asintoticamente normali. Per cui si ha:

$$\sqrt{n}(\hat{\boldsymbol{\xi}}_{FML} - \boldsymbol{\xi}_0) \rightarrow N(0, I^{-1}(\boldsymbol{\xi}_0))$$

dove $I^{-1}(\boldsymbol{\xi}_0)$ indica l'inversa della matrice di informazione di Fisher e $\boldsymbol{\xi}_0$ il vettore di parametri incognito.

La procedura appena descritta può risultare particolarmente dispendiosa da un punto di vista computazionale, soprattutto quando la densità congiunta in esame è parametrizzata da un vettore di ampie dimensioni, eventualità che accade ad esempio quando numerose sono le variabili casuali coinvolte e dunque le distribuzioni marginali da stimare. Il prossimo paragrafo sarà perciò dedicato ad illustrare una metodologia di stima, sempre basata sulla funzione di verosimiglianza, ma certamente più semplice e snella in quanto caratterizzata da due step successivi.

2.3.3 Il metodo IFM

Come abbiamo appena visto, il metodo descritto nel precedente paragrafo prevede la ricerca del massimo della funzione di log-verosimi-

glianza (2.9), che sarà un punto appartenente allo spazio parametrico Ξ, soluzione del sistema

$$\left(\frac{\partial l}{\partial \boldsymbol{\alpha}_1}, \ldots, \frac{\partial l}{\partial \boldsymbol{\alpha}_n}, \frac{\partial l}{\partial \boldsymbol{\theta}}\right) = \mathbf{0}$$

Osservando però proprio la (2.9) si nota che essa è data dalla somma di due quantità, la prima riferibile alla funzione copula e la seconda dipendente unicamente dalle distribuzioni marginali. E' quindi lecito supporre di poter separare i due step di stima, quello riferito ai parametri delle distribuzioni marginali e quello riferito ai parametri della copula considerata. Questo metodo è quello che viene detto *Inference for Margins* seguendo la terminologia di McLeish e Small [1988] e Xu [1996].
Dunque:
a) l'i-esima funzione di log-verosimiglianza l_i viene massimizzata separatamente dalle altre funzioni, ottenendo gli stimatori $\tilde{\boldsymbol{\alpha}}_i$ (questo passaggio si ripete per $i = 1, \ldots, n$);
b) la funzione $l(\boldsymbol{\theta}, \tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_n)$ viene massimizzata rispetto a $\boldsymbol{\theta}$ per ottenere lo stimatore $\tilde{\boldsymbol{\theta}}$.

Nel complesso si ottiene il punto di coordinate $(\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_n, \tilde{\boldsymbol{\theta}})$, quale soluzione del sistema

$$\left(\frac{\partial l_1}{\partial \boldsymbol{\alpha}_1}, \ldots, \frac{\partial l_n}{\boldsymbol{\alpha}_n}, \frac{\partial l}{\partial \boldsymbol{\theta}}\right) = \mathbf{0}$$

Gli stimatori ottenuti con questo metodo sono generalmente differenti dai precedenti e dunque c'è da chiedersi se le proprietà enunciate per i primi valgano anche per i secondi. Va comunque sottolineato che, proprio perchè più semplici da ottenere, possono essere considerati un valido punto di partenza in un qualunque procedimento iterativo per il calcolo degli stimatori FML.

Relativamente alle proprità Joe [1997] dimostra che, sotto le usuali ipotesi di regolarità, gli stimatori ottenuti sono ancora asintoticamente

normali e non distorti, con matrice di varianza e covarianza data dall'inversa della matrice di informazione di Godambe [Joe, 1997, pag. 301]. Ulteriori risultati sull'efficienza asintotica degli stimatori IFM possono essere ritrovati in Joe [2005].

Dunque, grazie a questo metodo di stima, è possibile fornire un modello per il comportamento congiunto di più variabili in due passi successivi, il primo diretto a valutare il comportamento delle singole variabili (definendo e stimando appropriate distribuzioni marginali), il secondo volto a definire la dipendenza e dunque l'andamento congiunto (definendo e stimando la funzione copula).

2.3.4 Altri metodi di stima

Alcune volte, per ovviare ai problemi computazionali connessi ai metodi descritti finora, si preferisce optare per metodi semi-parametrici, come quello che prevede ad esempio di massimizzare il logaritmo della funzione di pseudo-verosimiglianza, definita come segue:

$$L(\boldsymbol{\xi}) = \sum_{j=i}^{m} \log c_{\boldsymbol{\xi}}[F_{1m}(x_{1j}), \ldots, F_{nm}(x_{nj})]$$

dove F_{km} è la funzione di ripartizione empirica della k-esima variabile, basata su un campione di ampiezza m, ed eventualmente riscalata secondo un fattore di correzione pari ad $n/(n+1)$. [Genest et al., 1995]. Si sottolinea come sia possibile utilizzare anche metodi pienamente non parametrici, quali ad esempio tecniche di smoothing, non formulando ipotesi specifiche neanche sulla struttura di dipendenza.

Un altro metodo che generalmente non presenta particolari problemi da un punto di vista computazionale è il metodo dei momenti, che può essere impiegato per la stima dei parametri di associazione considerando una qualche misura di dipendenza. Ad esempio, supponendo di avere un solo parametro di associazione e avendo a disposi-

zione l'espressione del coefficiente *tau di Kendall*, che sarà funzione del suddetto parametro, si può eguagliarlo al suo corrispettivo campionario e ricavare lo stimatore cercato. Formalmente, il metodo prevede l'utilizzo della seguente espressione:

$$\tau_\theta(X, Y) = \tau_n(X, Y)$$

che risolta in termini di θ fornisce lo stimatore $\hat{\theta_n}$ di θ. Ovviamente, quando i parametri di associazione sono più di uno, il numero necessario di equazioni da imporre sarà connesso proprio alle dimensioni del vettore di parametri da stimare.

2.4 Alcune famiglie parametriche

In questa sezione verranno presentate alcune famiglie parametriche di uso comune e le relative proprietà, non prima di aver introdotto la classe delle cosiddette copule Archimediane, alla quale appartengono le funzioni che si andranno a descrivere nei paragrafi seguenti.

2.4.1 Le copule Archimediane

Le copule Archimediane sono un'importante classe di funzioni il cui utilizzo è largamente diffuso soprattutto grazie alla semplicità con la quale possono essere derivate, alla grande varietà di funzioni appartenenti a questa classe e alle proprietà di cui godono, tra cui la simmetria ovvero $C(u, v) = C(v, u)$, e la proprietà associativa, $C(C(u, v), w) = C(u, C(v, w))$.

Formalmente, sia ϕ una funzione continua, strettamente decrescente e convessa da $[0, 1]$ a $[0, \infty]$ tale che $\phi(1) = 0$. Sia inoltre $\phi^{[-1]}(t)$ la pseudo-inversa della funzione ϕ, ovvero
$\phi^{[-1]}(t) = \phi^{-1}(t)$ per $t \in [0, \phi(0)]$, e $\phi^{[-1]}(t) = 0$ per $t > \phi(0)$.

Allora, una copula Archimediana è una funzione C da $[0,1]^n$ a $[0,1]$ tale che:

$$C(u_1, \ldots, u_n) = \phi^{[-1]}(\phi(u_1) + \ldots + \phi(u_n))$$

e ϕ viene detto generatore della copula C[4].

2.4.2 Funzioni copula ad un parametro

Verranno ora descritte, per il caso bivariato (anche se molte possono essere estese al caso multivariato) alcune delle famiglie più note, in cui la distribuzione è caratterizzata da un unico parametro.

Normale bivariata. Si consideri una copula Archimediana con generatore dato dall'inversa Ψ^{-1} della funzione di ripartizione Ψ di una variabile casuale normale standardizzata. Allora si ottiene la copula con la seguente funzione di densità:

$$c(u, v; \theta) =$$
$$= (1-\theta^2)^{-1/2} exp\{-\frac{1}{2}(1-\theta^2)^{-1}[\tilde{u}^2 + \tilde{(}v)^2 - 2\theta\tilde{u}\tilde{v}]\} \cdot exp\{\frac{1}{2}[\tilde{u}^2 + \tilde{v}^2]\}$$

dove $0 \leqslant \theta \leqslant 1$, $\tilde{u} = \Phi^{-1}(u)$ e $\tilde{v} = \Phi^{-1}(v)$.
Questa funzione può essere estesa al caso multivariato e comprendere anche il caso di dipendenza negativa estendendo il range dei valori di θ. Inoltre raggiunge il limite inferiore C^- quando $\theta = -1$ e il limite superiore C^+ quando $\theta = 1$. Per $\theta = 0$ si ha l'indipendenza delle variabili casuali, ovvero $C(u, v) = uv$.

[4]Più precisamente in questo caso ϕ è un generatore additivo, ma può essere considerato anche un generatore moltiplicativo, per cui si avrebbe $C(u_1, \ldots, u_n) = \phi^{[-1]}(\phi(u_1) \cdot \ldots \cdot \phi(u_n))$.

Copula di Gumbel. Questa copula ha la seguente espressione:

$$C(u, v; \theta) = exp\{-(\log(u)^{\theta} + \log(v)^{\theta})^{1\ \theta}\}$$

Appartiene alla classe delle copule Archimediane ed anche alla classe delle *Extreme Value Copulas* (EVT). Il limite C^+ si ottiene quando $\theta \to \infty$, mentre il limite inferiore C^- non viene mai raggiunto e non contempla il caso di dipendenza negativa. Le marginali sono indipendenti se $\theta = 1$.

Copula di Frank. E' definita per $-\infty \leqslant \theta < \infty$ come segue:

$$C(u, v; \theta) = -\theta^{-1} \log([\eta - (1 - e^{-\theta u})(1 - e^{-\theta v})]/\eta)$$

dove $\eta = 1 - e^{\theta}$.
I limiti C^+, C^- e l'indipendanza si hanno rispettivamente per $\theta \to \infty$, $\theta \to -\infty$ e $\theta \to 0$. A differenza della precedente quindi può descrivere anche situazioni di dipendenza negativa.

Copula di Clayton. Nota anche come copula di Kimeldorf e Sampson [1975], è definita per $0 \leqslant \theta < \infty$, dall'espressione

$$C(u, v; \theta) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$$

Può essere parzialmente estesa fino al valore di $\theta = -1$ per ammettere dipendenza negativa, ma non raggiunge mai il limite inferiore C^-. Per $\theta \to \infty$ si ha $C \to C^+$ e per $\theta \to 0$ si ha $C = uv$. Anche quest'ultima copula appartiene alla classe delle copule Archimediane.

Tutte le copule presentate godono della proprietà di SI.

2.4.3 Funzioni copula a due parametri

Quando l'intento è quello di modellare più tipi di dipendenza, può essere opportuno utilizzare più di un parametro di associazione.

Un metodo per la costruzione di copule aventi due parametri è quello di considerare l'*interior* e l'*exterior power family* [Nelsen, 2006] associate al generatore ϕ in $\mathbf{\Omega}$ e date ripettivamente dagli insiemi:

$$\{\phi_{\alpha,1} \in \mathbf{\Omega} | \phi_{\alpha,1}(t) = \phi(t^{\alpha})\}$$

$$\{\phi_{1,\beta} \in \mathbf{\Omega} | \phi_{1,\beta}(t) = [\phi(t)]^{\beta}\}$$

con $\beta \geqslant 1$ e $0 < \alpha \leqslant 1$ (o anche semplicemente $\alpha > 0$ se ϕ è differenziabile due volte e $t\phi'(t)$ è non decrescente in $(0, 1)$), affinchè $\phi_{\alpha,1}$ e $\phi_{1,\beta}$ siano anch'essi elementi di $\mathbf{\Omega}$.

Utilizzare le funzioni $\phi_{\alpha,1}$ e $\phi_{1,\beta}$ quali generatori di copule Archimediane significa introdurre un ulteriore parametro di associazione che consente maggiore flessibilità nel descrivere le relazioni di dipendenza tra le variabili casuali. Da qui la possibilità di costruire numerose famiglie di copule a due parametri, proprio a partire dalle copule viste nella sezione precedente. Diverse funzioni possono essere ritrovate in Joe [1997], complete delle proprietà più salienti.

Capitolo 3

Misure di dissimilarità e distanze tra distribuzioni multivariate

3.1 Introduzione

La valutazione della discrepanza tra funzioni di probabilità è un aspetto rilevante in diversi contesti. Sicuramente risulta centrale nella costruzione di algoritmi di clustering ogni qualvolta gli oggetti da classificare sono descritti proprio da distribuzioni. Infatti, dovendo considerare un insieme di elementi E tra i quali definire una misura di dissimilarità o distanza, nella pratica può accadere che E sia un'insieme di n unità statistiche, descritte da un set di variabili, se si considerano strutture di dati classiche, oppure che sia un insieme di oggetti descritti da intervalli, distribuzioni o altro. A seconda della tipologia degli elementi, devono ovviamente essere definite misure appropriate.

In questo capitolo, in coerenza con gli obiettivi del presente lavoro di tesi, si focalizzarà l'attenzione su oggetti descritti da distribuzioni

e di conseguenza, verranno trattate nel dettaglio le misure atte a valutare la discrepanza tra distribuzioni di probabilità. In tale contesto, l'algoritmo di classificazione sarà diretto alla ricerca di una partizione dei dati in classi omogenee, ovvero tali che gli oggetti appartenenti alla medesima classe risultino il più possibile simili e quindi con funzioni di probabilità tra loro 'vicine', mentre gli oggetti appartenenti a classi diverse siano quanto più differenti possibile. A tal fine dunque, la scelta della metrica risulta fondamentale.

Altro ambito in cui è essenziale la definizione di una misura di discrepanza tra distribuzioni è quello in cui ci si propone di studiare il tasso di convergenza di certe misure in termini di probabilità. In questo particolare contesto, una review di importanti metriche definite su misure di probabilità, nonchè delle relazioni intercorrenti tra esse, può essere ritrovata in Gibbs and Su [2002]. Uno schema riassuntivo delle metriche trattate nel suddetto lavoro è riportato in tabella 3.1.

Nel prosieguo, verranno riprese e descritte in dettaglio alcune di queste misure e ne verranno introdotte delle altre, naturalmente inquadrabili nella cosiddetta teoria dell'informazione, della quale verranno richiamati i concetti fondamentali. Infine saranno brevemente presentate alcune proposte originali.

3.2 Una classe di misure di divergenza

Csiszár [1967] descrisse e analizzò una classe di misure di divergenza basata sul rapporto di due funzioni di probabilità o di densità. A questa classe appartengono molte delle misure che verranno di seguito citate. Al fine di snellire la trattazione, verranno riportate le formulazioni relative alle sole variabili casuali continue, ma l'estensione al caso discreto risulta immediata.

Definizione

Abbreviation	Metric
D	Discrepancy
H	Hellinger distance
KL	Relative entropy (or Kullback-Leibler divergence)
K	Kolmogorov (or Uniform) metric
L	Lévy metric
P	Prokhorov metric
S	Separation distance
TV	Total variation distance
W	Wasserstein (or Kantorovich) metric
χ^2	χ^2 distance

Tabella 3.1: Misure di distanza presenti in Gibbs and Su [2002]

Sia Ω uno spazio misurabile con $\sigma-$algebra A e sia $M(\Omega, A)$ un insieme di distribuzioni definite sullo spazio (Ω, A). Siano $F, G \in M(\Omega, A)$ due funzioni di ripartizione con rispettive funzioni di densità f e g. Si consideri il rapporto $\lambda(x) = f(x)/g(x)$.
Si definisce $\phi - divergence$ la seguente misura di divergenza:

$$\begin{aligned} d(F, G; \phi) &= E_F[\phi(\lambda(X))] = \\ &= \int_X \phi(\lambda(x)) dF(x) \end{aligned} \tag{3.1}$$

dove $\phi(\cdot)$ è una funzione convessa a valori reali tale che $\phi(1) = 0$.

Utilizzando la disuguaglianza di Jensen, si può dimostrare che quest'indice è sempre maggiore o al più pari a zero.

3.2.1 La divergenza del χ^2

Un caso particolare della $\phi - divergence$ è la nota divergenza del χ^2 che si ottiene quando si pone $\phi(\lambda) = (\lambda - 1)^2$.

In questo caso infatti si ha:

$$d_{\chi^2}(F, G) = \int_X \left(\frac{g(x)}{f(x)} - 1\right)^2 f(x)dx = \int_X \frac{(f(x) - g(x))^2}{f(x)}dx$$

Quest'indice è uno tra i più noti, essendo ampiamente utilizzato in alcune tecniche di analisi multivariata (soprattutto nella sua variante per variabili discrete), ma risulta essere asimmetrico e quindi non può essere considerato una misura di dissimilarità.

3.2.2 Il coefficiente di affinità di Hellinger

Il coefficiente di affinità di Hellinger è definito come segue:

$$d^{(s)}(F, G) = \int g(x)^s \cdot f(x)^{(1-s)}dx$$

e si ottiene facilmente dalla 3.1 ponendo $\phi(\lambda) = \lambda^s$ con $0 \leqslant s \leqslant 1$.

Questo coefficiente misura la similarità tra le funzioni f e g e da esso sono state derivate diverse misure di distanza come la distanza di Chernoff [Chernoff, 1952] di ordine s:

$$d(F, G) = -\log d^{(s)}(F, G)$$

e l'*information gain* di ordine s proposta da Rényi nel 1961 [Renyi, 1961]:

$$d(F, G) = \frac{\log d^{(s)}(F, G)}{s - 1}$$

Inoltre, come caso particolare di questa misura, si ottiene la *distanza di Bhattacharyya* se si pone $s = 1/2$ [Fukunaga, 1972]:

$$d^{(1/2)}(F, G) = \int \sqrt{g(x) \cdot f(x)} dx$$

Infine, ancora considerando $s = 1/2$ e contestualmente la seguente trasformazione del coefficiente di affinità:

$$\sqrt{2(1 - d(F, G))}$$

si ottiene la misura di distanza detta *distanza di Hellinger* [Gibbs and Su, 2002]:

$$d_H(F, G) = \left[2 \left(1 - \int \sqrt{g(x) \cdot f(x)} dx \right) \right]^{1/2}$$

3.2.3 Variation distance e Total Variation distance

Si consideri l'espressione della ben nota distanza di Minkoski a norma L_1 e si supponga di calcolare tale distanza tra due funzioni di densità f e g. L'espressione che si ottiene è la seguente:

$$d_V(F, G) = \int_X |f(x) - g(x)| dx$$

Questo caso particolare prende il nome di *Variation distance* e si ottiene dalla 3.1 ponendo $\phi(\lambda) = |\lambda - 1|$. La *Variation distance* è una misura di divergenza simmetrica che viene molto utilizzata anche in norma L_2.

La *Total Variation distance* viene definita sempre a partire dalle differenze delle funzioni di densità, in valore assoluto, ma più semplicemente viene considerato l'estremo superiore di tali differenze sullo spazio $A \subset \Omega$:

$$d_{TV}(F,G) = \sup_{A \subset \Omega} |f(x) - g(x)|$$

3.2.4 Distanze di Wasserstein e di Mahalanobis-Wasserstein

Date due distribuzioni univariate F e G, si definisce distanza di Wasserstein in norma L_2, la seguente:

$$d_W(F,G) = \left(\int_0^1 \left(F^{-1}(t) - G^{-1}(t) \right)^2 dt \right)^{\frac{1}{2}}$$

Questa misura gode di diverse proprietà tra le quali la decomposizione nelle tre componenti di posizione, variabilità e forma [Irpino and Romano, 2007] e la decomposizione dell'inerzia totale in inerzia entro i gruppi e inerzia tra i gruppi, nel caso in cui si consideri una partizione dell'insieme E.

Verde and Irpino [2008] utilizzano questa distanza nel contesto dell'analisi di dati a intervallo e propongono una nuova misura di distanza al fine di tener conto dell'interdipendenza tra le variabili considerate. Tale distanza, detta distanza di Mahalanobis-Wasserstein, viene ottenuta per ogni coppia (i, i') di osservazioni considerate, a partire dalla matrice di covarianza $\Sigma_{p \times p}$ (con p pari al numero di variabili a intervallo considerate), opportunamente calcolata rispetto alla tipologia di dati trattati ed è così definita:

$$d_{MW}(\boldsymbol{F_i}, \boldsymbol{F_{i'}}) = \left(\sum_{h=1}^{p} \sum_{k=1}^{p} \int_0^1 s_{hk}^{-1} \left(F_{ih}^{(-1)}(t) - F_{i'k}^{(-1)}(t) \right)^2 dt \right)^{\frac{1}{2}}$$

Per costruzione quindi, questa misura tiene conto della dipendenza lineare tra le variabili, sotto l'ipotesi di indipendenza delle distribuzioni marginali di ciascuna coppia (i, i') di osservazioni.

3.3 Distanze e Teoria dell'Informazione

Nelle prossime sezioni verranno brevemente illustrati alcuni concetti fondamentali della cosiddetta *Teoria dell'Infomazione*, una scienza nata a metà del '900 grazie agli studi intrapresi da Shannon [Shannon and Weaver, 1949] relativamente alla comunicazione dei segnali su un canale in presenza di rumore. Inizialmente incentrata sullo studio delle caratteristiche matematiche e probabilistiche dei sistemi di comunicazione, tale teoria è stata rapidamente estesa ad altri ambiti, e a tutt'oggi si inserisce in numerosi contesti applicativi. Sulla base dei concetti di seguito riportati sono state proposte diverse misure di dissimilarità e distanza tra distribuzioni, tra cui quella di Jensen-Shannon, illustrata nella parte finale della presente sezione.

3.3.1 L'Entropia

Concetto fondamentale nella Teoria dell'Informazione è l'*Entropia*. L'Entropia è una misura dell'incertezza associata ad una variabile casuale. In termini formali, data una variabile casuale discreta X, l'Entropia è definita come l'aspettativa, cambiata di segno, del logaritmo della funzione di probabilità $f(x)$:

$$H(X) = -E\{logf(x)\} = -\sum_{i=1}^{n} f(x_i) \log f(x_i). \tag{3.2}$$

Si può dimostrare [Shannon and Weaver, 1949] che è l'unica funzione delle realizzazioni di una variabile casuale (a meno di una costante positiva) che soddisfa i seguenti postulati:

- $H(X)$ è una funzione continua;
- dati eventi equiprobabili, con $f(x_i) = 1/n$ allora $H(X)$ è una funzione strettamente crescente di n;

- $H(X)$ è additiva: se l'evento finale può essere visto come il risultato della realizzazione di più eventi consecutivi, l'entropia globale deve essere pari alla somma pesata delle entropie corrispondenti a ciascuno step intermedio.

La funzione $H(X)$ gode inoltre di altre proprietà:

- $H(X) \geqslant 0$ dove l'uguaglianza vale solo se $f(x_i) = 1$ per un qualche valore di i;
- $H(X)$ assume valore massimo pari a $\log n$ quando gli eventi sono equiprobabili;
- se si considerano due variabili casuali X e Y è possibile calcolare l'Entropia congiunta $H(X, Y)$ come l'aspettativa, cambiata di segno, del logaritmo della funzione di probabilità congiunta:

$$H(X, Y) = -E[\log f(x, y)] = -\sum_x \sum_y f(x, y) \log f(x, y);$$

- Se si considera la distribuzione di probabilità condizionata $f(x|y)$, si può definire l'entropia condizionata:

$$H(X|Y) = -\sum_y f(y) \sum_x f(x|y) \log f(x|y);$$

Inoltre, si può verificare che vale la seguente relazione (*Chain rule*):

$$H(X, Y) = H(X|Y) + H(Y).$$

Inizialmente definito solo per variabili casuali discrete, il concetto di entropia è stato successivamente generalizzato per le cosiddette sorgenti di informazione continue. Considerando quindi una variabile casuale X con funzione di ripartizione $F(X) = Pr(X \leqslant x)$ e funzione di densità $f(x)$, si definisce *Entropia Differenziale* la seguente:

$$H(X) = -E\{logf(x)\} = -\int_X f(x) \cdot \log f(x)dx \tag{3.3}$$

Quest'ultima, anch'essa legata al concetto di incertezza circa la variabile casuale X, è per molti aspetti simile all'Entropia definita dalla (3.2), ma anche se sono strettamente legate [Cover and Thomas, 2006], le due cose non coincidono. Infatti, supponendo di dividere il supporto della variabile casuale X in intervalli di lunghezza Δ e considerando la quantizzazione X^Δ della variabile casuale X, così definita:

$$X^\Delta = x_i \quad se \quad i\Delta \leqslant X < (i+1)\Delta$$

(dove i indica di volta in volta l'intervallo considerato) si ha:

$$H\left(X^\Delta\right) = -\sum \Delta f(x_i) \log f(x_i) - \log \Delta$$

dato che la probabilità che la variabile casuale X^Δ sia pari ad x_i è pari ad $f(x_i)\Delta$. In aggiunta, se la funzione $f(x)$ è integrabile, si dimostra che

$$\lim_{\Delta \to 0}[H\left(X^\Delta\right) + \log \Delta] = \lim_{\Delta \to 0} -\sum \Delta f(x_i) \log f(x_i) = H(X)$$

Premesso ciò, è da sottolineare che, analogamente a quanto fatto per variabili casuali discrete, possono essere definite le Entropie differenziali congiunte e condizionate, considerando le opportune funzioni di densità.

Si vuole infine sottolineare una differenza fondamentale tra le entropie, calcolate per variabili casuali discrete, e le entropie differenziali. Queste ultime possono assumere valore negativo, a differenza delle prime: infatti se $f(x_i)$ è una funzione di densità, il suo logaritmo può assumere anche valori positivi, diversamente da quanto accade quando si considerano funzioni di probabilità, i cui valori sono sempre al più pari ad 1.

3.3.2 Entropia Relativa e Mutual Information

Altro concetto fondamentale nella Teoria dell'Informazione è l'*Entropia Relativa* o *divergenza di Kullback-Leibler* (d_{KL}). L'Entropia Relativa è una misura della divergenza tra due distribuzioni che quantifica la perdita di informazione che si ha supponendo che la vera distribuizione della variabile casuale X sia $g(x)$ anzichè $f(x)$.

In termini formali la d_{KL} è definita dalla seguente espressione:

$$d_{KL}(f|g) = E_f\left[log\frac{f(x)}{g(x)}\right] = \sum f(x)\log\frac{f(x)}{g(x)} \tag{3.4}$$

per variabili discrete, mentre

$$d_{KL}(f|g) = E_f\left[log\frac{f(x)}{g(x)}\right] = \int_X f(x)\log\frac{f(x)}{g(x)}dx \tag{3.5}$$

per variabili continue (ponendo $d_{KL} = \infty$ se $g(x) = 0$ e $d_{KL} = 0$ se $f(x) = g(x) = 0$).

La d_{KL} è sempre non negativa ed è nulla se e solo se $f(x) = g(x)$. La positività deriva direttamente dalla disuguaglianza di Jensen, secondo la quale, data una variabile casuale X si ha $E[f(X)] \geqslant f\left(E[X]\right)$ se e solo se $f(X)$ è strettamente concava.
Quindi, considerando la 3.5, si ha:

$$
\begin{aligned}
&-d_{KL}(f|g) = \\
&\quad = \int_X f(x) \log \frac{g(x)}{f(x)} dx \leqslant \text{ (per la disuguaglianza di Jensen)} \\
&\quad \leqslant \log \int_X f(x) \frac{g(x)}{f(x)} dx = \\
&\quad = \log \int_X g(x) dx = \log 1 = 0.
\end{aligned}
$$

(La dimostrazione è analoga per variabili casuali discrete).

Si evidenzia che la d_{KL} non è una misura simmetrica e inoltre non soddisfa la disuguaglianza triangolare. Di conseguenza non risulta essere una vera e propria metrica, anche se proprio a partire da essa sono state proposte numerose misure di distanza tra distribuzioni.

La d_{KL} assume un particolare significato quando, data una coppia di variabili casuali (X, Y) con funzione di probabilità o densità congiunta $f(x, y)$, si considera il caso in cui la funzione $g(x, y)$ sia pari al prodotto delle funzioni di densità marginali. In questo caso $g(x, y)$ altro non è che la funzione di densità congiunta qualora le variabili fossero indipendenti:

$$
I(X, Y) = \int_X \int_Y f(x, y) \cdot \log \frac{f(x, y)}{f(x) \cdot f(y)} dx dy. \tag{3.6}
$$

La d_{KL}, così calcolata, può essere interpretata come il gap di informazione che si ha supponendo erroneamente che le variabili siano indipendenti.

Tale quantità è centrale nella Teoria dell'Informazione e prende il nome di *Mutual Information* (MI). La MI è dunque connessa alla riduzione di incertezza che si ha relativamente ad una certa variabile casuale, conseguentemente alla conoscenza di una seconda variabile. Infatti, per costruzione essa è direttamente legata all'ammontare di

informazione che una variabile casuale contiene circa un'altra variabile casuale. L'espressione 3.6 può essere generalizzata considerando un vettore casuale. In questo caso, quanto più le componenti del vettore si avvicinano ad una situazione di reciproca indipendenza, tanto più la MI tenderà ad assumere valore nullo, stando ad indicare che la conoscenza del comportamento di una o più componenti del vettore non riduce l'incertezza riguardo il comportamento delle altre componenti.

Una definizione alternativa della MI può essere ottenuta partendo dall'entropia congiunta. Infatti si dimostra [Papoulis, 1991] che l'entropia congiunta è sempre al più pari alla somma delle entropie delle singole variabili. Ovvero, se si considerano due variabili casuali X e Y si ha:

$$H(X,Y) \leqslant H(X) + H(Y)$$

e l'uguaglianza vale solo nel caso in cui X e Y sono tra loro indipendenti. La differenza che intercorre tra i membri della precedente disuguaglianza è proprio pari alla MI.
Formalmente si ha:

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) = \\ &= -E[\log f(x)] - E[\log f(y)] + E[\log f(x,y)] = \\ &= E\left\{\log \frac{f(x,y)}{f(x)\cdot f(y)}\right\} \end{aligned}$$

E' chiaro come, anche dalla precedente espressione, può evidenziarsi il significato della MI finora espresso, in termini di differenza di entropie (marginali e congiunta) e dunque di riduzione di incertezza.

Nella figura 3.1 è mostrata, tramite un diagramma di Venn, la relazione intercorrente tra l'entropia congiunta, le entropie marginali e la MI.

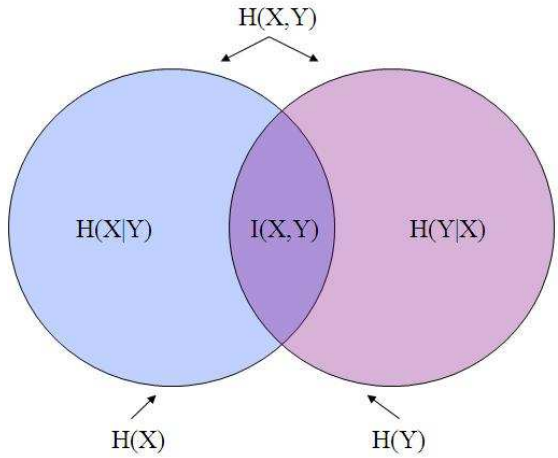

Figura 3.1: Relazione tra Entropia e MI

3.3.3 La d_{KL} e il Coefficiente J

Uno dei limiti della d_{KL} è la mancanza di simmetria, per cui non può essere considerata una misura di dissimilarità. Al fine di tener conto di questo, è stata proposta una sua versione simmetrica, il *coefficiente* J, anche detto *symmetrized* $\phi - divergence$:

$$J(f,g) = d_{KL}(f|g) + d_{KL}(g|f)$$

Si evidenzia come sia la d_{KL} che il coefficiente J possono essere ottenuti come casi particolari della $\phi - divergence$. La prima infatti si ottiene ponendo nella 3.1 $\phi(\lambda) = \lambda \log \lambda$, mentre la seconda si ha considerando $\phi(\lambda) = (\lambda - 1) \cdot \log(\lambda)$.

3.3.4 La divergenza di Jensen-Shannon

Tra le misure proposte al fine di quantificare la differenza tra due o più distribuzioni di probabilità, ce n'è una direttamente legata ai concetti finora descritti. E' la cosiddetta divergenza di Jensen-Shannon (d_{JS}), calcolata a partire dalla d_{KL}, come segue:

$$d_{JS}(f, g) = \pi \cdot d_{KL}(f|m) + (1 - \pi) \cdot d_{KL}(g|m) \tag{3.7}$$

dove

$$m(x) = \pi \cdot f(x) + (1 - \pi) \cdot g(x)$$

è la mistura delle due funzioni $f(x)$ e $g(x)$ di densità (o probabilità) con $\pi \in [0, 1]$.

Quindi, quando si considera la media ponderata delle d_{KL} di ciascuna funzione dalla mistura delle stesse, si ottiene la cosiddetta d_{JS}, che a differenza della precedente è una misura di dissimilarità, dato che valgono le seguenti:

Proprietà

- $d_{JS}(f, g) \geqslant 0$.

  Questa proprietà è conseguenza diretta di quanto dimostrato in precedenza per la d_{KL}; infatti la d_{JS} risulta essere non negativa in quanto somma pesata di quantità positive o al più nulle;

- $d_{JS}(f, g) = 0$ se e solo se $f = g$.

  Infatti è solo in questo caso che $f = g = m$ e conseguentemente $d_{KL}(f|m) = d_{KL}(g|m) = 0$;

- è simmetrica, ovvero $d_{JS}(f, g) = d_{JS}(g, f)$.

Queste proprietà la rendono una misura di dissimilarità ben definita.

Affinchè si possa parlare di metrica è necessario che la misura considerata soddisfi la disuguaglianza triangolare. E' stato dimostrato che la d_{JS} non è una vera e propria metrica, anche se lo è la sua radice quadrata [Endres and Schindelin, 2003]. Quindi può essere opportuno nelle applicazioni considerare questo risultato.

Anche la d_{JS}, essendo calcolata a partire dalla d_{KL}, è legata al concetto dell'entropia. Infatti si dimostra che la 3.7 può essere riscritta come segue:

$$\begin{aligned} d_{JS}(f,g) &= H(\pi \cdot f + (1-\pi)\cdot g) - \pi H(f) - (1-\pi)H(g) = \\ &= H(m) - \pi H(f) - (1-\pi)H(g) \end{aligned}$$

Inoltre, la precedente può essere estesa al caso di n funzioni di densità (o probabilità) $f_1, \ldots, f_n$:

$$\begin{aligned} d_{JS}(f_1,\ldots,f_n) &= H\left(\sum_{i=1}^{n} \pi_i \cdot f_i\right) - \sum_{i=1}^{n} \pi_i \cdot H(f_i) = \\ &= H(m) - \sum_{i=1}^{n} \pi_i \cdot H(f_i) \end{aligned} \tag{3.8}$$

dove $\pi_i \in [0,1]\ \forall i$ e $\sum_{i=1}^{n} \pi_i = 1$.

Quanto finora detto, così come accade per la d_{KL}, può essere ridefinito per i vettori casuali, considerando la d_{JS} tra le funzioni di densità congiunte.

Di seguito si dimostrerà che anche la d_{JS} tra n funzioni di densità congiunte può essere calcolata a partire dalle d_{KL} di ogni singola funzione dalla mistura delle stesse. Questo risultato risulterà utile nel prosieguo, allorchè verranno dimostrate altre affermazioni nel contesto della cluster analysis.

Teorema 1. *Siano $F_1, \ldots, F_n$ n funzioni di ripartizione congiunte appartenenti allo spazio $M(\Omega, A)$, con σ-algebra A definita sullo spazio campionario Ω, e siano $f_1, \ldots, f_n$ le rispettive funzioni di densità. La d_{JS} tra le n funzioni di densità $f_1, \ldots, f_n$ è pari alla media ponderata delle d_{KL} di ciascuna funzione dalla mistura di tutte le funzioni:*

$$d_{JS}(f_1, \ldots, f_n) = \sum_{i=1}^{n} \pi_i d_{KL}(f_i|m)$$

con $m = \sum_{i=1}^{m} \pi_i \cdot f_i$ e pesi $\pi_1, \ldots, \pi_n$.

Dimostrazione.

$$\begin{aligned}
&d_{JS}(f_1, \ldots, f_n) = \text{ (per la 3.8)} \\
&= H\left(\sum_{i=1}^{n} \pi_i \cdot f_i(\boldsymbol{x})\right) - \sum_{i=1}^{n} \pi_i \cdot H(f_i(\boldsymbol{x})) = \\
&= \int_{\boldsymbol{X}} \sum_{i=1}^{n} \pi_i \cdot f_i(\boldsymbol{x}) \cdot \ln \frac{1}{\sum_{i=1}^{n} \pi_i \cdot f_i(\boldsymbol{x})} d\boldsymbol{x} + \\
&+ \sum_{i=1}^{n} \pi_i \cdot \int_{\boldsymbol{X}} f_i(\boldsymbol{x}) \cdot \ln f_i(\boldsymbol{x}) d\boldsymbol{x} = \\
&= \int_{\boldsymbol{X}} \sum_{i=1}^{n} \pi_i \cdot f_i(\boldsymbol{x}) \ln \frac{1}{m(\boldsymbol{x})} d\boldsymbol{x} + \sum_{i=1}^{n} \int_{\boldsymbol{X}} \pi_i \cdot f_i(\boldsymbol{x}) \ln f_i(\boldsymbol{x}) d\boldsymbol{x} = \\
&= \int_{\boldsymbol{X}} \sum \pi_i \cdot f_i(\boldsymbol{x}) \left[\ln \frac{1}{m(\boldsymbol{x})} + \ln f_i(\boldsymbol{x})\right] d\boldsymbol{x} = \\
&= \int_{\boldsymbol{X}} \sum \pi_i \cdot f_i(\boldsymbol{x}) \left[\ln \frac{f_i(\boldsymbol{x})}{m(\boldsymbol{x})}\right] d\boldsymbol{x} = \\
&= \sum_{i=1}^{n} \pi_i \cdot d_{KL}(f_i|m)
\end{aligned}$$

□

Si sottolinea infine che le proprietà enunciate precedentemente per il caso di due sole funzioni di densità valgono anche nel caso in cui si considera la d_{JS} per un generico numero n di funzioni di densità.

3.4 Altre proposte

Verranno di seguito riportate due nuove proposte di dissimilarità che potrebbero risultare utili nell'ottica della classificazione di oggetti descritti da distribuzioni. Tali proposte rappresentano a tutt'oggi solo il risultato di alcune riflessioni in merito agli argomenti trattati nel presente lavoro di tesi e non verranno considerate nel prosieguo. Vengono ugualmente riportate in quanto potrebbero rappresentare uno spunto per ulteriori percorsi di ricerca.

3.4.1 Conditional Mahalanobis-Wasserstein

Quando si hanno elementi per supporre che una certa variabile (Y) influenzi una o più variabili (ipotesi di dipendenza asimmetrica), oppure quando si considera, per la descrizione degli n elementi dell'insieme E anche una variabile categorica, si può pensare di valutare la distanza di Mahalanobis-Wasserstein tra le distribuzioni condizionate di X dato Y, ovvero tra $F(X|Y)$ e $G(X|Y)$. In questo caso F e G possono essere invertite per considerare la differenza tra i quantili delle distribuzioni:

$$d_W(F,G) = \left(\int_0^1 \left(F^{(-1)}(t|y) - G^{(-1)}(t|y) \right)^2 dt \right)^{\frac{1}{2}}$$

Quindi si procede come precedentemente visto calcolando la matrice di varianza e covarianza e la distanza di Mahalanobis-Wasserstein:

$$d_{MW}(\boldsymbol{F_i}, \boldsymbol{F_{i'}}) = \left(\sum_{h=1}^{p} \sum_{k=1}^{p} \int_0^1 s_{hk}^{-1} \left(F_{ih}^{(-1)}(t|y) - F_{i'k}^{(-1)}(t|y) \right)^2 dt \right)^{\frac{1}{2}}.$$

La variabile Y assumerà un ruolo prioritario nel processo di clustering, venendosi a configurare come variabile di classificazione.

3.4.2 Weighted - Wasserstein

Si è già detto che la MI è una quantità che misura la dipendenza reciproca tra due variabili e può essere interpretata come l'informazione relativa ad X contenuta in Y (o equivalentemente l'informazione relativa ad Y contenuta in X). Si può perciò pensare di pesare la distanza di Wasserstein con gli elementi della matrice di MI, al fine di tener conto non solo della dipendenza di tipo lineare, ma anche di quella non lineare:

$$d_{WW}(\boldsymbol{F_i}, \boldsymbol{F_{i'}}) = \left(\sum_{h=1}^{p} \sum_{k=1}^{p} \int_0^1 i_{hk}^{-1} \left(F_{ih}^{-1}(t) - F_{i'h}^{-1}(t) \right)^2 dt \right)^{\frac{1}{2}}$$

Si può facilmente verificare che la d_{WW} è una misura di dissimilarità. Infatti:

- $d_{WW}(\boldsymbol{F_i}, \boldsymbol{F_{i'}}) \geqslant 0$;
- $d_{WW}(\boldsymbol{F_i}, \boldsymbol{F_{i'}}) = 0$ se e solo se $\boldsymbol{F_i} = \boldsymbol{F_{i'}}$
  (ovvero se $F_{ij} = F_{i'j} \ \ \forall i = 1, \ldots, n$ e $\forall j = 1, \ldots, p$);
- $d_{WW}(\boldsymbol{F_i}, \boldsymbol{F_{i'}}) = d_{WW}(\boldsymbol{F_{i'}}, \boldsymbol{F_i})$.

Come nel caso della d_{MW}, la distanza tra i due vettori $\mathbf{F}_i$ ed $\mathbf{F}_{i'}$ sarà inversamente proporzionale alla dipendenza tra le variabili considerate. In questo caso però, a differenza del precedente, verrà tenuta in considerazione anche la dipendenza non-lineare.

Capitolo 4

Classificazione di dati descritti da distribuzioni multivariate

4.1 Introduzione

Il presente capitolo sarà dedicato al problema della classifcazione di dati non standard, e più nello specifico ai metodi di clustering non gerarchici per oggetti descritti da distribuzioni. Verranno presentati i due approcci principali: l'algoritmo di clustering dinamico e la classificazione dinamica su tabelle di distanza. Successivamente verranno presentati alcuni strumenti per l'interpretazione dei risultati. Nell'ultima parte del capitolo verrà descritto l'algoritmo di clustering dinamico basato sulla distanza di Jensen-Shannon e sull'utilizzo delle funzioni copula.

4.2 L'algoritmo di clustering dinamico

L'algoritmo di clustering dinamico (DCA) [Diday, 1971] è un algoritmo non gerarchico di tipo iterativo che ha come obiettivo quello di suddividere un insieme di oggetti $\omega_i \in E$ in gruppi omogenei. Nel contempo lo scopo è anche quello di fornire un'appropriata rappresentazione di ciascun cluster al fine di avere una conoscenza tale da poter allocare eventuali nuovi oggetti.

4.2.1 L'Algoritmo

Alcune caratteristiche generali di questo algoritmo sono il numero prefissato k di clusters, e la presenza di due passi successivi: il passo di allocazione e il passo di rappresentazione. Come vedremo, il passo di allocazione richiede la specificazione di una funzione di allocazione, che è direttamente legata alla misura di dissimilarità prescelta, mentre il passo di rappresentazione richiede che sia definita una modalità di descrizione sintetica di ciascun cluster attraverso un elemento 'centrale' o un 'prototipo'. In particolare, va detto che il concetto di prototipo come elemento rappresentativo del cluster è stato introdotto di recente nel contesto dell'estensione del DCA a dati simbolici [de Carvalho et al., 2008]. Nel presente lavoro, si continuerà a far riferimento ai prototipi, in quanto i dati trattati sono di tipo non puntuale, perchè descritti da distribuzioni, e in quanto tali associabili al concetto di dati simbolici.

L'algoritmo si sviluppa quindi secondo lo schema seguente:

- *Inizializzazione*:
  Si considera una partizione iniziale degli elementi di E, ottenuta in maniera casuale. Alternativamente possono essere scelti, sempre casualmente, k elementi di E, ma in questo caso, per inizializzare l'algoritmo, è richiesto un passo di allocazione, durante il quale vengono formati i clusters iniziali considerando

la prossimità di ciascun elemento di E da quelli casualmente individuati.

- *Passo di rappresentazione*:
  è volto all'identificazione di un oggetto rappresentativo di ciascun cluster, detto anche centro o 'prototipo'. Data una partizione P degli elementi di E in k classi $(C_1, \ldots, C_k)$, verrà identificato un vettore $L = (G_1, \ldots, G_k)$ di oggetti che possano in qualche modo sintetizzare e descrivere gli elementi appartenenti a ciascuna classe.

  Per costruzione, il centro G_h della classe C_h sarà tale se minimizza il seguente criterio:

$$f_{C_h}(G_h) = \sum_{\omega_i \in C_h} D(\omega_i, G_h), G_h \in \Lambda \tag{4.1}$$

  dove $D(\omega_i, G_h)$ è una misura di dissimilarità tra l'oggetto considerato e il prototipo dell'$h - esimo$ cluster e Λ è lo spazio dei prototipi.

- *Passo di allocazione*:
  è diretto alla costruzione della partizione degli elementi dell'insieme E, mediante l'attribuzione di ciascun elemento ad un cluster. In questo passo ciascun oggetto sarà assegnato alla classe il cui prototipo risulta essere più prossimo. La prossimità viene stabilita secondo una predeterminata funzione di allocazione ψ.

- *Regola di arresto*:
  l'algoritmo si arresta quando la partizione identificata al generico step t è uguale a quella trovata allo step precedente ($P^{(t)} = P^{(t-1)}$).

Nel complesso dunque, l'algoritmo ha l'obiettivo di partizionare gli oggetti in maniera ottimale, secondo un criterio predefinito $\Delta(P, L)$

che misura l'adattamento tra i clusters individuati e i rispettivi prototipi, in modo che sia massimizzata l'omogeneità tra gli elementi appartenenti al medesimo cluster e tale che:

$$\Delta(P, L) = \sum_{h=1}^{k} \sum_{\omega_i \in C_h} D(\omega_i, G_h). \tag{4.2}$$

In termini più formali si può dire che il DCA è diretto alla ricerca di una partizione $P^* = (C_1, \ldots, C_k)$ dell'insieme E in k cluster e, contemporaneamente di un vettore $L^* = (G_1, \ldots, G_k)$ di k prototipi tale che sia ottimizzata la funzione criterio Δ:

$$\Delta(P^*, L^*) = \min\{\Delta(P, L) | P \in P_k, L \in L_k\}$$

dove P_k rappresenta l'insieme di tutte le possibili partizioni di dimensione k di E ed L_k è l'insieme di tutti i possibili vettori di prototipi.

4.2.2 Condizioni di convergenza

In generale si può affermare che, se la funzione $\Delta(P, L)$ può essere definita rispetto alla funzione di allocazione ψ in maniera additiva, come segue:

$$\Delta(P, L) = \sum_{h=1}^{k} \sum_{\omega_i \in C_h} \psi(\omega_i, G_h)$$

l'algoritmo converge se, per ogni classe C_h esiste ed è unico il prototipo G_h. In questo caso, la funzione criterio $\Delta(P, L)$ decresce sia ad ogni passo di allocazione, sia ad ogni passo di rappresentazione. Infatti, ad ogni generica iterazione t verrà individuata una nuova coppia $(P^{(t)}, L^{(t)})$, costituita dalla partizione corrente e dal corrispondente vettore di prototipi, in modo che il valore della funzione $\Delta(P^{(t)}, L^{(t)})$ sia inferiore al valore $\Delta(P^{(t-1)}, L^{(t-1)})$, ottenuto al passo precedente.

Al fine di garantire il decremento della funzione Δ ad ogni passo devono però essere soddisfatte alcune condizioni circa l'esistenza e l'unicità del prototipo e del cluster di appartenenza per ogni oggetto ω_i.

Relativamente all'unicità del cluster di appartenenza, non vi sono in generale problemi particolari: l'unico inconveniente potrebbe essere rappresentato dalla presenza di due clusters tali che la distanza di ω_i da ciascuno dei corrispondenti prototipi risulti identica. In questo caso si può eliminare l'indecisone ad esempio scegliendo a priori di attribuire l'oggetto al cluster avente indice inferiore.

L'esistenza e l'unicità del prototipo pone invece maggiori problemi. Infatti, la convergenza dell'algoritmo è garantita se esiste un unico prototipo G_h tale che sia minima la funzione criterio 4.1. Questa condizione può essere però difficile da dimostrare analiticamente e la valutazione deve essere fatta caso per caso a seconda della misura di dissimilarità prescelta.

Come si vedrà nel prosieguo, la misura di dissimilarità scelta per l'algoritmo che verrà proposto, consente di identificare un prototipo che soddisfa le condizioni suddette.

4.2.3 Il prototipo

Ne caso di dati non puntuali, sia la misura di prossimità, così come evidenziato nel precedente capitolo, che la definizione del prototipo assumono particolare rilevanza. Il problema è stato ampiamente affrontato nel contesto dell'analisi dei dati simbolici [de Carvalho et al., 2008]. Il prototipo, in quanto tale, deve generalizzare le caratteristiche degli elementi da partizionare e può essere un elemento dello spazio di rappresentazione degli stessi. In questo senso, dovrà essere coerente con la tipologia di oggetti trattati, essendo esso stesso consistente con la descrizione degli elementi di E. Nell'algoritmo proposto, ad esem-

pio, la classificazione verrà fatta su oggetti descritti da distribuzioni, e dunque anche il prototipo sarà descritto da una distribuzione.

Nell'algoritmo DCA, la misura di prossimità e il prototipo risultano strettamente legati. Infatti, come si è detto, il prototipo G_h di un cluster C_h è definito secondo il criterio $f_{C_h}(G_h)$ indicato nella 4.1, ma dato che la funzione $f_{C_h}(G_h)$ è basata sulla funzione di dissimilarità D scelta per confrontare ogni oggetto appartenente al cluster h e il corrispondente prototipo, quest'ultimo andrà definito e identificato proprio tramite la funzione D. Come vedremo in seguito, quindi, l'identificazione dell'elemento rappresentativo di ciascun cluster avverrà in maniera analitica a partire dalla misura di dissimilarità prescelta.

4.3 Classificazione dinamica su tabelle di distanza

A differenza del metodo precedente, in cui i dati di input sono gli elementi dell'insieme E, nel metodo di classificazione dinamica su tabelle di distanza (DCLUST) il punto di partenza è costituito da una matrice di dissimilarità o distanze. L'obiettivo è sempre quello di trovare una partizione degli oggetti in un numero predefinito di gruppi, ma in questo caso la procedura è interamente basata sulla prossimità di ogni coppia di individui. Anche la funzione criterio da ottimizzare sarà basata esclusivamente sulla somma delle dissimilarità tra gli elementi appartenenti allo stesso gruppo.

4.3.1 L'Algoritmo

L'algoritmo si sviluppa secondo lo schema seguente:

- *Inizializzazione:*
  Vengono scelti casualmente k elementi dell'insieme E, che costituiscono il vettore iniziale $L^{(0)} = (G_1^{(0)}, \ldots, G_k^{(0)})$ di prototipi;

- *Passo di allocazione*:
  Ciascun oggetto è assegnato alla classe il cui prototipo è più prossimo, secondo la funzione di allocazione ψ. Quindi l'oggetto ω_i sarà assegnato alla classe C_h se e solo se

$$h = \arg\min\{\psi(\omega_i, G_l^{(t-1)}) | l = 1, \ldots, k\};$$

- *Passo di rappresentazione*:
  come oggetto che rappresenta la classe si sceglierà quello che rende minima la somma delle distanze da ciascun oggetto appartenente alla classe stessa. Quindi il prototipo $G_h^{(t)}$ del cluster $C_h^{(t)}$ sarà l'oggetto ω_i se

$$i = \arg\min\{\sum_{\omega_l \in C_h^{(t)}} \psi(\omega_l, \omega_j) | \omega_j \in C_h^{(t)}; l \neq j\}. \qquad (4.3)$$

  Questa procedura equivale ad assumere come spazio di rappresentazione l'insieme degli elementi ω_i di E che minimizzano la funzione somma delle distanza degli elementi della classe C_h da ω_i. Lo spazio di rappresentazione è quindi dato da tutte le parti dell'insieme E costituite dagli elementi appartenenti a ciascun cluster.

- *Regola di arresto*:
  l'algoritmo si arresta quando la partizione identificata al generico step t è uguale a quella trovata allo step precedente ($P^{(t)} = P^{(t-1)}$).

4.3.2 Il prototipo

Una differenza sostianzale tra l'algoritmo DCA e la classificazione su tabelle di distanza risiede proprio nella definizione del prototipo. Nel primo caso infatti il prototipo generalmente non è un oggetto osservato, ovvero appartenente all'insieme E, ma è un oggetto 'fittizio' derivato dalla minimizzazione della funzione 4.1. Nel secondo caso, invece, il prototipo corrisponde sempre ad un oggetto osservato, essendo un oggetto scelto tra tutti gli oggetti da partizionare.

In generale, in questo algoritmo, l'identificazione del prototipo risulta di facile attuazione, non essendo basata su un procedimento analitico, ma semplicemente sulla matrice di distanze di input. Inoltre non si pone il problema della convergenza dell'algoritmo, in quanto esiste sempre un oggetto per cui risulta soddisfatta la 4.3.

4.4 Interpretazione dei risultati: la bontà della partizione

Gli strumenti classici per la valutazione della bontà della partizione ottenuta durante il processo di classificazione sono generalmente basati sul criterio dell'inerzia [Celeux et al., 1989], secondo cui se si suppone di poter decomporre l'inerzia totale in inerzia entro i gruppi ed inerzia tra i gruppi, si può considerare tanto più 'buona' una partizione, quanto più la componente dovuta all'inerzia entro i gruppi è piccola rispetto all'inerzia totale.

E' possibile generalizzare tale concetto al caso di dati non standard [de Carvalho et al., 2008]. Infatti, avendo definito una misura di dissimilarità D, è possibile calcolare il valore della funzione 4.2, la quale può essere considerata una misura di inerzia entro i cluster. L'inerzia totale invece sarà data dal valore assunto dalla funzione $f_E(G)$, con

$$f_E(G) = \sum_{i=1}^{n} D(\omega_i, G)$$

dove G indica il prototipo generale.

Dato che l'identificazione di una partizione degli elementi di E permette di poter descrivere l'intero insieme E attraverso il vettore dei prototipi $(G_1, \ldots, G_k)$, è interessante valutare la differenza di inerzia che si ottiene considerando come rappresentazione degli oggetti il prototipo generale G piuttosto che i k prototipi $(G_1, \ldots, G_k)$ della partizione P.

Tale differenza non è mai negativa, in quanto per costruzione si ha:

$$\sum_{h=1}^{k} f_{C_h}(G_h) \leqslant \sum_{h=1}^{k} f_{C_h}(G)$$

dove

$$\sum_{h=1}^{k} f_{C_h}(G_h)$$

non è altro che la funzione criterio $\Delta(P, L)$ e

$$\sum_{h=1}^{k} f_{C_h}(G)$$

è la funzione $f_E(G)$.

Il confronto tra le due quantità sopra riportate può essere interpretato come comparazione tra l'ipotesi 'Assenza di struttura', o equivalentemente 'Presenza di un unico cluster', e l'ipotesi 'Partizione nelle k classi' identificate minimizzando la funzione criterio Δ. Infatti esso è basato sulla valutazione del grado di omogeneità degli oggetti

ottenuto, una volta, considerando un unico centro, e, un'altra volta, considerando i k prototipi $G_1, \ldots, G_k$.

In definitiva gli indicatori che verranno utilizzati per interpretare la partizione saranno:

- $f_{C_h}(G_h)$ come misura di omgeneità del cluster C_h;
- $\Delta(P, L)$ come misura di omogeneità entro i clusters della partizione P;
- $f_E(G)$ quale una misura di omogeneità totale degli oggetti appartenenti all'insieme E.

Come misura globale di bontà della partizione ottenuta può perciò essere considerata la seguente quantità:

$$Q(P) = 1 - \frac{\sum_{h=1}^{k} f_{C_h}(G_h)}{\sum_{h=1}^{k} f_{C_h}(G)} = 1 - \frac{\Delta(P, L)}{f_E(G)}$$

Quest'indice varia tra 0 e 1. Sarà pari a 0 quando tutti gli oggetti coincidono con il prototipo generale G. Infatti in questo caso si ottiene:

$$f_{C_h}(G_h) = f_{C_h}(G)$$

e dunque

$$\Delta(P, L) = f_E(G).$$

L'indice $Q(P)$ segnalerà in questo caso particolare la presenza di un unico cluster contenente tutti gli oggetti dell'insieme E, tra loro coincidenti e perciò rappresentati esattamente dal prototipo G.

Al contrario, il caso in cui $Q(P) = 1$ si ottiene quando i clusters sono costituiti da oggetti identici tra loro, con riferimento alla dissimilarità

considerata, ma differenti da cluster a cluster. Ciascun oggetto coinciderà allora con il prototipo del cluster di appartenenza e la funzione $f_{C_h}(G_h)$ assumerà valore nullo per ogni $h = 1, \ldots, k$. Di conseguenza la partizione individuata sarà una partizione perfetta in quanto ciascun oggetto sarà rappresentato esattamente dal corrispondente prototipo.

Dunque l'indice $Q(P)$ misura la parte di omogeneità dell'insieme E spiegata dalla partizione P, potendosi interpretare come il rapporto tra inerzia tra i gruppi e inerzia totale.

Va comunque sottolineato che la misura $Q(P)$ è crescente rispetto al numero di clusters e quindi la stessa non può essere utilizzata per fare valutazioni circa il numero ottimale di clusters da considerare, oppure per confrontare partizioni di dimensioni diverse.

4.5 La divergenza di Jensen-Shannon nell'algoritmo di clustering dinamico

Il metodo di clustering dinamico proposto in questo lavoro di tesi è volto alla classificazione di oggetti descritti da distribuzioni multivariate. In quanto tale, vi è innanzitutto l'esigenza di individuare una misura di dissimilarità, tra quelle descritte nel precedente capitolo, che possa fornire indicazioni relativamente alla prossimità degli oggetti suddetti.

Per le sue proprietà e le caratteristiche che verranno evidenziate di seguito, si è optato per l'utilizzo della divergenza di Jensen-Shannon. Verranno quindi ridefiniti i concetti introdotti finora in termini di tale distanza. Saranno poi riportati alcuni risultati originali, relativi alle condizioni di unicità che garantiscono la convergenza dell'algoritmo DCA.

4.5.1 Definizione

Come si è detto, la divergenza di Jensen-Shannon è una misura della discrepanza tra due funzioni di densità o probabilità e si può ottenere in maniera additiva dalla divergenza di Kullback-Leibler. Proprio per la sua natura dunque, può essere impiegata nell'algoritmo DCA quando gli elementi dell'insieme E sono descritti da distribuzioni, e più in particolare, quando si vuole considerare anche la dipendenza tra le variabili casuali in esame, pervenendo ad un confronto tra le distribuzioni di densità congiunte utilizzate per descrivere gli oggetti $\omega_i \in E$.

Questa misura di divergenza, consente infatti non solo di individuare una funzione di allocazione utile ai fini del partizionamento degli oggetti, ma ammette l'esistenza di un prototipo per la descrizione di ciascuna classe, che come vedremo risulta essere unico. Insieme, queste caratteristiche garantiscono la convergenza dell'algoritmo DCA.

Si considerino $\omega_1, \ldots, \omega_n$ oggetti costituenti l'insieme E. Sia T un data-set contenente n righe e p colonne. Si supponga che l'$i-esima$ riga $(i = 1, \ldots, n)$ corrisponda ad un oggetto ω_i e si consideri un vettore di variabili casuali $X^{(1)}, \ldots, X^{(p)}$ di interesse per la descrizione di tale oggetto. Con riferimento alle suddette variabili casuali, si supponga che ogni cella di T contenga una distribuzione marginale $F_i^{(j)}$ $(j = 1, \ldots, p)$ e si assuma di avere informazioni aggiuntive circa la relazione di dipendenza tra le variabili considerate, in modo tale da poter ottenere, per ogni oggetto di E, la corrispondente funzione di densità congiunta f_i, sempre con riferimento al vettore $X^{(1)}, \ldots, X^{(p)}$.

Con riferimento alla 4.1, si consideri quale funzione $f_{C_h}(G_h)$ la somma pesata delle divergenza di Kullback-Leibler tra ciascuna funzione di densità congiunta corrispondente al singolo oggetto appartenente al cluster C_h e la funzione di densità congiunta m_h corrispondente al prototipo:

$$f_{C_h}(G_h) = \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} d_{KL}(f_i|m_h)$$

con

$\pi^{(h)} = \sum_{\omega_i \in C_h} \pi_i$

L'aver definito la funzione $f_{C_h}(G_h)$ consente l'individuazione immediata della funzione di allocazione $\psi(\omega_i, G_h)$. Infatti questa sarà data dalla divergenza di Kullback-Leibler $d_{KL}(f_i|m_h)$ moltiplicata per il corrispondente peso $\frac{\pi_i}{\pi^{(h)}}$. Ciò significa che, durante il passo di allocazione, ciascun oggetto sarà assegnato al cluster il cui prototipo risulta più prossimo in termini di divergenza di Kullback-Leibler.

Allora anche la funzione criterio $\Delta(P, L)$ può essere espressa in termini di divergenza di Kullback-Leibler, come segue:

$$\Delta(P, L) = \sum_{h=1}^{k} \sum_{\omega_i \in C_h} \pi_i d_{KL}(f_i|m_h)$$

L'aver assunto quale funzione criterio 4.1 la somma pesata delle divergenze di Kullback-Leibler consente, come vedremo nella sezione successiva, anche l'identificazione di un prototipo per ciascuna classe.

4.5.2 L'individuazione del prototipo

Al fine di identificare il prototipo G_h di ciascuna classe C_h va innanzitutto sottolineato che, per costruzione, se esiste, esso sarà l'oggetto a cui corrisponde la funzione di densità congiunta $m_h(\boldsymbol{x})$ tale che sia minimo il criterio di adattamento:

$$f_{C_h}(G_h) = \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} \cdot d_{KL}(f_i|m_h)$$

Come verrà tra breve dimostrato, tale funzione non solo esiste ma è unica. Infatti l'unica funzione che rende minima tale quantità è la mistura delle funzioni di densità congiunte.

Teorema 2. *Si consideri un gruppo C_h di elementi di E. Allora la quantità*

$$m_h(\boldsymbol{x}) = \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} f_i(\boldsymbol{x})$$

rende minima la funzione $f_{C_h}(G_h)$ definita dalla 4.1, ovvero:

$$\sum_{i \in C_h} \frac{\pi_i}{\pi^{(h)}} d_{KL}(f_i|s) \geqslant \sum_{i \in C_h} \frac{\pi_i}{\pi^{(h)}} d_{KL}(f_i|m_h)$$

$\forall s(\boldsymbol{x}) \neq m_h(\boldsymbol{x})$

Dimostrazione.

$$\begin{aligned}
&\sum_{i \in C_h} \frac{\pi_i}{\pi^{(h)}} d_{KL}(f_i|s) - \sum_{i \in C_h} \frac{\pi_i}{\pi^{(h)}} d_{KL}(f_i|m_h) = \\
&= \sum_{i \in C_h} \frac{\pi_i}{\pi^{(h)}} \left[d_{KL}(f_i|s) - d_{KL}(f_i|m_h) \right] = \\
&= \sum_{i \in C_h} \frac{\pi_i}{\pi^{(h)}} \left[\int_{\boldsymbol{x}} f_i(\boldsymbol{x}) \cdot \ln \frac{f_i(\boldsymbol{x})}{s(\boldsymbol{x})} d\boldsymbol{x} - \int_{\boldsymbol{x}} f_i(\boldsymbol{x}) \cdot \ln \frac{f_i(\boldsymbol{x})}{m_h(\boldsymbol{x})} d\boldsymbol{x} \right] = \\
&= \sum_{i \in C_h} \frac{\pi_i}{\pi^{(h)}} \left[\int_{\boldsymbol{x}} f_i(\boldsymbol{x}) \cdot \ln \left(\frac{f_i(\boldsymbol{x})}{s(\boldsymbol{x})} \cdot \frac{m_h(\boldsymbol{x})}{f_i(\boldsymbol{x})} \right) d\boldsymbol{x} \right] = \\
&= \sum_{i \in C_h} \frac{\pi_i}{\pi^{(h)}} \left[\int_{\boldsymbol{x}} f_i(\boldsymbol{x}) \cdot \ln \frac{m_h(\boldsymbol{x})}{s(\boldsymbol{x})} d\boldsymbol{x} \right] = \int_{\boldsymbol{x}} m_h(\boldsymbol{x}) \ln \frac{m_h(\boldsymbol{x})}{s(\boldsymbol{x})} d\boldsymbol{x} = \\
&= d_{KL}(m_h|s) \geqslant 0 \quad \forall s \; e \; \forall \boldsymbol{x}
\end{aligned}$$

e l'uguaglianza vale se e solo se $m_h(\boldsymbol{x}) = s(\boldsymbol{x})$ $\forall(\boldsymbol{x})$. □

Quindi la mistura delle densità corrispondenti agli oggetti appartenenti al generico cluster h è quell'elemento che minimizza la funzione criterio 4.1. Ma per definizione, quando si considera la mistura delle funzioni considerate, la 4.1 coincide con la divergenza di Jensen-Shannon tra le funzioni appartenenti al cluster C_h. Quindi l'algoritmo sarà diretto alla ricerca della partizione $P = (C_1, \ldots, C_k)$ e contestualmente delle misture $m_1, \ldots, m_k$, tali che sia minimizzata la divergenza di Jensen-Shannon entro i cluster.

Questa proprietà è di notevole rilevanza, in quanto garantisce la convergenza dell'algoritmo DCA nel momento in cui si prende come elemento rappresentativo di ciascun cluster la mistura delle densità apparteneti al cluster stesso.

Può essere dimostrato che ottimizzare la funzione criterio Δ precedentemente definita è equivalente a massimizzare la divergenza di Jensen-Shannon tra i clusters (d_{JS}^B) e contestualmente minimizzare la divergenza di Jensen-Shannon entro i clusters (d_{JS}^W). Infatti una delle proprietà di questa misura è che la divergenza totale tra gli oggetti considerati, può essere decomposta in due quantità, la prima relativa alla dissimilarità tra gli oggetti appartenenti allo stesso cluster, e la seconda attribuibile alla differenza esistente tra i clusters.

Nel seguente teorema è ricavata la dimostrazione formale di quanto appena affermato.

Teorema 3. *Si consideri una partizione degli elementi E in k gruppi. Allora la divergenza di Jensen-Shannon tra n funzioni $f_1, \ldots, f_n$ può essere scomposta nella divergenza di Jensen-Shannon entro i gruppi (d_{JS}^W) e nella divergenza di Jensen-Shannon tra i gruppi (d_{JS}^B):*

$$d_{JS}^{TOT} = d_{JS}^W + d_{JS}^B \tag{4.4}$$

Dimostrazione.

$$
\begin{aligned}
&d_{JS}(f_1(\boldsymbol{x}),\ldots,f_n(\boldsymbol{x})) = \sum_{i=1}^{n} \pi_i \cdot d_{KL}(f_i|m) = \\
&= \sum_{h=1}^{k} \sum_{i\in C_h} \pi_i \cdot d_{KL}(f_i|m) = \\
&= \sum_{h=1}^{k} \sum_{i\in C_h} \pi_i \cdot \int_{\boldsymbol{x}} f_i(\boldsymbol{x}) \cdot \ln \frac{f_i(\boldsymbol{x})}{m(\boldsymbol{x})} d\boldsymbol{x} = \\
&= \sum_{h=1}^{k} \sum_{i\in C_h} \pi_i \cdot \int_{\boldsymbol{x}} f_i(\boldsymbol{x}) \cdot \ln \left(\frac{f_i(\boldsymbol{x})}{m_h(\boldsymbol{x})} \cdot \frac{m_h(\boldsymbol{x})}{m(\boldsymbol{x})} \right) d\boldsymbol{x} = \\
&= \sum_{h=i}^{k} \sum_{i\in C_h} \pi_i \cdot \int_{\boldsymbol{x}} f_i(\boldsymbol{x}) \cdot \ln \frac{f_i(\boldsymbol{x})}{m_h(\boldsymbol{x})} d\boldsymbol{x} + \sum_{h=i}^{k} \sum_{i\in C_h} \pi_i \cdot \int_{\boldsymbol{x}} f_i(\boldsymbol{x}) \cdot \ln \frac{m_h(\boldsymbol{x})}{m(\boldsymbol{x})} d\boldsymbol{x} = \\
&= \sum_{h=1}^{k} \sum_{i\in C_h} \pi_i \cdot d_{KL}(f_i|m_h) + \sum_{h=1}^{k} \int_{\boldsymbol{x}} \left(\sum_{i\in C_h} \pi_i \cdot f_i(\boldsymbol{x}) \right) \cdot \ln \frac{m_h(\boldsymbol{x})}{m(\boldsymbol{x})} d\boldsymbol{x} = \\
&= \sum_{h=1}^{k} \sum_{i\in C_h} \pi_i \cdot d_{KL}(f_i|m_h) + \sum_{h=1}^{k} \pi^{(h)} \cdot \int_{\boldsymbol{x}} m_h(\boldsymbol{x}) \cdot \ln \frac{m_h(\boldsymbol{x})}{m(\boldsymbol{x})} d\boldsymbol{x} = \\
&= \sum_{h=1}^{k} \sum_{i\in C_h} \pi_i \cdot d_{KL}(f_i|m_h) + \sum_{h=1}^{k} \pi^{(h)} \cdot d_{KL}\left(m_h|m\right) = \\
&= \sum_{h=1}^{k} \pi^{(h)} \sum_{i\in C_h} \frac{\pi_i}{\pi^{(h)}} \cdot d_{KL}(f_i|m_h) + \sum_{h=1}^{k} \pi^{(h)} \cdot d_{KL}\left(m_h|m\right) = \\
&= \sum_{h=1}^{k} \pi^{(h)} \cdot d_{JS}^{(h)} + \sum_{h=1}^{k} \pi^{(h)} \cdot d_{KL}\left(m_h|m\right) = \\
&= d_{JS}^{W} + d_{JS}^{B}
\end{aligned}
$$

□

Questo risultato, in accordo con la teoria classica, consentirà inoltre di ottenere alcuni strumenti utili per la valutazione della bontà della partizione ottenuta.

Infine va sottolineato che, benchè l'utilizzo della divergenza di Jensen-Shannon consenta di attribuire un peso diverso ad ogni funzione di densità e quindi di attribuire un'importanza differente a ciascun oggetto dell'insieme E, generalmente non sussistono motivi per farlo e quindi i pesi π_i possono essere supposti costanti e pari ad $1/n$. Nel prosieguo comunque, ci si riferirà sempre al caso generale, considerando un generico sistema di pesi $\pi_1, \ldots, \pi_n$.

4.5.3 La bontà della partizione

Visto il risultato ottenuto nel Teorema 3, è possibile ridefinire la misura di bontà della partizione $Q(P)$ come rapporto tra la d_{JS}^{B} e la d_{JS}^{TOT}, secondo l'espressione seguente:

$$Q(P) = \frac{d_{JS}^{B}}{d_{JS}^{TOT}} = \frac{\sum_{h=1}^{k} \pi^{(h)} \cdot d_{KL}\left(m_h | m\right)}{\sum_{i=1}^{n} \pi_i \cdot d_{KL}(f_i | m)} \tag{4.5}$$

Quindi $Q(P)$ sarà pari a zero se tutte le funzioni di densità considerate sono tra loro identiche e perciò identiche alla loro mistura, mentre sarà pari ad uno se gli oggetti appartenenti a ciascuno dei k clusters individuati saranno descritti da funzioni di densità tra loro uguali e dunque con mistura uguale, ma diverse da cluster a cluster. In questo ultimo caso, ogni classe sarà rappresentata esattamente dalla corrispondente mistura.

4.5.4 L'algoritmo in termini di funzione copula

E' stato più volte sottolineato come la divergenza di Jensen-Shannon, in quanto basata sulle funzioni di densità congiunte relative a ciascun oggetto dell'insieme E, permette di tenere in considerazione la dipendenza che intercorre tra le variabili considerate. La scelta di modellare tale dipendenza utilizzando la funzione copula ha come conseguenza anche quella di poter riscrivere la divergenza di Jensen-Shannon in termini di funzione copula e riconsiderare quindi l'intero algoritmo DCA.

Un primo risultato in questo senso riguarda la scomposizione dell'entropia congiunta. Infatti l'entropia $H(f_i)$ può essere decomposta in una prima parte attribuibile alle entropie marginali $H(f_i^{(j)})$ (per $j = 1, \ldots, p$) (dove $f_i^{(j)}$ indica la funzione di densità marginale corrispondente alla $j - esima$ variabile casuale considerata per l'oggetto $i - esimo$), e in una seconda parte dovuta alla Mutual Information [Papoulis, 1991].

E' facile verificare che la Mutual Information altro non è che l'entropia della funzione copula, cambiata di segno [1]:

$$
\begin{aligned}
I(X_1, \ldots, X_n) &= E\left[\log \frac{f_i(x_1, \ldots, x_n)}{f_i^{(1)}(x_1) \cdot \ldots \cdot f_i^{(n)}(x_n)}\right] = \\
&= E\left[\log \frac{c_i(F(x_1), \ldots, F(x_n)) \cdot f_i^{(1)}(x_1) \cdot \ldots \cdot f_i^{(n)}(x_n)}{f_i^{(1)}(x_1) \cdot \ldots \cdot f_i^{(n)}(x_n)}\right] = \\
&= E\left[\log c_i(F^{(1)}(x_1), \ldots, F^{(n)}(x_n))\right] = H(c_i).
\end{aligned}
$$

dove c_i è la funzione di densità copula corrispondente all'oggetto i-

[1] si veda anche Ma and Sun [2008]

esimo ω_i e $H(c_i)$ è la sua entropia.
Nel complesso quindi si può riscrivere l'entropia congiunta come segue:

$$H(f_i) = \sum_{j=1}^{p} H\left(f_i^{(j)}\right) + H(c_i)$$

Inoltre l'entropia della mistura di n funzioni di densità è data da:

$$H(m) = H\left[\sum_{i=1}^{n}\left(\pi_i c_i \cdot \prod_{j=1}^{p} f_i^{(j)}\right)\right].$$

Quindi, in definitiva, è possibile riscrivere la divergenza di Jensen-Shannon secondo la seguente espressione:

$$\begin{aligned}
&d_{JS}(f_1, \ldots, f_n) = \\
&= H\left[\sum_{i=1}^{n}\left(\pi_i c_i \cdot \prod_{j=1}^{p} f_i^{(j)}\right)\right] - \sum_{i=1}^{n} \pi_i \cdot \left(\sum_{j=1}^{p} H\left(f_i^{(j)}\right) + H(c_i)\right) = \\
&= H\left[\sum_{i=1}^{n}\left(\pi_i c_i \cdot \prod_{j=1}^{p} f_i^{(j)}\right)\right] - \sum_{j=1}^{p}\sum_{i=1}^{n} \pi_i \cdot H\left(f_i^{(j)}\right) - \sum_{i=1}^{n} \pi_i \cdot H(c_i) = \\
&= H(m) - \sum_{j=1}^{p} \bar{H}(f^{(j)}) - \bar{H}(c)
\end{aligned} \tag{4.6}$$

dove $\bar{H}(f^{(j)})$ è l'entropia marginale media relativa alla $j-esima$ variabile e $\bar{H}(c)$ è la media delle entropie delle funzioni copula.

Una volta riscritta la divergenza di Jensen-Shannon in termini di copula non resta che valutare tale divergenza nel contesto della DCA.

Si è detto che, data la 4.1, la funzione $\Delta(P, L)$ è pari dalla somma pesata delle divergenze di Kullback-Leibler tra ogni funzione di densità

f_i corrispondente all'oggetto ω_i e la mistura delle densità relative a tutti gli oggetti appartenenti al cluster C_h:

$$\Delta(P, L) = \sum_{h=1}^{k} \sum_{\omega_i \in C_h} \pi_i \cdot d_{KL}(f_i | m_h)$$

con

$$m_h = \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} \cdot f_i = \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} \cdot c_i \cdot \prod_{j=1}^{p} f_i^{(j)}$$

e

$$\pi^{(h)} = \sum_{\omega_i \in C_h} \pi_i.$$

Ma, come è stato sottolineato in precedenza, data la scomposizione ottenuta nel Teorema 3, essa è esattamente pari alla divergenza entro i clusters. Tale divergenza può essere riscritta, considerando le funzioni copula relative a ciascun oggetto ω_i, come segue:

$$
\begin{aligned}
d_{JS}^{W} &= \sum_{h=1}^{k} \pi^{(h)} d_{JS}^{(h)} = \\
&= \sum_{h=1}^{k} \pi^{(h)} \cdot \left[H(m_h) - \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} H(f_i) \right] = \\
&= \sum_{h=1}^{k} \pi^{(h)} \cdot \left\{ H(m_h) - \sum_{\omega_i \in C_h} \frac{\pi_i}{\pi^{(h)}} \left[\sum_{j=1}^{p} H(f_i^{(j)}) + H(c_i) \right] \right\} = \\
&= \sum_{h=1}^{k} \pi^{(h)} \cdot H(m_h) - \sum_{i=1}^{n} \sum_{j=1}^{p} \pi_i \cdot H(f_i^{(j)}) - \sum_{i=1}^{n} \pi_i \cdot H(c_i) = \\
&= \sum_{h=1}^{k} \pi^{(h)} \cdot H(m_h) - \sum_{j=1}^{p} \bar{H}(f^{(j)}) - \bar{H}(c)
\end{aligned}
\tag{4.7}
$$

Analogamente, la divergenza tra i clusters può essere riscritta come:

$$
\begin{aligned}
d_{JS}^{B} &= \sum_{h=1}^{k} \pi^{(h)} d_{KL}(m_h | m) = \\
&= H(m) - \sum_{h=1}^{k} \pi^{(h)} H(m_h)
\end{aligned}
\tag{4.8}
$$

Anche l'indice $Q(P)$ definito dalla 4.5 può essere riscritto in tal senso, come segue:

$$
\begin{aligned}
Q(P) &= 1 - \frac{d_{JS}^{W}}{d_{JS}^{TOT}} = \\
&= 1 - \frac{\bar{H}(m_h)}{d_{JS}^{TOT}} + \frac{\bar{H}(c)}{d_{JS}^{TOT}} + \frac{\sum_{j=1}^{p} \bar{H}(f^{(j)})}{d_{JS}^{TOT}}
\end{aligned}
\tag{4.9}
$$

dove $\bar{H}(m_h) = \sum_{h=1}^{k} \pi^{(h)} \cdot H(m_h)$.

Quindi, la bontà della partizione dipenderà solo in parte dalla classificazione ottenuta, attraverso la media sui k clusters delle entropie di ciascuna mistura. Dipenderà inoltre direttamente dalle entropie delle funzioni di densità marginali e dalle entropie delle funzioni copula attraverso la loro media calcolata su tutti gli n elementi di E. Perciò sarà direttanmente legata alla dipendenza esistente tra le variabili marginali utilizzate per descrivere l'oggetto ω_i.

Questo appare un risultato lecito se si pensa al fatto che l'omogeneità entro i clusters è sì dovuta alla partizione scelta, ma è ovviamente anche dovuta alle caratteristiche intrinseche delle funzioni di densità congiunte utilizzate per la descrizione degli elementi di E.

4.5.5 La stima

Tutte le quantità ottenute nella precedente sezione, in quanto esprimibili esclusivamente attraverso entropie, possono essere stimate se si sceglie uno stimatore per l'entropia differenziale.

In letteratura numerose sono le proposte in tal senso. Una revisione dei metodi non parametrici più diffusi è proposta in Beirlant et al. [1997]. Nel lavoro citato vengono presentati e discussi numerosi stimatori sostanzialmente basati sulla stima preliminare della funzione di densità con tecniche di tipo non parametrico, quali ad esempio

l'utilizzo di stimatori kernel. Per tutti gli stimatori considerati sono presentate e comparate le diverse proprietà.

Per quanto riguarda invece i metodi parametrici, gli stimatori proposti si basano generalmente sulla funzione di verosimiglianza. In particolare Moddemeijer [2000] descrive i cosidetti stimatori MALL (Maximum Average Log-Likelihood) dei quali ne discute le proprietà.

Una proposta alternativa può essere ritrovata invece in Miller [2003].

Va detto che, nel momento in cui l'obiettivo finale è quello di ottenere una stima della divergenza di di Jensen-Shannon, la scelta dello stimatore dell'entropia richiede particolare cautela. Infatti, solo se vengono soddisfatte alcune condizioni, gli stimatori della divergenza risulteranno accettabili.

A questo proposito si ricorda che, la positività della divergenza di Jensen-Shannon, è garantita dalla disuguaglianza di Jensen e quindi richiede come condizione la concavità delle funzione $H(\cdot)$. Infatti solo se $H(f)$ è una funzione concava in f si ha:

$$H\left[\sum \pi_i f_i\right] \geqslant \sum \pi_i H(f_i)$$

In effetti, l'entropia differenziale è una funzione concava e ciò garantisce $d_{JS}(f_1, \ldots, f_n) \geqslant 0$.

Al fine di ottenere stime positive della divergenza di Jensen-Shannon e quindi accettabili, è necessario che anche lo stimatore $\hat{H}$ dell'entropia abbia questa caratteristica, ovvero sia funzione concava della funzione $\hat{f}$ di densità stimata. Questa condizione non è verificata per tutti gli stimatori proposti in letteratura. Ad esempio, se si considera lo stimatore MALL, si ha:

$$\hat{H}(f) = -\frac{1}{s}\sum_{s=1}^{S} \log f(x_s; \hat{\boldsymbol{\theta}})$$

con S pari all'ampiezza campionaria e $\hat{\boldsymbol{\theta}}$ pari alla stima di massima verosimiglianza del vettore di parametri incognito. Ma $\hat{H}(f)$ è una funzione convessa di f, come si verifica immediatamente. Infatti:

$$\frac{\partial \hat{H}(f)}{\partial f} = -\frac{1}{s}\sum_{s=1}^{S}\frac{1}{f(x_s;\hat{\boldsymbol{\theta}})};$$
$$\frac{\partial^2 \hat{H}(f)}{\partial f^2} = \frac{1}{s}\sum_{s=1}^{S}\frac{1}{[f(x_s;\hat{\boldsymbol{\theta}})]^2} \geqslant 0 \text{ sempre.}$$

Ciò significa che questo stimatore, seppure corredato di diverse proprietà desiderabili e certamente adatto ai fini della stima puntuale dell'entropia differenziale, non può essere utilizzato per costruire uno stimatore della divergenza di Jensen-Shannon.

Diverso è il caso dello stimatore di tipo plug-in descritto in Beirlant et al. [1997], che invece soddisfa la condizione richiesta. Altra valida possibilità è rappresentata dallo stimatore naturale dell'entropia differenziale, il quale, assumendo per definizione la stessa forma funzionale dell'entropia stessa, risulta certamente ammissibile. A differenza dello stimatore plug-in suddetto, si può pensare di procedere attraverso un approccio parametrico per stimare la funzione di densità congiunta. Più in particolare, considerando la modellizzazione attraverso la funzione copula, si può procedere con il metodo IFM, stimando i parametri delle distribuzioni marginali in un primo step e quelli della funzione copula in un secondo. Inoltre, al fine di ottenere la stima dell'entropia differenziale, si può procedere con i metodi di integrazione numerica (si veda ad esempio Kuonen [2003]).

Capitolo 5

Applicazione a dati simulati e reali

5.1 Un caso reale

Al fine di mostrare l'applicazione dell'algoritmo proposto ad una caso reale, sono stati utilizzati i dati climatici giornalieri relativi a venticinque città europee. I dati utilizzati sono stati reperiti on-line sul sito web (http://eca.knmi.nl/) del progetto *European Climate Assessment & Dataset* (ECA&D) [Klein Tank and et Al., 2002].

Tra le serie di dati disponibili per ciascun osservatorio, sono state considerate quelle relative alle temperature massime giornaliere, registrate nell'arco delle 24 ore, e quelle relative ai livelli di precipitazioni medie giornaliere. Come periodo di riferimento è stato considerato quello dal 1951 al 2008 [1].

[1] Per alcune delle città considerate i dati disponibili erano a partire dal 1961 (Stoccolma) oppure fino al 1999 (Barcellona e Lisbona) o 1995 (Marsiglia e Praga)

5.1.1 I dati iniziali

Come proposto da Schölzel and Friederichs [2008], per ridurre la dipendenza temporale e quindi il problema dell'autocorrelazione, sono state calcolate le temperature massime (in gradi centigradi) e le precipitazioni medie settimanali (in cm), su 5 giorni, introducendo un gap di due giorni, tra una valutazione e la successiva. Si è dunque considerato il vettore bivariato, la cui prima componente riguarda le temperature massime settimanali e la cui seconda componente è relativa alle precipitazioni medie settimanali.

5.1.2 Analisi preliminare

Al fine di avere un'idea circa il comportamento delle singole variabili, nonchè relativamente alla struttura di dipendenza, è stata effettuata un'analisi descrittiva preliminare dei dati.

Innanzitutto, è stata effettuata una valutazione grafica delle distribuzioni marginali, considerando gli istogrammi di frequenza delle temperature massime e delle precipitazioni medie di ogni città. Il comportamento delle variabili risulta abbastanza simile per tutte le città considerate. In particolare, la distribuzione delle precipitazioni massime si presenta sempre tendenzialmente unimodale, e con leggera asimmetria a sinistra, mentre la distribuzione delle precipitazioni medie risulta essere sempre decrescente. A titolo di esempio in figura 5.1 sono riportati gli istogrammi di frequenza per la città di Roma.

Al fine invece di avere informazioni circa la dipendenza esistente tra le variabili considerate per ogni città sono stati calcolati il coefficiente di correlazione r di Pearson, il coefficiente τ di Kendall e il coefficiente ρ di Spearman. I risultati ottenuti sono riportati nella tabella 5.1.

Dai risultati ottenuti per i tre indici in generale sembra esservi un basso grado di dipendenza tra le due variabili, alcune volte di tipo positivo altre volte negativo. In alcuni casi, effettuando un test con

Città	r	τ	ρ
Roma	-0.226	-0.317	-0.225
Parigi	0.008	-0.068	-0.044
Madrid	-0.268	-0.326	-0.233
Berlino	0.069	0.014	0.012
Vienna	0.081	0.034	0.024
Stoccolma	0.074	-0.031	-0.018
Barcellona	-0.026	-0.112	-0.081
Lussemburgo	0.001	-0.033	-0.02
Lisbona	-0.38	-0.54	-0.394
Lugano	0.071	0.123	0.082
Helsinki	0.034	-0.056	-0.034
Reykjavik	-0.014	-0.02	-0.016
Praga	0.258	0.222	0.155
Milano	-0.043	-0.046	-0.033
Marsiglia	-0.126	-0.249	-0.177
Monaco	0.197	0.17	0.117
Bucharest	0.022	-0.042	-0.029
Vilnius	0.142	0.05	0.04
Tirana	-0.265	-0.337	-0.237
Cagliari	-0.27	-0.455	-0.325
Corfù	-0.32	-0.498	-0.361
Zagabria	0.063	0.035	0.0233
Dublino	-0.066	-0.099	-0.065
Bordeaux	-0.173	-0.194	-0.135
Oslo	0.083	0.032	0.025

Tabella 5.1: Alcuni indici di dipendenza

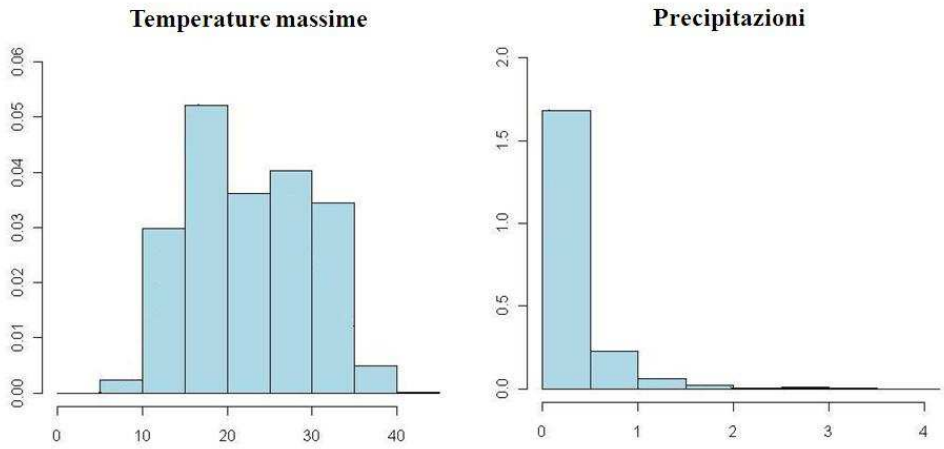

Figura 5.1: Istogrammi di frequenza delle temperature massime e delle precipitazioni medie giornaliere della città di Roma

ipotesi nulla tale che il singolo indice sia pari a zero, il p-value che si ottiene porta ad accettare l'ipotesi suddetta (considerando l'usuale soglia del 5%).

Inoltre, anche in questo contesto, un'analisi grafica del fenomeno è stato effettuata considerando lo scatterplot relativo ad ogni città per le due variabili. Ancora a titolo esemplificativo, è riportato in figura 5.2 il grafico ottenuto per la città di Roma, ma si evidenzia che il comportamento è molto simile per tutte le altre città. Quello che si può notare è che sembra esservi assenza di dipendenza di coda,

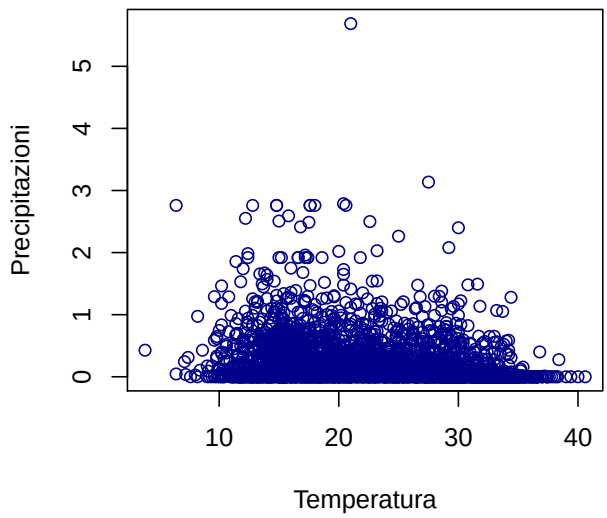

Figura 5.2: Scatterplot delle temperature e delle precipitazioni della città di Roma

sia destra che sinistra, ovvero a livelli alti di precipitazioni non si accompagnano valori alti di temperatura, così come a livelli bassi di precipitazioni non si associano in genere valori bassi di temperatura. Queste tendenze riscontrate nei dati risultano essenziali per la scelta della funzione copula.

5.1.3 I descrittori

In accordo con la teoria dei valori estremi [Coles, 2001], per descrivere la temperatura massima è stata utilizzata la distribuzione di Gumbel, la cui funzione di ripartizione è data da:

$$F(x;\alpha,\beta) = \exp\left\{-\exp\left[-\left(\frac{x-\beta}{\alpha}\right)\right]\right\}$$

Per descrivere invece la distribuzione delle precipitazioni medie è stata adottata la distribuzione Gamma:

$$F(y,\delta,\lambda) = \frac{\gamma(\delta, x/\lambda)}{\Gamma(\delta)}$$

dove $\gamma(\delta, x/\lambda)$ è la funzione gamma incompleta e δ è un parametro di forma, mentre λ è un parametro di scala.

Al fine di modellare la struttura di dipendenza tra le due variabili considerate, alcune osservazioni preliminari sono state effettuate. Nel caso specifico va sottolineato innanzitutto che la dipendenze tra temperatura massima e livello di precipitazioni medio risulta essere generalmente di basso livello, e può essere sia di tipo positivo che negativo. Dunque la scelta di una funzione copula che leghi le due distribuzioni marginali suddette deve necessariamente riflettere tali caratteristiche. Inoltre anche la dipendenza di coda risulta essere abbastanza contenuta o assente. Sotto queste condizioni quindi, sembra essere appropriato

l'impiego della copula di Frank. La funzione di densità copula per tale modello, considerando due distribuzioni marginali è la seguente:

$$c(u,v;\theta) = \frac{\theta \cdot [1-\exp(-\theta)] \cdot \exp[-\theta \cdot (u+v)]}{\{1-\exp(-\theta)-[1-\exp(-\theta \cdot u)] \cdot [1-\exp(-\theta \cdot v)]\}^2}$$

Quindi la densità bivariata delle temperature massime e delle precipitazioni avrà la seguente espressione:

$$\begin{aligned} &h(x,y;\alpha,\beta,\gamma,\lambda,\theta) = \\ &\left\{\theta \cdot \left[1-e^{-\theta}\right] \cdot e^{-\theta \cdot \left[\exp\left\{-\exp\left[-\left(\frac{x-\beta}{\alpha}\right)\right]\right\}+\frac{\gamma(\delta,x/\lambda)}{\Gamma(\delta)}\right]}\right\} \cdot \\ &\left\{1-e^{-\theta}-\left[1-e^{-\theta \cdot \exp\left\{-\exp\left[-\left(\frac{x-\beta}{\alpha}\right)\right]\right\}}\right] \cdot \left[1-e^{-\theta \cdot \frac{\gamma(\delta,x/\lambda)}{\Gamma(\delta)}}\right]\right\}^{-2} \end{aligned}$$

La stima dei parametri

Per la stima dei parametri delle funzioni di ogni città, è stato utilizzato il cosiddetto metodo IFM, ottenendo quindi in un primo tempo, tramite il metodo della massima verosimiglianza, i parametri delle distribuzioni marginali, e in un secondo tempo il parametro della funzione copula. I valori ottenuti per ogni singola città sono riportati in tabella 5.2.

La figura 5.3 riporta le funzioni di densità stimate corrispondenti alle distribuzioni di figura 5.1. Inoltre è riportato il grafico con i contorni di densità della distribuzione bivariata stimata attraverso l'utilizzo della copula di Frank.

5.1.4 La classificazione

La procedura di classificazione delle città prescelte è stata ripetuta sotto diverse ipotesi. Innazitutto è stata fatta variare l'ampiezza della

	Gumbel		Gamma		Frank Copula
Città	α	β	γ	λ	θ
Roma	19.224	6.313	0.161	1.401	-0.807
Parigi	14.745	7.113	0.242	0.720	-0.187
Madrid	17.770	7.008	0.126	0.997	-1.026
Berlino	11.850	8.504	0.275	0.589	0.001
Vienna	12.878	9.086	0.249	0.685	0.076
Stoccolma	8.531	8.077	0.262	0.558	-0.245
Barcellona	18.389	6.866	0.112	1.564	-0.300
Lussemburgo	11.645	7.807	0.262	0.898	-0.107
Lisbona	20.472	5.002	0.128	1.608	-1.856
Lugano	15.539	7.094	0.158	2.796	0.467
Helsinki	6.379	8.360	0.301	0.587	-0.311
Reykjavik	7.137	4.482	0.351	0.609	0.274
Praga	12.260	8.770	0.249	0.505	0.314
Milano	15.927	8.225	0.136	1.999	0.005
Marsiglia	18.585	5.814	0.125	1.281	-0.794
Monaco	11.951	8.737	0.280	0.913	0.273
Bucharest	14.740	10.320	0.172	0.954	-0.100
Vilnius	8.126	9.582	0.278	0.623	-0.211
Tirana	19.699	6.535	0.178	1.805	-0.590
Cagliari	20.500	5.600	0.150	0.762	-1.570
Corfù	20.500	5.734	0.126	2.328	-1.701
Zagabria	14.201	9.006	0.213	1.125	0.103
Dublino	13.232	4.628	0.375	0.562	-0.135
Bordeaux	17.567	6.609	0.225	1.142	-0.211
Oslo	8.198	8.307	0.246	0.866	-0.063

Tabella 5.2: Stime dei parametri delle distribuzioni marginali e della funzione copula

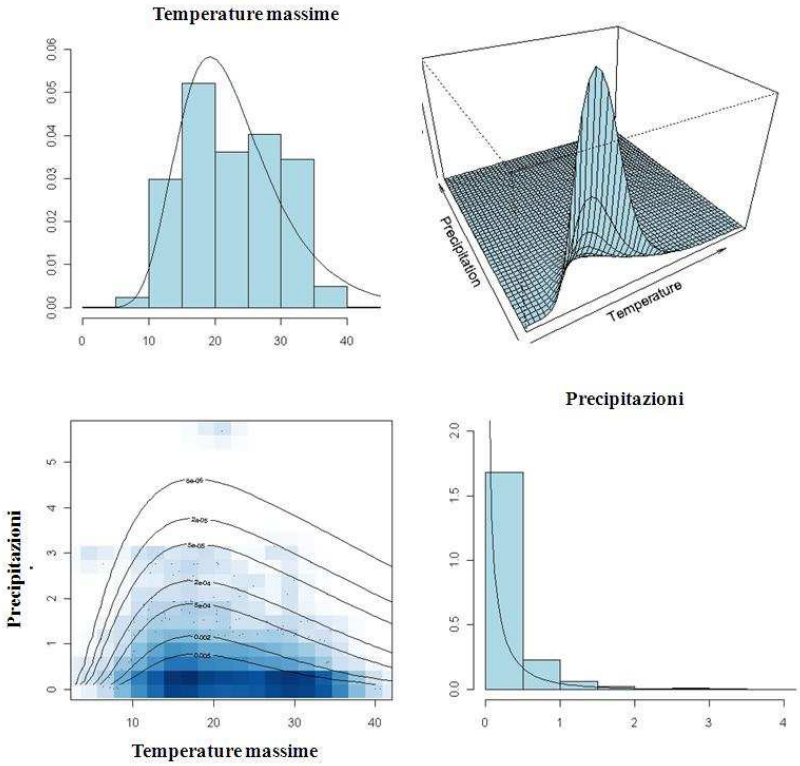

Figura 5.3: Funzioni di densità congiunta e densità marginali stimate per la città di Roma

partizione, fissando il numero di clusters inizialmente pari a due e successivamente pari a tre. I risultati ottenuti sono stati poi messi a confronto con gli analoghi ottenuti sotto l'ipotesi di indipendenza tra la temperatura e le precipitazioni, ovvero considerando quali descrittori degli oggetti le distribuzioni multivariate ottenute attraverso l'utilizzo della copula *prodotto*.

Il caso di due gruppi

Per la classificazione delle 25 città è stata inizialmente considerata una bi-partizione degli oggetti, ovvero è stato posto un numero di cluster k pari a 2. L'applicazione dell'algoritmo di clustering dinamico utilizzando come misura di dissimilarià tra le distribuzioni multivariate quella di Jensen-Shannon ha portato alla seguente classificazione:

Cluster 1: Roma, Madrid, Barcellona, Lisbona, Lugano, Milano, Marsiglia, Tirana, Cagliari, Corfù, Bordeaux;

Cluster 2: Parigi, Berlino, Vienna, Stoccolma, Lussemburgo, Helsinki, Reykjavik, Praga, Monaco, Bucharest, Vilnius, Zagabria, Dublino, Oslo.

In tabella 5.3 sono riportate le distanze ottenute per ogni singola città dal baricentro del cluster di appartenenza.

Come si può notare anche dalla figura 5.4, l'algoritmo ha consentito di identificate una fascia climatica 'meridionale' ed una 'settentrionale'.

E' da notare come la partizione ottenuta segua la contiguità geografica delle città considerate. I due gruppi inoltre risultano essere non molto distanti tra loro, come si evince dal rapporto tra distanza entro i gruppi e distanza totale. Infatti la distanza entro i gruppi ($d_{JS}^{W} = 0.1171$) rappresenta il 42% della distanza complessiva

Cluster 1		Cluster 2	
Città	Distanza	Città	Distanza
Roma	0.0306	Parigi	0.1636
Madrid	0.0325	Berlino	0.0339
Barcellona	0.0450	Vienna	0.0524
Lisbona	0.1360	Stoccolma	0.0517
Lugano	0.1542	Lussemburgo	0.0595
Marsiglia	0.0313	Helsinki	0.1754
Tirana	0.0533	Reykjavik	0.3754
Cagliari	0.1223	Praga	0.0226
Corfù	0.1068	Monaco	0.0507
Bordeaux	0.1587	Bucharest	0.2093
Milano	0.1377	Vilnius	0.0978
		Zagabria	0.1559
		Dublino	0.3667
		Oslo	0.1044

Tabella 5.3: Distanze di ogni città dal baricentro del cluster di appartenenza

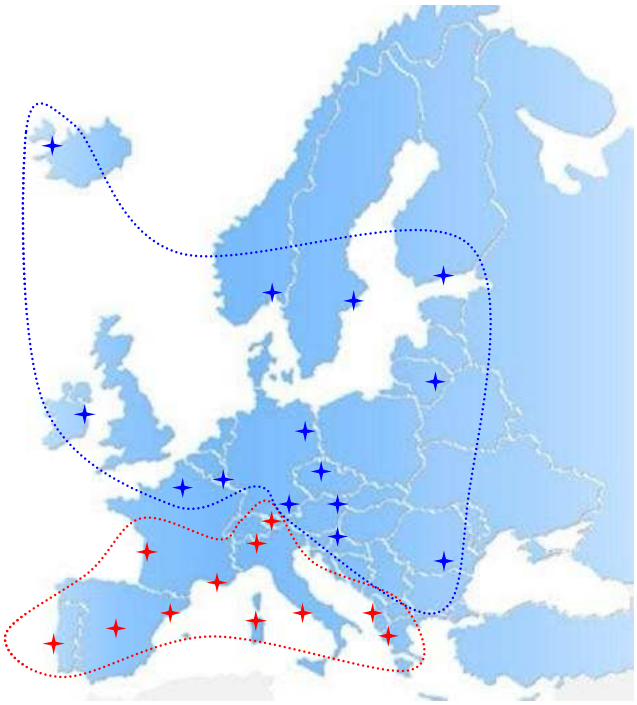

Figura 5.4: Risultati della classificazione in due gruppi di 25 città europee

Cluster 1		Cluster 2		Cluster 3	
Città	Distanza	Città	Distanza	Città	Distanza
Roma	0.0232	Lugano	0.1169	Berlino	0.0420
Madrid	0.0665	Milano	0.1017	Vienna	0.0850
Barcellona	0.0602	Bucharest	0.0497	Stoccolma	0.0289
Lisbona	0.0765	Zagabria	0.0662	Lussemburgo	0.0803
Marsiglia	0.0178	Bordeaux	0.1416	Helsinki	0.1224
Tirana	0.0471	Parigi	0.1331	Reykjavick	0.3010
Cagliari	0.0796			Praga	0.0392
Corfù	0.0555			Monaco	0.0694
				Vilnius	0.0732
				Dublino	0.3529
				Oslo	0.0937

Tabella 5.4: Distanze di ogni città dal baricentro del cluster di appartenenza

($d_{JS}^{Tot} = 0.276$), mentre il restante 58% è attribuibile alla distanza tra i gruppi.

Il caso di tre gruppi

L'algoritmo è stato impiegato nuovamente ponendo il numero di clusters uguale a 3. Il raggruppamento che si ottiene per le città considerate è mostrato in figura 5.5, mentre le distanze di ogni città dal nuovo prototipo del gruppo di appartenenza sono riportate in tabella 5.4.

Anche in questo caso i gruppi sono formati da città geograficamente contigue. Si viene a creare una terza fascia climatica intermedia tra le località più fredde e quelle più calde, costituita da tre città (Lugano, Milano e Bordeaux) precedentemente allocate nel cluster 1 e da altre tre (Bucharest, Zagabria e Parigi) che invece si trovavano nel cluster 2. Come ci si attende, la divergenza entro i gruppi si riduce ($d_{JS}^{W} =$

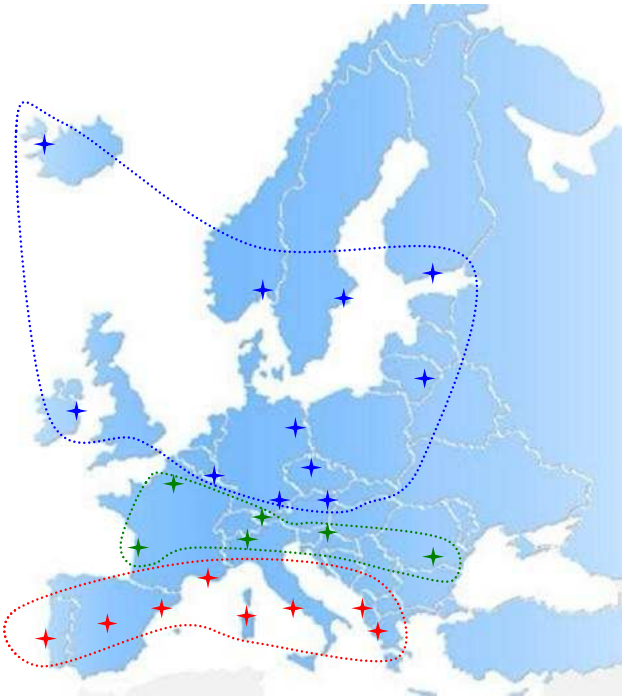

Figura 5.5: Risultati della classificazione in tre gruppi di 25 città europee

0.0929) e quindi l'indice di bontà della partizione aumenta, essendo pari a $Q(P) = 0.663$. Perciò, soltanto il 33.7% della divergenza totale è attribuibile alla divergenza entro i clusters, mentre il restante 66.3% è dovuto alla dissimilarità tra i clusters.

Anche in questo caso la classificazione ottenuta sembra essere plausibile rispetto alla localizzazione geografica delle città costituenti ogni singolo cluster.

Il caso dell'indipendenza

L'algoritmo è stato infine impiegato supponendo che temperatura e livelli di precipitazioni siano tra loro indipendenti. Quali descrittori multivariati di ciascun oggetto quindi sono stati considerati i prodotti delle distribuzioni marginali. La classificazione ottenuta è risultata differente rispetto alla precedente ed è rappresentata in figura 5.6

Si sottolinea che il Cluster 1, contenente in precedenza solo città il cui parametro di dipendenza era negativo, adesso include anche le città di Milano e Lugano, che invece sembrano essere caratterizzate da dipendenza positiva. Gli altri due clusters includono sia città la cui corrispondente funzione copula era caratterizzata da parametro con segno negativo e sia città con funzione copula avente parametro positivo.

Una seconda osservazione riguarda la bontà della partizione ottenuta. L'indice di qualità, anche se di poco, registra un leggero miglioramento, passando da $Q(P) = 0.663$ ad un valore pari a $Q(P) = 0.691$. Si fa presente che anche la divergenza di Jensen-Shannon totale è leggermente inferiore e pari a $d_{JS}^{TOT} = 0.275$ (in precedenza si aveva $d_{JS}^{TOT} = 0.276$), presumibilmente perchè le funzioni in questo caso si differenziano solo per i parametri delle distribuzioni marginali e non più per il parametro della funzione copula e dunque risultano nel complesso più simili.

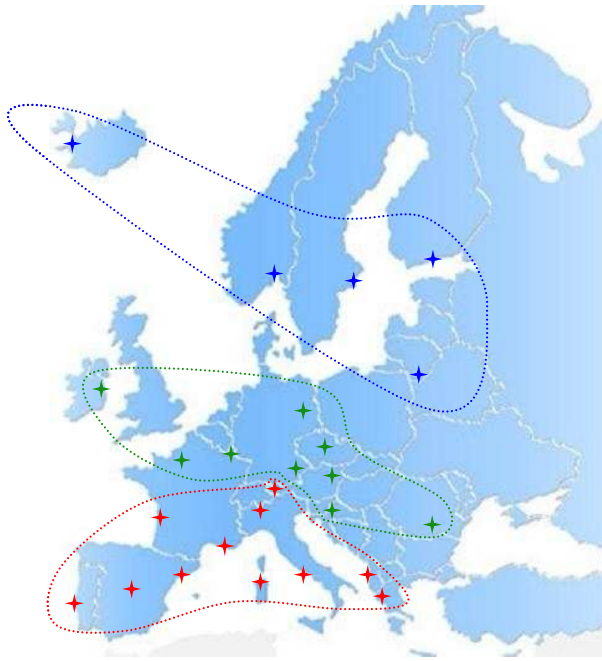

Figura 5.6: Risultati della classificazione ottenuti ipotizzando l'indipendenza tra temperature e precipitazioni

In definitiva, data anche la situazione particolare di bassa dipendenza presente nei dati, non si evince un'influenza netta del parametro della funzione copula nel processo di classificazione e quindi in questo caso particolare, l'informazione aggiuntiva circa la dipendenza determina una diversa classificazione, ma non può concludersi se tale classificazione sia migliore o peggiore.

Ovviamente, la procedura proposta risulta sensibile alle diverse specificazioni dei modelli per la descrizione degli oggetti da classificare. Tanto più è alto l'adattamento delle funzioni multivariate prescelte ai dati, tanto maggiore sarà la veridicità dei risultati ottenuti dall'algoritmo DCA.

5.2 Simulazione

Al fine di avere ulteriori elementi per la valutazione della performance dell'algoritmo proposto, si è predisposta una simulazione. L'idea è stata quella di scegliere distribuzioni marginali e funzioni copule per 21 oggetti 'fittizi' e formulare da queste le corrispondenti distribuzioni multivariate. Come dati di input per l'algoritmo DCA proposto è stato considerato il database in tabella 5.5.

Dato che la procedura per la stima della distanza tra un oggetto e l'altro è basata su metodi di integrazione numerica, al fine di valutare la precisione di tali metodi nel fornire distanze positive si sono scelte funzioni molto simili tra loro. Infatti le distribuzioni marginali di ogni oggetto sono state supposte identiche e si sono fatti variare di poco i parametri associati. Di volta in volta si è fatto variare anche il parametro della funzione copula. In questo modo le distanze ottenute sono prossime allo zero, ma questo ha permesso di verificare che anche in questo caso estremo non si ottengono valori negativi.

Il numero di clusters è stato fissato uguale a due e i risultati ottenuti riflettono una classificazione che sembra essere influenzata esclusiva-

Oggetto	$Y^{(1)}$	$Y^{(2)}$	$Y^{(3)}$	Copula
ω_1	$Exp(3)$	$Exp(3)$	$Exp(3)$	$Clayton(0.1)$
ω_2	$Exp(3)$	$Exp(3)$	$Exp(3)$	$Clayton(0.5)$
ω_3	$Exp(3)$	$Exp(3)$	$Exp(3)$	$Clayton(1.5)$
ω_4	$Exp(3)$	$Exp(3)$	$Exp(3)$	$Clayton(3)$
ω_5	$Exp(3)$	$Exp(3)$	$Exp(3)$	$Clayton(5)$
ω_6	$Exp(3)$	$Exp(3)$	$Exp(3)$	$Clayton(10)$
ω_7	$Exp(3)$	$Exp(3)$	$Exp(3)$	$Clayton(20)$
ω_8	$Exp(5)$	$Exp(5)$	$Exp(5)$	$Clayton(0.1)$
ω_9	$Exp(5)$	$Exp(5)$	$Exp(5)$	$Clayton(0.5)$
ω_{10}	$Exp(5)$	$Exp(5)$	$Exp(5)$	$Clayton(1.5)$
ω_{11}	$Exp(5)$	$Exp(5)$	$Exp(5)$	$Clayton(3)$
ω_{12}	$Exp(5)$	$Exp(5)$	$Exp(5)$	$Clayton(5)$
ω_{13}	$Exp(5)$	$Exp(5)$	$Exp(5)$	$Clayton(10)$
ω_{14}	$Exp(5)$	$Exp(5)$	$Exp(5)$	$Clayton(20)$
ω_{15}	$Exp(20)$	$Exp(20)$	$Exp(20)$	$Clayton(0.1)$
ω_{16}	$Exp(20)$	$Exp(20)$	$Exp(20)$	$Clayton(0.5)$
ω_{17}	$Exp(20)$	$Exp(20)$	$Exp(20)$	$Clayton(1.5)$
ω_{18}	$Exp(20)$	$Exp(20)$	$Exp(20)$	$Clayton(3)$
ω_{19}	$Exp(20)$	$Exp(20)$	$Exp(20)$	$Clayton(5)$
ω_{20}	$Exp(20)$	$Exp(20)$	$Exp(20)$	$Clayton(10)$
ω_{21}	$Exp(20)$	$Exp(20)$	$Exp(20)$	$Clayton(20)$

Tabella 5.5: Dati Simulati

mente dal parametro della funzione copula prescelta. Infatti il primo cluster risulta composto dagli oggetti descritti dalle funzioni con parametro di dipendenza inferiore a 5, mentre il secondo è caratterizzato dalle distribuzioni aventi parametro della copula maggiore o uguale a 5.

Inoltre sono stati calcolati i valori della divergenza di Jensen-Shannon totale, e della divergenza di Jensen-Shannon entro e tra i gruppi, al fine di calcolare la misura di bontà della partizione proposta nel precedente capitolo. Il valore di quest'ultima è il seguente:

$$Q(P) = 1 - \frac{d_{JS}^{W}}{d_{JS}^{TOT}} = 1 - \frac{0.581}{7.722} = 0.925$$

Il valore ottenuto riflette una partizione quasi ottimale, in cui i due prototipi rappresentano quasi perfettamente gli oggetti appartenenti ai rispettivi cluster. In questo caso la divergenza totale è quasi tutta dovuta alla dissimilarità tra i due gruppi.

In definitiva, oltre alla valutazione circa la positività delle distanze stimate, la particolare struttura di questi dati ha permesso di evidenziare il fatto che esistono situazioni in cui la classificazione è determinata esclusivamente dal legame esistente tra i descrittori di ciascun oggetto. Tale evidenza pone in risalto la necessità di ulteriori approfondimenti riguardo il ruolo specifico della copula nel processo di classificazione, questione che può rappresentare lo spunto di riflessione per ulteriori sviluppi futuri.

Conclusioni

Con il presente lavoro si è tentato di fornire una soluzione per la descrizione, il confronto e la classificazione di dati non puntuali, che siano descritti da distribuzioni.

L'informazione aggiuntiva relativa alla dipendenza, spesso disponibile per questo tipo di dati, è stata considerata attraverso l'approccio delle funzioni copula. Tramite questo strumento, infatti, è possibile modellare la struttura di dipendenza presente nei dati e, a partire dalle distribuzioni marginali, ricavare le distribuzioni multivariate associate a ciascun oggetto.

La proposta di valutare il grado di dissimilarità tra gli oggetti utilizzando la divergenza di Jensen-Shannon, consente di considerare anche questa informazione aggiuntiva e quindi di predisporre comparazioni tra oggetti che tengano conto anche della interconnessione tra i descrittori utilizzati. Infatti tale misura è funzione del rapporto delle densità multivariate associate agli oggetti in esame e, proprio per questo, dipende sia dalle funzioni marginali, che dalla funzione copula.

Nel presente lavoro di tesi, è stato verificato che tale misura può essere proficuamente utilizzata nel contesto del clustering dinamico esteso al trattamento dei dati descritti da distribuzioni multivariate. Infatti gode di alcune proprietà che la rendono adatta ad essere impiegata nella costruzione di un algoritmo di tipo dinamico. Il suo utilizzo non solo permette di identificare un elemento rappresentativo di cia-

scun cluster, ma si è dimostrato che questo è unico e ciò garantisce la convergenza dell'algoritmo stesso, in coerenza con lo schema classico del DCA.

Inoltre è stato verificato che, in accordo con i metodi standard, la divergenza totale tra tutti gli oggetti in esame può essere decomposta in divergenza tra ed entro i gruppi. Quest'ultima proprietà ha consentito di costruire un indicatore per la valutazione della bontà della partizione ottenuta, basato sul confronto tra la divergenza *beetwen* e la divergenza totale.

Nel complesso, dunque, l'algoritmo proposto consente di trovare simultaneamente la migliore partizione dell'insieme di oggetti considerati, secondo il criterio prefissato, e un modello adatto a descrivere la strutture di dipendenza tra le variabili considerate.

Lo studio di un caso reale ha messo in evidenza l'applicabilità di tale algoritmo e il vantaggio derivante dal suo utilizzo quando si è in presenza di grandi moli di dati. La descrizione dei dati elementari attraverso le rispettive funzioni di densità e successivamente l'applicazione di un algoritmo capace di trattare direttamente con questa tipologia di informazioni consente di ritenere una quantità di dati estremamente ridotta rispetto a quelli di partenza.

Per concludere, insieme alle precedenti considerazioni, si vogliono in questa sede anche proporre spunti di riflessione per eventuali sviluppi futuri. La prima riguarda la complessità computazionale dell'algoritmo proposto. Il calcolo della dissimilarità tra gli oggetti è stato basato su metodi di integrazione numerica. Ciò porta a costi generalmente elevati, in termini di tempo. Un miglioramento potrebbe essere rappresentato dall'utilizzo di stimatori alternativi per la divergenza di Jensen Shannon. Inoltre, uno studio delle proprietà degli stimatori potrebbe indirizzare la scelta su stimatori non necessariamente naturali.

Il secondo spunto per ulteriori valutazioni riguarda il peso della dipendenza sulla determinazione della divergenza: potrebbe essere di

notevole interesse riuscire a quantificare quale sia l'impatto della copula e quale quello delle distribuzioni marginali sulla distanza tra due oggetti e perciò sull'intero processo di classificazione. Cercare di scindere analiticamente i suddetti fattori potrebbe fornire un contributo rilevante nella valutazione del ruolo svolto dalle diverse distribuzioni nella determinazione della divergenza complessiva.

Infine, resta aperta la possibilità di applicare la metodologia proposta a dati di diversa natura, ad esempio dati descritti da variabili ad istogramma, assimilabili a distribuzioni empiriche.

Appendice A

Il codice in linguaggio R

La classificazione è ottenuta applicando la funzione DCA_JS, avente tre argomenti:

- *inputList*:
  è la lista di oggetti da classificare costituita da elementi del tipo:
  $object <- list(fun = f, from = fFrom, to = fTo, name='oggetto')$
  dove f è l'espressione della funzione di densità, *fFrom* e *fTo* sono gli estremi da utilizzare durante le procedure di integrazione numerica e *name* è il nome assegnato all'oggetto;

- h:
  è il numero prefissato di cluster

- t:
  è il valore da assegnare all'argomento *minpts* della funzione *adapt*, per ottenere la precisione desiderata nel calcolo degli integrali.

La funzione prevede l'utilizzo del pacchetto *adapt* che deve essere preliminarmente installato.

```
DCA_JS <- function(inputList,h,t) {
    require(adapt);
    # Funzione per la scrittura del file di log
    logWrite <- function(string) {
        cat(string, file="stima.log", append=TRUE, fill=
            TRUE);
    }
    n <- length(inputList); # numero di oggetti da
        classificare

    # Definizione dei range per il calcolo delle distanze
    minRange <- function(cluster) {
        dimension <- length(cluster[[1]]$from);
        result <- rep(Inf,dimension);
        for (i in 1:dimension) {
            for (j in 1: length(cluster)) {
                result[i] <- min(cluster[[j]]$from[i],
                    result[i]);
            }
        }
        result;
    }
    maxRange <- function(cluster) {
        dimension <- length(cluster[[1]]$to);
        result <- rep(-Inf,dimension);
        for (i in 1:dimension) {
            for (j in 1: length(cluster)) {
                result[i] <- max(cluster[[j]]$to[i],result[
                    i]);
            }
        }
        result;
    }

    # Si generano casualmente h prototipi
    randomList <- sample(c(1:n),h,replace=FALSE);
    logWrite("Funzioni appartenenti al cluster:");
    for (i in 1: length(randomList)) {
```

```
    logWrite(paste("f",randomList[i],sep=""));
}
# Vettore contenente per ciascuna funzione, l'indice
    del cluster di appartenenza
clusterIndex <- rep(-1,n);

# Vettore contenente le distanza di ciascuna funzione
    dal cluster di appartenenza.
distanceList <- rep(Inf,n);

# Associazione delle funzioni scelte ai cluster
for (i in 1:h) {
    clusterIndex[randomList[i]] <- i;
    distanceList[randomList[i]] <- 0;
}

# Calcolo dell'entropia associata a ciascuna funzione
entropie_f <- NULL;
for (i in 1:n) {
    ent <- function(x) {
        -(inputList[[i]]$fun)(x)*(log(inputList[[i]]
            $fun(x)))
    }
    Hf <- adapt(2, lo=inputList[[i]]$from, up=inputList
        [[i]]$to, functn = ent, minpts = t, eps=1);
    entropie_f[i] <- Hf$value;
}

# Composizione del nuovo cluster
clusterIndexNew <- NULL;
while (!identical(clusterIndex,clusterIndexNew)) {

    if (!is.null(clusterIndexNew)) {
        clusterIndex <- clusterIndexNew;
    }
    cluster <- list();
    for (i in 1:h) {
        cluster[[i]]<-list();
```

```
}

# Si determina il cluster di riferimento per
    ciascuna funzione
for (i in 1:n) {
    if (clusterIndex[i] != -1) {
        cluster[[clusterIndex[i]]][length(cluster[[
            clusterIndex[i]]])+1] <- inputList[i];
    }
}

for (i in 1:n) {
    logWrite(paste("\nRicerca del cluster di
        riferimento per la funzione f",i,sep=""));

    k <- -1; # indice della f di riferimento a cui
        associare la f in esame
    min <- Inf; # Minima distanza

    # Si calcola la distanza della i-esima funzione
        da ciascun cluster
    for (j in 1:h) {

        clusterTemp <- cluster[[j]];

        if (clusterIndex[i] != j) {
            clusterTemp[length(clusterTemp)+1] <-
                inputList[i];
        }

        logWrite(paste("composizione temporanea del
            cluster",j));
        for (m in 1:length(clusterTemp)) {
            logWrite(paste(clusterTemp[[m]]$name,"
                "));
        }

        ftot <- function(x) {
```

```
            result<-0;
            for(i in 1:length(clusterTemp)) {
                result<-result+clusterTemp[[i]]$fun
                    (x);
            }
        }

        fm <- function(x) {
            (1/length(clusterTemp))*ftot(x)
        }
        low <- minRange(clusterTemp);
        upp <- maxRange(clusterTemp);

        intKL <- function(x) {
            inputList[[i]]$fun(x)*log(fm(x))
        }

        argKL <- adapt(2, lo=low, up=upp, functn =
            intKL, minpts = t, eps=1);

        distanza <- -entropie_f[i] - argKL$value;

        logWrite(paste("distanza della funzione f",
            i," dal cluster temporaneo ", j, ": ",
            distanza,sep=""));

        if (distanza < min) {
            min <- distanza;
            k <- j;
        }
    }

    if (clusterIndex[i] != k) {
        # la funzione ha cambiato cluster.
        logWrite(paste("la funzione f",i," viene
            associata al cluster ",k,sep=""));
        clusterIndexNew[i] <- k;
    } else {
```

```
            # la funzione rimane nel cluster precedente
            logWrite(paste("la funzione f",i," rimane
                nel cluster ",clusterIndex[i],sep=""));
            clusterIndexNew[i] <- clusterIndex[i];
        }
        # Si aggiorna la distanza
        distanceList[i] <- min;
    }
}

# stampa dei risultati
print("Risultati ottenuti:");
for (i in 1:h) {
    tmp <- paste("il cluster ",i," contiene i seguenti
        oggetti:");
    logWrite(tmp);
    print(tmp);
    for(j in 1: length(cluster[[i]])) {
        tmp <- paste("       ", cluster[[i]][[j]]$name);
        logWrite(tmp);
        print(tmp);
    }
}

# Calcolo della distanza within
dw<-rep(0,h);
for (j in 1: h){
    for (i in 1:n) {
        if (clusterIndex[i]==j) {
            dw[j]<- dw[j]+1/n*distanceList[i];
        }
    }
}
d_W=sum(dw);
logWrite(paste("distanza Within=",d_W));
print(paste("distanza Within=",d_W));

# Calcolo della distanza totale
```

```
    ftot <- function(x) {
        result<-0;
        for(i in 1:n) {
            result<-result+inputList[[i]]$fun(x);
        }
    }

    fm <- function(x) {
        (1/n)*ftot(x)
    }
    low <- minRange(inputList);
    upp <- maxRange(inputList);

    ent_m <- function(x) {
        -fm(x)*log(fm(x))
    }
    Hf_m <- adapt(2, lo=low, up=upp, functn = ent_m, minpts
        = t, eps=1);

    dTot <- Hf_m$value - ((1/n)*sum(entropie_f));

    logWrite(paste("distanza totale: ", dTot,sep=""));
    ratio <- d_W/dTot;
    logWrite(paste("rapporto d_w/dTot: ", ratio,sep=""));
    print(paste("rapporto d_w/dTot: ", ratio,sep=""));

}
```

Bibliografia

A. Arnault and P. Nicole. *La Logique ou l'art Depensur*. Wiley, 1662.

J. Beirlant, E. J. Dudewicz, L. Gyorfi, and E. C. van der Meulen. Nonparametric entropy estimation: an overview. *International journal of mathematical and statistical scinces*, 6:17–39, 1997.

L. Billard and E. Diday. *Symbolic data analysis*. Wiley, 2007.

H. H. Bock and E. Diday. *Analysis of Symbolic data*. Springer, 2000.

H. Carley. Maximum and minimum extensions of finite subcopulas. *Comm. Statist. Theory Methods*, 31:2151–2166, 2002.

G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy. *Classification automatique des donnes*. Dunod, 1989.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statis.*, 23: 493–507, 1952.

S. Coles. *An introduction to statistical modeling of extreme values*. Springer-Verlag, London, 2001.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

I. Csiszár. Information-type measures of differences of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

F. de Carvalho, Y. Lechevallier, and R. Verde. *Symbolic Data Analysis and the Sodas Software*, chapter Clustering methods in Symbolic Data Analysis. John Wiley & Sons, Ltd, 2008.

E. Diday. La méthode des nuées dynamiques. *Reveu de Statistique Appliquée*, 19:19–34, 1971.

E. Diday. Introduction à l'approche symbolique en analyse des données. In *Première Journées Simbolique-Numerique*, pages 21–56. CEREMADE, Université Paris IX- Dauphine, 1987.

D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49:1858–1860, 2003.

K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, New York, 1972.

C. Genest, K. Ghoudi, and L. P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552, 1995.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002.

A. Irpino and E. Romano. Optimal histogram representation of large data.set: Fisher vs piecewise linear approximations. *RNTI E*, 9: 99–110, 2007.

H. Joe. *Multivariate Models and Dependence Concepts.* Chapman & Hall, 1997.

H. Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419, 2005.

G. Kimeldorf and A. R. Sampson. Uniform representations of bivariate distribution. *Comm. Statist.*, 4:617–627, 1975.

A. M. G. Klein Tank and et Al. Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *Int. J. of Climatol.*, 22:1441–1453, 2002.

D. Kuonen. Numerical integration in s-plus or r: A survey. *Journal of Statistical Software*, 8(13):1–14, 2003. URL `http://www.jstatsoft.org/v08/i13`.

J. Ma and Z. Sun. Mutual information is copula entropy, 2008. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:0808.0845`.

D. L. McLeish and C. G. Small. The theory and applications of statistical inference functions. *Lecture Notes in Statistics*, 44, 1988.

E. G. Miller. A new class of entropy estimators for multi-dimensional densities. In *International Conference on Acoustic, Speech, and signal processing*, 2003.

R. Moddemeijer. The distribution of entropy estimators based on maximum mean log-likelihood. In 21^{st} *Symposium on Information theory in the Benelux*, 2000.

R. B. Nelsen. *An introduction to Copulas*. Springer, 2006.

A. Papoulis. *Probability, Random Variables nd Stochastic Processes*. McGrow Hill, 1991.

A. Renyi. On measures of entropy and information. In *Proc.* 4^{th} *Berkely Symp. Math. Statist. Probab.*, volume 1. J. Neyman (ed.). Berkeley, Calif., 1961.

E. Rosh. *Principles of Categorization*, chapter Cognition and Categorization. (Ed.E. Rosh and B. B. Lloyd) Erlbaum Associates, Hillsdale, NJ, 1978.

A. M. G. Schölzel and P. Friederichs. Multivariate non-normally distributed random variables in climate research - introduction to the copula approach. *Nonlinear Process in Geophysics*, 15:761–772, 2008.

C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois press, 1949.

A. Sklar. Fonction de répartition à n dimensions et leurs marges. *Publ Inst Statist Univ Paris*, 8:229–231, 1959.

R. Verde and A. Irpino. Comparing histogram data using mahalanobis-wasserstein distance. In *Compstat 2008: Proceedings in Computational Statistics*. Heidelberg, Physica-Verlag Springer, 2008.

J. J. Xu. *Statistical Modelling and Inference for Multivariate and Longitudinal Discrete Response Data*. PhD thesis, Department of Statistics, University of British Columbia, 1996.