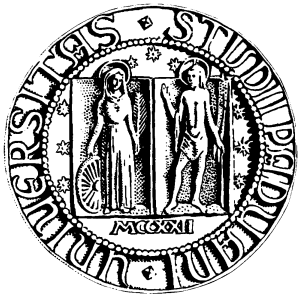
UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA

IN STATISTICA E INFORMATICA



NETWORK BAYESIANI: UN APPROCCIO NON PARAMETRICO BASATO SULL'ENTROPIA PER LA SELEZIONE DEL MODELLO

Relatore: Chiar.mo Prof. Adriana Brogini

Correlatore: Chiar.mo Prof. Fortunato Pesarin

Candidato: Marco Scutari

ANNO ACCADEMICO 2006–2007

”It’s not being stubborn
when you’re right.”

Jaenelle Angelline

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

In letteratura esistono molte strutture formali che, sotto ipotesi diverse, permettono di rappresentare ed analizzare l'informazione contenuta in un insieme di dati riguardanti un fenomeno in esame. Tra questi i *network bayesiani*, che rappresentano la convergenza della *metodologia statistica* (che descrive il fenomeno sulla base dell'informazione contenuta nelle osservazioni) e dell'*intelligenza artificiale* (che si pone l'obiettivo di automatizzare il trattamento dei dati per mezzo di calcolatori), risultano interessanti per la loro capacità di esprimere in modo semplice insiemi di relazioni complesse.

Un network bayesiano descrive una distribuzione di probabilità multivariata definita su un insieme di variabili aleatorie attraverso due componenti:

- un *grafo orientato aciclico*, in cui i *nodi* rappresentano le variabili aleatorie del dominio e gli *archi* rappresentano relazioni di dipendenza condizionale.

- un insieme di *distribuzioni locali* di probabilità, ciascuna associata ad una variabile e condizionata ad ogni configurazione dei suoi *genitori* (ovvero dei nodi con archi orientati in direzione del nodo *figlio* associato alla variabile in questione).

In particolare verranno trattati network bayesiani discreti, in cui le variabili aleatorie seguono una distribuzione multinomiale. I loro parametri identificano le distribuzioni di probabilità locali, rappresentate da tabelle di probabili-

tà condizionate dei nodi figlio rispetto ad ogni combinazione di valori dei suoi genitori.

La selezione di un modello di questo tipo si traduce nell'*apprendimento automatico* della struttura del grafo (supposta non nota) e, condizionatamente a quest'ultima, nella stima delle probabilità associate alle variabili aleatorie. In letteratura questa operazione viene eseguita applicando ad un insieme di dati osservati due classi di metodi: uno basato sulla classificazione della bontà complessiva del modello tramite un punteggio (metodi *score-based*), l'altro sull'analisi locale dei vincoli di indipendenza condizionale (metodi *constraint-based*).

Gli obiettivi di questa tesi sono:

1. evidenziare la relazione che lega i criteri decisionali che caratterizzano i metodi *score-based* e *constraint-based* in ambito discreto, e che porta alla scelta dell'informazione reciproca come indicatore statistico su cui fondare la selezione del modello.

2. confrontare tramite simulazione la stima parametrica dell'informazione reciproca, utilizzata comunemente in letteratura, con quella non parametrica per evidenziare le migliori proprietà di quest'ultima.

La tesi è strutturata nel modo seguente.

Nel **Capitolo 1** (*Network bayesiani*) vengono caratterizzati i network bayesiani, le ipotesi su cui si fondano e le loro proprietà.

Nel **Capitolo 2** (*Informazione ed entropia*) si definiscono alcune quantità della teoria dell'informazione, con particolare attenzione all'informazione reciproca (*mutual information*) ed alle sue relazioni con i principali indicatori di indipendenza statistica utilizzati nell'ambito discreto.

Nel **Capitolo 3** (*Statistica non parametrica*) vengono introdotte la *distribuzione di permutazione* e la *combinazione non parametrica*, che verranno utilizzate per la stima dell'informazione reciproca.

Nel **Capitolo 4** (*Apprendimento di network bayesiani*) vengono esplorati gli algoritmi per l'apprendimento della struttura dei network bayesiani in ambito discreto, raccolti nelle classi *score-based* e *constraint-based*.

Nel **Capitolo 5** (*Informazione reciproca e network bayesiani*) vengono sviluppati i due punti centrali di questa tesi, ovvero la relazione che lega le due classi di metodi esposte nel capitolo precedente sulla base dell'informazione reciproca, e la stima di quest'ultima in ambito parametrico e non parametrico (studiata tramite simulazione).

Nelle **Appendici** (*Implementazione* e *Problemi aperti*) viene riportato l'ambiente, il codice utilizzato nelle simulazioni ed i problemi aperti incontrati nella scrittura di questa tesi.

Simboli e notazioni

Variabili aleatorie e probabilità

$\mathbf{U}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$	insiemi di variabili aleatorie
$X, Y, Z, X_1, \ldots, X_n$	variabili aleatorie
$X \perp Y$	indipendenza stocastica
$X \perp Y \mid Z,\ X \perp_P Y \mid Z$	indipendenza stocastica condizionale
$XY \stackrel{d}{=} X \cdot Y$	uguaglianza in distribuzione
$\mathbb{X}, \mathbb{Y}, \mathbb{Z}$	dominio delle variabili X, Y, Z
$\lvert\mathbb{X}\rvert, \lvert\mathbb{Y}\rvert, \lvert\mathbb{Z}\rvert$	cardinalità dei domini
$\boldsymbol{\theta}_X$	vettore dei parametri di X
$\lvert\boldsymbol{\theta}_X\rvert$	numero dei parametri di X

Teoria dei grafi

$\mathcal{G} = (V, A)$	grafo con nodi V ed archi A
u, v, w	nodi di una connesione
$A, B, \ldots,$	nodi di un grafo
$X \perp_G Y \lvert Z$	separazione grafica
Π_X	genitori di X

Teoria dell'informazione

$\mathrm{I}(x),\ i \in \mathbb{X}$	informazione di Shannon
$\mathrm{H}(X)$	entropia

$\mathrm{MI}\,(X, Y)$	informazione reciproca
$\mathrm{D}\,(P \parallel Q)$	distanza di Kullback-Leibler

Statistica non parametrica

$\mathbf{X} = \{x_1, \ldots, x_n\}$	dati osservati
$\mathbf{X}^*$	permutazione dei dati
$\mathcal{X}_{/X}$	spazio delle permutazioni
$H_{0_i}, H_{1_i}, i = 1, \ldots, k$	ipotesi (nulle ed alternative) parziali
λ	livello di significatività osservato
$\lambda_1, \ldots, \lambda_B$	livelli di significatività delle permutazioni

Tabelle di contignenza

$n_{ij},\ i \in \mathbb{X}, j \in \mathbb{Y}$	numerosità delle celle
$n_{i+},\ n_{+j}$	numerosità marginali
$\pi_{ij},\ i \in \mathbb{X}, j \in \mathbb{Y}$	probabilità delle celle
$\pi_{i+},\ \pi_{+j}$	probabilità marginali
$f_{ij},\ i \in \mathbb{X}, j \in \mathbb{Y}$	frequenze attese
n	numerosità campionaria

Capitolo 1

Network bayesiani

I network bayesiani sono caratterizzati dall'interazione tra un *grafo orientato aciclico* (*directed acyclic graph* o DAG) ed una distribuzione di probabilità su un insieme di variabili aleatorie $\mathbf{U} = \{X_1, \ldots, X_n\}$; gli aspetti probabilistici sono fondati sulle proprietà grafiche, che a loro volta sono legate alla distribuzione stocastica dei dati.

1.1 Aspetti grafici

Un *grafo*, indicato con $\mathcal{G} = (V, A)$, è una struttura composta da un insieme di *nodi* o *vertici* $V = \{v_1, v_2, \ldots, v_n\}$ e da un insieme di *archi* $A \subset V \times V$ che li collegano. Nel caso dei network bayesiani gli archi sono *orientati* ($v_1 \rightarrow v_2$), ovvero si ha che $(v_1, v_2) \in A \wedge (v_2, v_1) \notin A$; si parla quindi di *grafi orientati*.

La direzionalità di questa relazione permette di definire:

- *genitori* di v tutti quei nodi $u \in V : (u, v) \in A$, ovvero i nodi u da cui parte un arco verso v.
- *figli* di v tutti quei nodi $u \in V : (v, u) \in A$, ovvero i nodi u in cui arriva un arco da v.

- *coniugi* di v tutti quei nodi $u \in V : (u, w) \in A \wedge (v, w) \in A$, ovvero tutti i nodi u che condividono un figlio con v.

- *discendenti* di v tutti quei nodi $u \in V$ per cui esiste un cammino che porta da v a u ($v \rightarrow \ldots \rightarrow u$).

- *predecessori* di v tutti quei nodi $u \in V$ per cui esiste un cammino che viceversa porta da u a v ($u \rightarrow \ldots \rightarrow v$).

Anche se due nodi non sono *adiacenti* (ovvero se non esiste un arco che li collega) si dicono *connessi* se esiste un *cammino* (*path*) che li unisce, ovvero una sequenza di nodi $v_1, v_2, \ldots, v_k$ a due a due adiacenti tale che:

$$\begin{cases} (u, v_1) \in A \\ (v_i, v_{i+1}) \in A \vee (v_{i+1}, v_i) \in A, \ i = 1, 2, \ldots, k-1 \\ (v_k, v) \in A \end{cases}$$

indipendentemente dalla direzione dei singoli archi. L'unico vincolo a questo riguardo è che nessun cammino può formare un *ciclo*, un cammino in cui il punto di partenza coincide con quello di arrivo e tutti gli archi sono concordi; in questo caso il grafo è detto *aciclico*.

Sotto queste condizioni il comportamento complessivo del grafo può essere ricostruito sulla base di tre costrutti fondamentali (Jensen, 2001), che descrivono i possibili rapporti tra due nodi non adiacenti (u, w):

- *connessione seriale* ($u \rightarrow v \rightarrow w$): u influenza v, che a sua volta influenza w. Dato che u non influenza direttamente w, se lo stato di v è noto u e w sono tra loro indipendenti.

- *connessione divergente* ($u \leftarrow v \rightarrow w$): v influenza u e w, dato che sono suoi figli. Anche in questo caso se lo stato di v è noto u e w sono tra loro indipendenti, poichè u non influenza direttamente w,

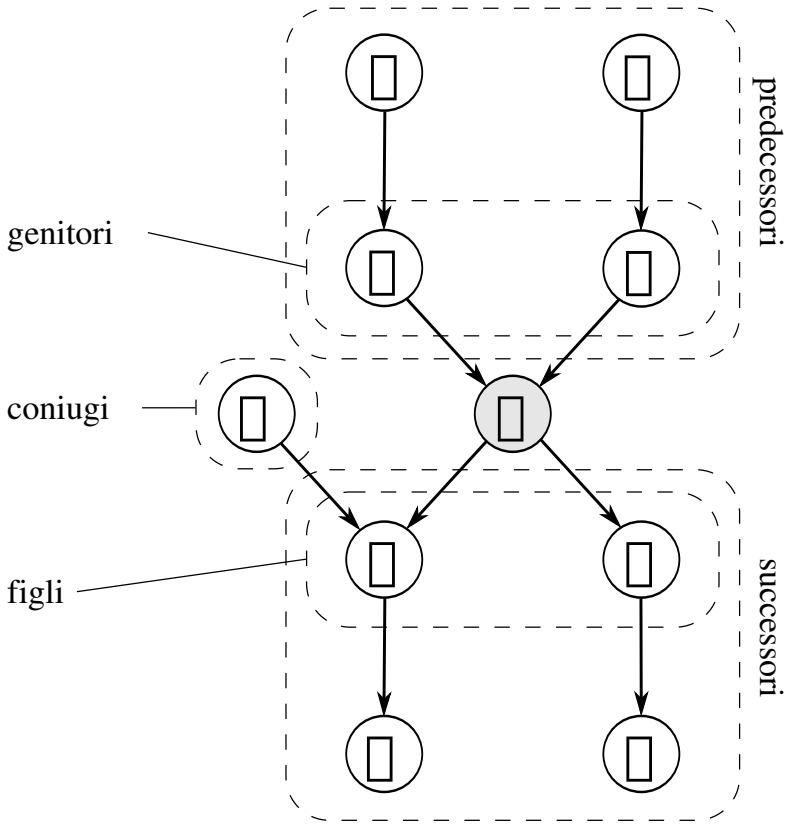


Figura 1.1: Grafo orientato aciclico.

- *connessione convergente* ($u \rightarrow v \leftarrow w$): u e w influenzano contemporaneamente v; se il suo stato non è noto si possono considerare indipendenti, dato che la somma delle loro influenze non è vincolata dal valore di v.

1.2 Indipendenza stocastica e separazione grafica

Un grafo orientato aciclico è in grado di esprimere le relazioni che legano le variabili aleatorie di un modello probabilistico tramite *dipendenze* ed *indipendenze condizionali*, e permette quindi di trattare in modo intercambiabile nodi del grafo e variabili aleatorie.

L'indipendenza condizionale tra tre insiemi disgiunti di variabili $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, è

caratterizzata in modo completo e generale da quattro assiomi (Pearl, 1988):

SIMMETRIA	$\mathbf{X} \perp \mathbf{Y} \,|\, \mathbf{Z} \Leftrightarrow \mathbf{Y} \perp \mathbf{X} \,|\, \mathbf{Z}$
DECOMPOSIZIONE	$\mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \,|\, \mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{Y} \,|\, \mathbf{Z} \wedge \mathbf{X} \perp \mathbf{W} \,|\, \mathbf{Z}$
UNIONE DEBOLE	$\mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \,|\, \mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{Y} \,|\, (\mathbf{Z} \cup \mathbf{W})$
CONTRAZIONE	$\mathbf{X} \perp \mathbf{Y} \,|\, \mathbf{Z} \wedge \mathbf{X} \perp \mathbf{W} \,|\, (\mathbf{Z} \cup \mathbf{Y}) \Rightarrow \mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \,|\, \mathbf{Z}$

quindi per qualunque relazione li soddisfi vale:

$$\mathrm{P}(\mathbf{X} \,|\, \mathbf{Y}, \mathbf{Z}) = \mathrm{P}(\mathbf{X} \,|\, \mathbf{Z}) \iff \mathbf{X} \perp \mathbf{Y} \,|\, \mathbf{Z}$$

Nei grafi orientati questa corrispondenza si ha per la *d-separazione* (*direction dependent separation*), che è definita nel modo seguente:

> **Definizione (d-separazione).** Siano $\mathbf{X}$, $\mathbf{Y}$ e $\mathbf{Z}$ tre insiemi disgiunti di nodi in un grafo orientato aciclico; $\mathbf{X}$ e $\mathbf{Y}$ si dicono *d-separati* da $\mathbf{Z}$ ($\mathbf{X} \perp_G \mathbf{Y} \,|\, \mathbf{Z}$) se non esiste un cammino tra $\mathbf{X}$ e $\mathbf{Y}$ tale che:
>
> - ogni nodo con archi convergenti appartiene a $\mathbf{Z}$ o ha un discendente che appartiene a $\mathbf{Z}$.
> - qualsiasi altro nodo non appartiene a $\mathbf{Z}$.

La sua applicazione è evidente nel caso delle tre connessioni definite precedentemente: nelle connessioni seriali e divergenti i nodi u e w sono d-separati da v (a causa della seconda condizione), mentre nella connessione convergente u e w sono d-separati solo se v non è *istanziata* (ovvero se il suo stato non è noto e non è oggetto di condizionamento). In altre parole la presenza (o l'assenza, nel caso delle connessioni convergenti) di *evidenza* sullo stato di v o di un condizionamento esplicito rispetto ai suoi possibili valori *blocca* il flusso di informazione tra u e w, rendendole indipendenti.

La d-separazione inoltre permette di stabilire una relazione tra grafi orientati e modelli probabilistici, da cui discende la definizione formale dei *network bayesiani*:

Definizione (mappe). Un grafo orientato G è una *mappa di dipendenza* (*dependency map* o *d-map*) di un modello probabilistico P se esiste una corrispondenza biunivoca tra le variabili aleatorie $\mathbf{U}$ del modello ed i nodi V del grafo, e se per ogni possibile terna $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ si ha

$$\mathbf{X} \perp_P \mathbf{Y} \,|\, \mathbf{Z} \Longrightarrow \mathbf{X} \perp_G \mathbf{Y} \,|\, \mathbf{Z}$$

ovvero che l'indipendenza condizionale implica la separazione grafica. Viceversa G è una *mappa di indipendenza* (*independency map* o *i-map*) di P se

$$\mathbf{X} \perp_P \mathbf{Y} \,|\, \mathbf{Z} \Longleftarrow \mathbf{X} \perp_G \mathbf{Y} \,|\, \mathbf{Z}$$

Infine G è una *mappa perfetta* (*perfect map*) di P se

$$\mathbf{X} \perp_P \mathbf{Y} \,|\, \mathbf{Z} \Longleftrightarrow \mathbf{X} \perp_G \mathbf{Y} \,|\, \mathbf{Z}$$

ovvero se è sia una *i-map* che una *d-map*; in questo caso P si dice *isomorfo* rispetto a G o *causale*.

Definizione (network bayesiani). Sia P una distribuzione di probabilità su un insieme di variabili $\mathbf{U}$; allora un grafo orientato aciclico $\mathcal{G} = (\mathbf{U}, A)$ è un *network bayesiano* (*bayesian network*) di P se e solo se $\mathcal{G}$ è una mappa di indipendenza minimale di P.

La d-separazione è anche alla base della definizione di *Markov blanket* per i network bayesiani, che individua per ogni nodo l'intorno $\mathbf{S} \subset \mathbf{U}$ che lo d-separa dal resto del grafo e che sotto condizioni di regolarità relativamente blande può

essere ricondotto al *Markov boundary* (che in quel caso è unico):

Definizione (Markov blanket). Si dice *Markov blanket* di X ($Bl(X)$) di un elemento $X \in \mathbf{U}$ ogni sottoinsieme $S \subset \mathbf{U}$ di elementi per cui:

$$X \perp_G (\mathbf{U} - \mathbf{S} - X) \,|\, \mathbf{S}, \qquad X \notin \mathbf{S}$$

Se S è minimale (ovvero nessun suo sottoinsieme è ancora un Markov blanket) è detto *Markov boundary* ($B_I(X)$).

In ogni network bayesiano l'unione dei seguenti nodi costituisce un Markov blanket di X: i suoi genitori, i suoi figli ed i suoi coniugi.

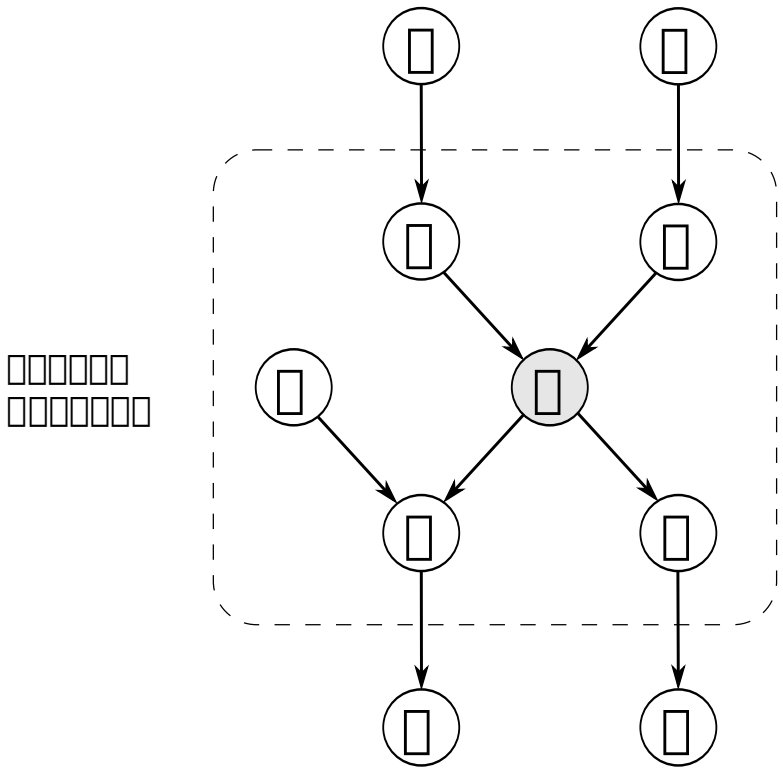


Figura 1.2: *Markov blanket* del nodo A dato il network bayesiano $\mathcal{G}$.

1.3 Aspetti probabilistici

La distribuzione di probabilità congiunta su $\mathbf{U} = \{X_1, \ldots, X_n\}$ in generale non è nota o è difficilmente quantificabile per via delle interazioni tra le variabili ed il grande numero di parametri coinvolti.

Tuttavia i network bayesiani (grazie alla d-separazione) ne permettono una rappresentazione compatta, espressa in funzione delle singole variabili X_i e dei loro genitori Π_{X_i}:

$$\mathrm{P}(\mathbf{U}) = \prod_{X_i \in \mathbf{U}} \mathrm{P}(X_i \mid \Pi_{X_i})$$

Questa scomposizione (detta *chain rule*, e nota anche come *condizione di markovianità*) si basa sull'indipendenza di ogni variabile dai suoi non-discendenti condizionatamente ai suoi genitori, che permette di esplicitare ricorsivamente le singole componenti partendo da un qualsiasi nodo A senza figli (Jensen, 2001):

$$\begin{aligned}\mathrm{P}(\mathbf{U}) &= \mathrm{P}(A \mid \mathbf{U} - \{A\})\,\mathrm{P}(\mathbf{U} - \{A\}) \\ &= \mathrm{P}(A \mid \Pi_A)\,\mathrm{P}(\mathbf{U} - \{A\})\end{aligned}$$

la cui esistenza è garantita dall'aciclicità del grafo.

In questo modo è possibile rappresentare e trattare la distribuzione congiunta delle variabili tramite un insieme di *distribuzioni di probabilità locali*, riducendo la dimensionalità del problema e facilitando la specificazione del modello probabilistico. Le interazioni tra le variabili sono rispecchiate dalla struttura delle dipendenze condizionali, e non necessitano formalizzazioni che ne limiterebbero la versatilità e la potenza (come accade nei modelli lineari).

In particolare in un network bayesiano discreto, dove ogni variabile segue una distribuzione multinomiale, la distribuzione di probabilità congiunta ha un numero

di parametri pari al prodotto del numero delle possibili combinazioni di modalità:

$$\prod_{i \in \mathbf{U}} |\boldsymbol{\theta}_{X_i}|$$

mentre dopo la scomposizione il numero di parametri dipende in larga parte dal numero di configurazioni dei genitori di $X_1, \ldots, X_n$:

$$\sum_{i \in \mathbf{U}} \prod_{j \in \Pi_{X_i}} |\boldsymbol{\theta}_{X_i}|$$

Un altro aspetto che è importante sottolineare è come ogni distribuzione congiunta possa essere ricondotta a numerose configurazioni di probabilità (e quindi essere associata a diversi grafi orientati) grazie alla commutatività della *chain rule*:

$$\mathrm{P}(A)\,\mathrm{P}(B \mid A) = \mathrm{P}(A, B) = \mathrm{P}(B)\,\mathrm{P}(A \mid B)$$

dando luogo a delle *classi di equivalenza* (Verma and Pearl, 1991) nello spazio dei modelli causali, ovvero sotto ipotesi di isomorfismo:

> **Definizione (equivalenza markoviana).** Due modelli causali definiti su uno stesso insieme di variabili $\mathbf{U}$ sono *equivalenti in senso markoviano* (*Markov equivalent*) se contengono gli stessi archi, indipendentemente dalla loro direzione, e le stesse connessioni convergenti.

All'interno di ognuna di queste classi i singoli network bayesiani non sono quindi statisticamente distinguibili, dato che corrispondono a parametrizzazioni equivalenti della stessa funzione di verosimiglianza (Chickering, 1995).

Capitolo 2

Informazione ed entropia

2.1 Definizioni e proprietà

2.1.1 Entropia

La quantità informazione contenuta in una variabile aleatoria discreta può essere quantificata in molti modi. Uno di questi è l'*informazione di Shannon*, che è definita come:

$$\mathrm{I}(X) = -\log_2 \mathrm{P}(X = x) = -\log_2 \mathrm{P}(x), \quad x \in \mathbb{X}$$

Il suo valore atteso nel caso delle variabili discrete è detto *entropia*:

$$\mathrm{H}(X) = \mathrm{E}(-\log_2 X) = -\sum_{x \in \mathbb{X}} \mathrm{P}(x) \log_2 \mathrm{P}(x) = \sum_{x \in \mathbb{X}} \mathrm{P}(x) \log_2 \frac{1}{\mathrm{P}(x)}$$

e può anche essere definito congiuntamente per due (o più) variabili:

$$\mathrm{H}(X, Y) = -\sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \mathrm{P}(x, y) \log_2 \mathrm{P}(x, y) = \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \mathrm{P}(x, y) \log_2 \frac{1}{\mathrm{P}(x, y)}$$

In quest'ultimo caso l'entropia può essere scomposta esplicitando l'*entropia*

condizionale, che esprime la relazione tra le variabili in funzione della quantità di informazione apportata dalla distribuzione condizionata:

$$\begin{cases} \mathrm{H}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y \mid X) = \mathrm{H}(Y) + \mathrm{H}(X \mid Y) \\ \mathrm{H}(Y \mid X) = -\sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \mathrm{P}(x,y) \log_2 \mathrm{P}(y \mid x) \\ \mathrm{H}(X \mid Y) = -\sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \mathrm{P}(x,y) \log_2 \mathrm{P}(x \mid y) \end{cases}$$

L'entropia gode delle seguenti proprietà:

- è finita e positiva in casi non degeneri, ovvero con $\mathrm{P}(x) \in (0,1), \forall x \in \mathbb{X}$; è comunque non negativa.
- l'entropia condizionata è inferiore all'entropia non condizionata, a meno che le variabili in questione non siano stocasticamente indipendenti:

$$\mathrm{H}(Y \mid X) \leqslant \mathrm{H}(Y)$$

- anche se formalmente andrebbe espressa in termini di $\log_2$, la base effettivamente utilizzata per il logaritmo è irrilevante dato che:

$$\mathrm{H}_b(X) = (\log_b a)\, \mathrm{H}_a(X)$$

2.1.2 Informazione reciproca

Un altro modo per misurare la forza della relazione tra due variabili è l'*informazione reciproca* (*mutual information*), che è definita come:

$$\begin{aligned} \mathrm{MI}(X,Y) &= \mathrm{E}_{\{X,Y\}} \left(\log \frac{XY}{X \cdot Y} \right) \\ &= \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \mathrm{P}(x,y) \log \frac{\mathrm{P}(x,y)}{\mathrm{P}(x)\,\mathrm{P}(y)} \end{aligned}$$

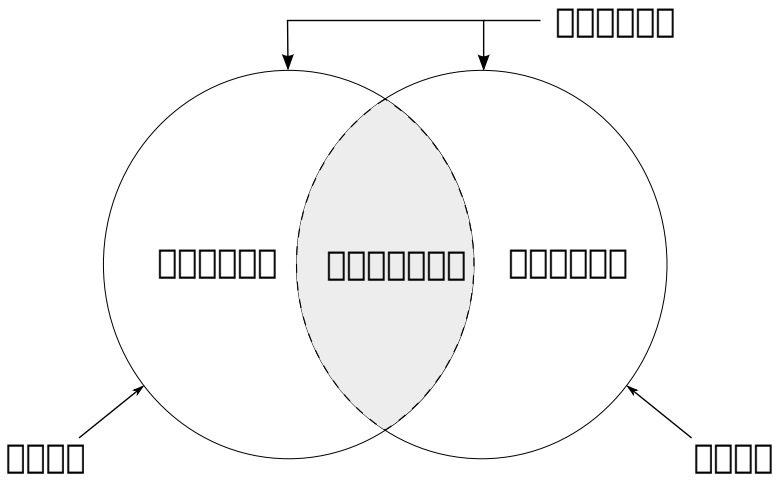


Figura 2.1: Rapporto tra entropia, entropia condizionata ed informazione reciproca.

e che non è altro che una distanza di Kullback-Leibler (detta anche *entropia relativa*) tra la distribuzione congiunta ed il prodotto delle distribuzioni marginali (Cover and Thomas, 2006):

$$\begin{cases} \mathrm{MI}\,(X,Y) = \mathrm{D}\,(\mathrm{P}\,(X,Y)\,\|\mathrm{P}\,(X)\,\mathrm{P}\,(Y)) \\ \mathrm{D}\,(P \parallel Q) = \sum_{i\in\mathbb{X}} \sum_{j\in\mathbb{Y}} p_{ij} \log \dfrac{p_{ij}}{q_{ij}} \end{cases}$$

Anche questa quantità può essere espressa in forma condizionata:

$$\begin{aligned} \mathrm{MI}\,(X,Y \mid Z) &= \mathrm{E}_{\{X,Y,Z\}} \left(\log \frac{XY \mid Z}{X \mid Z \cdot Y \mid Z} \right) \\ &= \sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} \sum_{z\in\mathbb{Z}} \mathrm{P}\,(x,y,z) \log \frac{\mathrm{P}\,(x,y \mid z)}{\mathrm{P}\,(x \mid z)\,\mathrm{P}\,(y \mid z)} \\ &= \sum_{z\in\mathbb{Z}} \mathrm{P}\,(z) \sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} \mathrm{P}\,(x,y \mid z) \log \frac{\mathrm{P}\,(x,y \mid z)}{\mathrm{P}\,(x \mid z)\,\mathrm{P}\,(y \mid z)} \\ &= \sum_{z\in\mathbb{Z}} \mathrm{P}\,(z)\,\mathrm{MI}\,(X,Y \mid z) \end{aligned}$$

ed in questo caso è detta *informazione reciproca condizionale*.

L'informazione reciproca, sia nella sua forma originale che in quella condizionata, gode delle seguenti proprietà:

- è finita e non negativa in casi non degeneri
- è nulla se e solo se le variabili in questione sono stocasticamente indipendenti, in virtù della concavità del logaritmo e della disuguaglianza di Jensen.
- è invariante rispetto all'ordine delle variabili ($\mathrm{MI}(X,Y) = \mathrm{MI}(Y,X)$).
- può essere scritta in funzione dell'entropia:

$$\begin{aligned}\mathrm{MI}(X,Y) &= \mathrm{H}(X) - \mathrm{H}(X \mid Y) \\ &= \mathrm{H}(Y) - \mathrm{H}(Y \mid X) \\ &= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y)\end{aligned}$$

2.2 Applicazioni statistiche

Le quantità derivate dall'informazione di Shannon, essendo funzioni della distribuzione di probabilità sottostante ai dati, sono facilmente utilizzabili in applicazioni statistiche, ed in particolare nelle verifiche di ipotesi.

Ad esempio l'informazione reciproca è un buon indicatore per l'indipendenza tra variabili aleatorie discrete:

$$H_0 : XY \stackrel{d}{=} X \cdot Y \qquad\qquad H_1 : XY \stackrel{d}{\neq} X \cdot Y$$

e in particolare per l'indipendenza tra due variabili con distribuzione multinomiale, organizzate in una tabella di contingenza di probabilità $\{\pi_{ij}\}$ per $i \in \mathbb{X}, j \in \mathbb{Y}$

e con probabilità marginali $\{\{\pi_{i+}\}, \{\pi_{+j}\}\}$:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \qquad\qquad H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$$

Il suo comportamento in quest'ultimo caso è assimilabile a quello del X^2 di Pearson, del log-rapporto di verosimiglianza e del test esatto di Fisher, che sono alcuni test di indipendenza comunemente utilizzati (Agresti, 2002).

2.2.1 X^2 di Pearson

L'informazione reciproca è approssimativamente equivalente al X^2 di Pearson sotto l'ipotesi nulla di indipendenza (Kullback, 1959); questo risultato, ampiamente utilizzato in letteratura, permette di utilizzare il χ^2 come distribuzione asintotica per $\mathrm{MI}\,(X, Y)$.

La relazione tra queste due quantità si può ricavare sfruttando l'approssimazione lineare del logaritmo in $x = 1$:

$$\log x \leqslant x - 1, \qquad x > 0$$

che permette di costruire il seguente intervallo:

$$\frac{a-b}{a} \leqslant \log\left(\frac{a}{b}\right) \leqslant \frac{a-b}{b}$$

dove l'uguaglianza vale se e solo se $a = b$. Sulla base di questo risultato il logaritmo può essere approssimato con la media dei suoi vincoli:

$$\log\left(\frac{a}{b}\right) \simeq \frac{1}{2}\left(\frac{a-b}{a} + \frac{a-b}{b}\right) = \frac{a^2 - b^2}{2ab}$$

che porta alla seguente approssimazione sotto l'ipotesi di indipendenza (ovvero

nel caso in cui $\mathrm{P}(x,y) \simeq \mathrm{P}(x)\,\mathrm{P}(y)$):

$$\begin{aligned}\mathrm{MI}(X,Y) &= -\sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} \mathrm{P}(x,y)\log\frac{\mathrm{P}(x)\,\mathrm{P}(y)}{\mathrm{P}(x,y)} \\ &\simeq -\sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} \mathrm{P}(x,y)\,\frac{\mathrm{P}(x,y)^2-(\mathrm{P}(x)\,\mathrm{P}(y))^2}{2\mathrm{P}(x,y)\,\mathrm{P}(x)\,\mathrm{P}(y)} \\ &\simeq \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} \frac{(\mathrm{P}(x,y)-\mathrm{P}(x)\,\mathrm{P}(y))^2}{\mathrm{P}(x)\,\mathrm{P}(y)} = X^2\end{aligned}$$

La precisione di questa approssimazione si riduce rapidamente al crescere della differenza tra la probabilità congiunta ed il prodotto delle probabilità marginali; tuttavia sotto l'ipotesi nulla permette di definire la distribuzione approssimata per l'informazione reciproca:

$$\mathrm{MI}(X,Y) \dot{\sim} \chi^2_{(|\mathbb{X}|-1)(|\mathbb{Y}|-1)}$$

2.2.2 Log-rapporto di verosimiglianza

La verifica dell'ipotesi di indipendenza può essere effettuata anche confrontando l'aderenza del modello ipotizzato ai dati osservati rispetto quello stimato partendo dai dati stessi; si può quindi far ricorso al *log-rapporto di verosimiglianza*, le cui proprietà sono ampiamente descritte in letteratura (Azzalini, 2001).

Le variabili aleatorie multinomiali in generale hanno densità:

$$f(\{n_i\};\{\pi_i\}) = \frac{n!}{\prod_i n_i!}\prod_i \pi_i^{n_i} \qquad\qquad i \in \mathbb{X}$$

dove $\{n_i\}$ e $\{\pi_i\}$ sono rispettivamente le numerosità e le probabilità associate alle varie modalità ed n è la numerosità campionaria. La log-verosimiglianza

corrispondente nel caso di una tabella di contingenza è:

$$\log L(\{\pi_{ij}\}) = \log n! - \sum_i \sum_j \log n_{ij}! + \sum_i \sum_j n_{ij} \log \pi_{ij}$$

quindi il log-rapporto di verosimiglianza è pari a:

$$\begin{aligned} G^2 &= -2 \left[\left(\log n! - \sum_i \sum_j \log n_{ij}! + \sum_i \sum_j n_{ij} \log \pi_i \pi_j \right) \right. \\ &\quad \left. - \left(\log n! - \sum_i \sum_j \log n_{ij}! + \sum_i \sum_j n_{ij} \log \pi_{ij} \right) \right] \\ &= 2 \sum_i \sum_j n_{ij} \log \frac{\pi_{ij}}{\pi_i \pi_j} \\ &= 2n \sum_i \sum_j \pi_{ij} \log \frac{\pi_{ij}}{\pi_i \pi_j} = 2n \mathrm{MI}\,(X, Y) \end{aligned}$$

e risulta essere direttamente proporzionale all'informazione reciproca. Pertanto anche in questo caso si ha che quest'ultima, a meno di un fattore che non dipende dalla distribuzione dei dati, ha distribuzione asintotica χ^2:

$$2n \mathrm{MI}\,(X, Y) \dot{\sim} \chi^2_{(|\mathbb{X}|-1)(|\mathbb{Y}|-1)}$$

coerentemente con il fatto che la differenza tra X^2 e G^2 converge a zero in probabilità (Agresti, 2002). Questa uguaglianza si mantiene anche se si opera condizionando rispetto ad una o più variabili:

$$\begin{aligned} G^2\,(X, Y \mid Z) &= \sum_{z \in \mathbb{Z}} G^2(X, Y \mid Z = z) = \sum_{z \in \mathbb{Z}} 2n_z \mathrm{MI}\,(X, Y \mid Z = z) \\ &= 2n \sum_{z \in \mathbb{Z}} \mathrm{P}\,(Z = z)\, \mathrm{MI}\,(X, Y \mid Z = z) = 2n \mathrm{MI}\,(X, Y \mid Z) \end{aligned}$$

come accade, per esempio, nell'analisi di tabelle di contingenza marginalizzate.

2.2.3 Test esatto di Fisher

Il test esatto di Fisher (Agresti, 2002) è basato sulla distribuzione ipergeometrica multidimensionale, che descrive la probabilità di una tabella di contingenza $\{n_{ij}\}$ condizionatamente alle sue numerosità marginali $\{n_{i+}\}$ e $\{n_{+j}\}$:

$$F(\{n_{ij}\}) = \frac{\prod_i n_{i+}! \prod_j n_{+j}!}{n! \prod_{ij} n_{ij}!}$$

Queste ultime sono anche statistiche sufficienti per le probabilità marginali $\{\pi_{i+}\}$ e $\{\pi_{+j}\}$, e permettono di individuare la distribuzione esatta sullo spazio delle possibili tabelle sotto l'ipotesi nulla di indipendenza. L'ordinamento in probabilità che ne consegue privilegia quindi le tabelle di contingenza con frequenze vicine a quelle attese:

$$f_{ij} = \frac{n_{i+} n_{+j}}{n}$$

rispetto a quelle in cui si evidenzia una maggiore dipendenza tra le variabili.

Attraverso l'approssimazione di Stirling

$$\log n! = n \log n - n$$

derivata dall'omonima espansione in serie della funzione $\Gamma(x)$, si ricava che il logaritmo del test esatto di Fisher è:

$$\begin{aligned}
\log F(\{n_{ij}\}) \simeq & \sum_i (n_{i+} \log n_{i+} - n_{i+}) + \sum_j (n_{+j} \log n_{+j} - n_{+j}) \\
& - (n \log n - n) - \sum_i \sum_j (n_{ij} \log n_{ij} - n_{ij}) \\
= & \sum_i n_{i+} \log n_{i+} + \sum_j n_{+j} \log n_{+j} \\
& - n \log n - \sum_i \sum_j n_{ij} \log n_{ij}
\end{aligned}$$

Sviluppando ognuna delle sommatorie in modo da esplicitare l'entropia si ottiene:

$$\begin{aligned}
\sum_i n_{i+} \log n_{i+} &= n \sum \frac{n_{i+}}{n} \log \frac{n_{i+}}{n} n \\
&= n \sum \pi_i \log n\pi_i \\
&= n \left(\sum \pi_i \log n + \sum \pi_i \log \pi_i \right) \\
&= n \log n + n \sum \pi_i \log \pi_i \\
&= n \log n - n\mathrm{H}(X) \\
\sum_j n_{+j} \log n_{+j} &= n \log n - n\mathrm{H}(Y) \\
\sum_i \sum_j n_{ij} \log n_{ij} &= n \log n - n\mathrm{H}(X, Y)
\end{aligned}$$

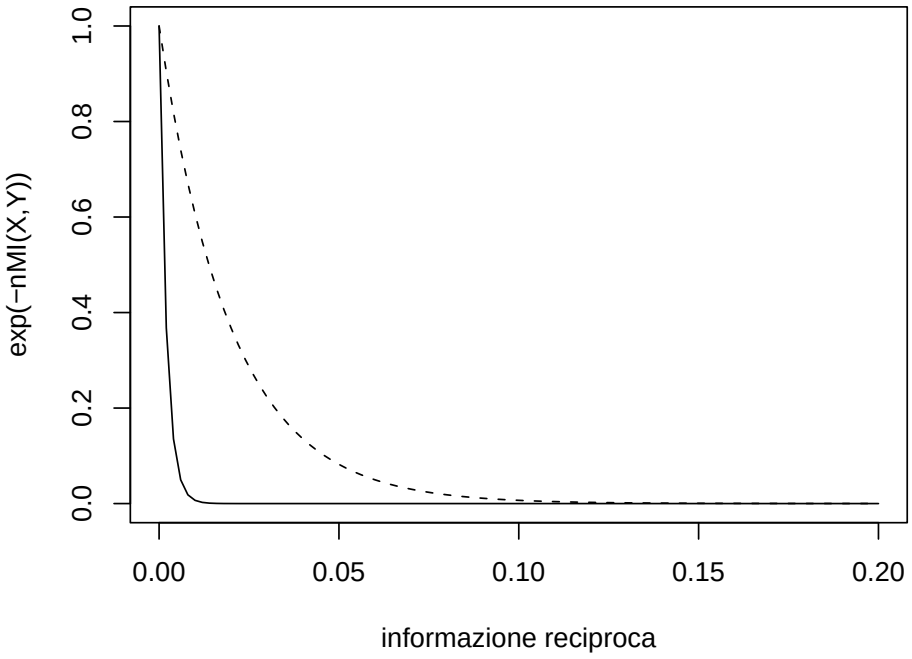


Figura 2.2: Approssimazione del test di Fisher. La linea continua è calcolata con una numerosità campionaria pari a $n = 500$, quella tratteggiata con $n = 50$.

Sostituendo le quantità così ottenute nella formula originale si ottiene la relazione che lega il test di Fisher e l'informazione reciproca.

$$\log F(\{n_{ij}\}) \simeq -n\left(\mathrm{H}\left(X\right)+\mathrm{H}\left(Y\right)-\mathrm{H}\left(X,Y\right)\right)=-n\mathrm{MI}\left(X,Y\right)$$

o equivalentemente:

$$F(\{n_{ij}\}) \simeq e^{-n\mathrm{MI}(X,Y)}$$

La coerenza tra l'ordinamento in probabilità imposto dalla distribuzione ipergeometrica multidimensionale e l'approssimazione basata sull'informazione reciproca risulta evidente dal grafico alla pagina precedente e dall'analisi del comportamento al limite; quando le variabili X e Y sono dipendenti infatti si ha:

$$\lim_{\mathrm{MI}(X,Y)\to+\infty} e^{-n\mathrm{MI}(X,Y)} = 0^{+}$$

mentre in caso di indipendenza si ha:

$$\lim_{\mathrm{MI}(X,Y)\to 0^{+}} e^{-n\mathrm{MI}(X,Y)} = 1^{-}$$

Capitolo 3

Statistica non parametrica

La *statistica non parametrica* è una branca della statistica che si occupa dell'inferenza (stima e verifica di ipotesi) nei casi in cui la distribuzione sottostante ai dati non è nota (ad esempio in distribuzioni multivariate con componenti tra loro dipendenti) o non è trattabile con i metodi classici (ad esempio a causa del rapporto tra numerosità campionaria e numero di parametri del modello).

In particolare ai fini di questa tesi risultano di interesse due aspetti di questa disciplina: la *distribuzione di permutazione* di un indicatore e la *combinazione non parametrica* di più statistiche test, non necessariamente indipendenti, per la verifica di un'ipotesi globale complessa.

3.1 Distribuzione di permutazione

Si supponga di voler verificare un'ipotesi di uguaglianza in distribuzione tra le variabili aleatorie relative a C gruppi:

$$H_0 : X_1 \stackrel{d}{=} X_2 \stackrel{d}{=} \ldots \stackrel{d}{=} X_C$$

applicando un indicatore $T(\mathbf{X})$, che si suppone significativo per valori alti senza perdita di generalità, ad un insieme di dati osservati $\mathbf{X} = \{x_1, \ldots, x_n\}$.

Sotto l'ipotesi nulla questi ultimi provengono tutti da una stessa distribuzione X_{H_0}, e possono essere quindi ricondotti indifferentemente ad ognuno dei gruppi $1, \ldots, C$. Questa assunzione, detta di *scambiabilità*, discende naturalmente da quella di uguaglianza in distribuzione, ed operando condizionatamente ai dati permette di creare una insieme di permutazioni delle osservazioni $\mathbf{X}^*$ statisticamente equivalenti al campione originale $\mathbf{X}$. Lo *spazio delle permutazioni* $\mathcal{X}_{/X}$ che si viene così a creare consente di stimare l'indicatore necessario alla verifica di ipotesi e di ricavarne la significatività λ tramite la sua *distribuzione di permutazione*:

$$\lambda = \mathrm{P}\left(T(\mathbf{X}^*) \geqslant T(\mathbf{X}) \,|\, \mathbf{X}\right)$$

in modo analogo a quanto accadrebbe utilizzando una distribuzione parametrica.

La distribuzione di permutazione permette una stima esatta di λ nel caso in cui essa sia valutata su tutto lo spazio delle permutazioni $\mathcal{X}_{/X}$, dato che l'insieme dei dati osservati costituisce in ogni caso una statistica sufficiente per la distribuzione della popolazione di riferimento e quindi per la statistica test T. Qualora esso fosse troppo ampio per rendere praticabile questa operazione, è comunque possibile ottenere una stima approssimata di λ la cui precisione dipende solamente dalla porzione di $\mathcal{X}_{/X}$ che viene presa in considerazione, quantificabile nel numero di permutazioni (solitamente indicato con B) utilizzate nell'inferenza.

Questa caratteristica rende preferibile l'uso della distribuzione di permutazione rispetto alle distribuzioni parametriche nei casi in cui queste siano approssimate o asintotiche, poichè la qualità della stima della significatività osservata:

- non è influenzata dall'aderenza dei dati alle assunzioni che occorrerebbe fare in ambito parametrico, come l'indipendenza o l'appartenenza ad una famiglia esponenziale regolare.

- non dipende dalla numerosità campionaria e dal tasso di convergenza della statistica alla distribuzione di riferimento.

3.2 Combinazione non parametrica

La *combinazione non parametrica* è una tecnica statistica che permette la verifica di ipotesi complesse, anche nel caso in cui le loro componenti non siano tra loro indipendenti, tramite l'analisi sullo spazio delle permutazioni e la successiva combinazione secondo funzioni opportune.

Si supponga di voler verificare un'*ipotesi globale* multivariata di uguaglianza in distribuzione:

$$H_0 : \mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_2 \stackrel{d}{=} \ldots \stackrel{d}{=} \mathbf{X}_C$$

la cui ipotesi nulla possa essere scomposta nell'intersezione di più *ipotesi parziali*:

$$H_0 : \bigcap_{i=1}^{k} \{H_{0_i}\} \qquad\qquad i = 1, \ldots, k$$

e la cui ipotesi alternativa possa essere scritta in modo analogo come unione delle relative ipotesi alternative parziali:

$$H_1 : \bigcup_{i=1}^{k} \{H_{1_i}\} \qquad\qquad i = 1, \ldots, k$$

In questo caso è possibile utilizzare gli indicatori $T_i(\mathbf{X})$ relativi alle ipotesi parziali per costruire un unico indicatore $\mathbf{T}''(\mathbf{X})$ per l'ipotesi globale tramite una *funzione di combinazione* ψ dei loro livelli di significatività osservata λ_i (Pesarin, 2001). Alcuni esempi di questo tipo di funzione sono la *funzione di Fisher*

$$\mathbf{T}''_F = -\sum_{i=1}^{k} \log \lambda_i$$

e la *funzione di Tippett*

$$\mathbf{T}''_T = \max_{i \in \{1 \ldots k\}} (1 - \lambda_i)$$

ed in generale può essere utilizzata qualsiasi quantità che goda delle seguenti

caratteristiche:

- deve essere non-crescente in ognuno dei suoi argomenti:

$$\lambda_i < \lambda_i' \Rightarrow \psi(\dots, \lambda_i, \dots) \geqslant \psi(\dots, \lambda_i', \dots) \qquad i = 1, \dots, k$$

- deve raggiungere il suo estremo superiore quando almeno uno dei suoi argomenti tende a zero:

$$\lim_{\lambda_i \to 0} \psi(\dots, \lambda_i, \dots) = \bar{\psi}$$

- per $\forall \alpha > 0$, il valore critico di ψ si assume finito e strettamente inferiore del suo valore massimo $\bar{\psi}$:

$$\mathbf{T}_\alpha < \bar{\psi}$$

Ulteriori proprietà di questa funzione derivano da quelle degli indicatori utilizzati nella verifica delle ipotesi parziali; se ad esempio questi ultimi sono dei test esatti, sono non distorti, o se almeno uno di essi è consistente, lo è anche la loro combinazione.

La distribuzione della funzione di combinazione può essere ricavata sfruttando la distribuzione di permutazione. Stimando congiuntamente tutti gli indicatori relativi alle ipotesi parziali per ogni permutazione dei dati osservati, e combinandoli tramite la funzione ψ scelta, si ottiene infatti la distribuzione di permutazione di $\mathbf{T}''(\mathbf{X})$, grazie alla quale è possibile verificare l'attendibilità l'ipotesi globale.

La valutazione simultanea di tutte le ipotesi parziali sui singoli elementi dello spazio delle permutazioni permette di mantenere i legami di dipendenza presenti all'interno dei dati e nelle ipotesi stesse, al contrario di metodi come Bonferroni e Bonferroni-Holms, e consente un controllo esatto del livello di significatività.

Capitolo 4

Apprendimento di network bayesiani

La selezione del modello nel contesto dei network bayesiani si traduce nell'individuazione della struttura del grafo orientato aciclico che meglio descrive le relazioni tra le variabili aleatorie presenti nel modello stesso. Questa operazione, nota come *apprendimento* (*learning*), è composta da due fasi:

- l'*apprendimento della struttura* del grafo, ed in particolare dei suoi archi
- l'*apprendimento dei parametri* che regolano il comportamento delle distribuzioni di probabilità

e può essere portata a termine in due modi:

- utilizzando degli esperti e conoscenze pregresse sul fenomeno in esame.
- tramite l'applicazione di opportuni algoritmi a dati osservati.
- combinando i due metodi precedenti.

Dato che la prima soluzione in molti casi non è applicabile, perchè troppo costosa o poco affidabile, questa tesi si concentrerà sui metodi di apprendimento automatico da un insieme di dati osservati.

4.1 Ipotesi di lavoro

L'individuazione della struttura di un network bayesiano può essere effettuata applicando a dei dati osservati due classi di metodi:

- la selezione del modello sulla base di un *punteggio* (*score-based methods*) che ne descriva in modo sintetico la bontà complessiva di adattamento.
- l'individuazione delle relazioni tra le variabili utilizzando dei test locali (tra sottoinsiemi di $\mathbf{U} = \{X_1, \ldots, X_n\}$), e la loro composizione attraverso gli assiomi dell'indipendenza condizionale (*constraint-based methods*).

In entrambi i casi la definizione stessa di network bayesiano impone alcune ipotesi per garantire la validità del modello, che derivano dalla sua condizione di *i-map* minimale e dal rispetto degli assiomi di indipendenza condizionale:

- *markovianità* (*markov assumption*): ogni nodo deve essere indipendentemente da tutti i suoi non-discendenti condizionatamente ai suoi genitori, in modo da poter applicare la *chain rule*.
- *causalità* (*causal sufficiency*): non devono esistere variabili latenti; in caso contrario verrebbe meno la corrispondenza biunivoca tra i nodi del grafo le variabili stesse. Per lo stesso motivo le variabili considerate devono essere tra loro distinte.
- *accuratezza* (*faithfulness*): il grafo deve essere una mappa perfetta della distribuzione di probabilità considerata; quest'ipotesi è necessaria nei metodi *constraint-based*, e viene spesso assunta anche per i metodi *score-based*.

Ai fini di questa tesi si fanno alcune ipotesi ulteriori:

- tutte le variabili devono avere distribuzione multinomiale; in questo caso esistono delle soluzioni in forma esplicita e la completezza della d-separazione è formalmente dimostrata (Meek, 1995).

- non vi devono essere dati mancanti; la loro imputazione infatti rischia di introdurre ulteriori dipendenze tra le variabili aleatorie, impedendo la fattorizzazione in probabilità locali.
- non vi devono essere parametri nulli per costruzione, per garantire la stabilità numerica dei test statistici e le loro proprietà asintotiche.

4.2 Metodi *Score-based*

I metodi *score-based* sono affini ai metodi di selezione (*stepwise regression*, *forward selection*, *backward elimination*) che si applicano usualmente ai modelli lineari (McCullagh and Nelder, 1989) o ai network markoviani (Edwards, 2000), a meno delle differenze dovute alle caratteristiche specifiche dei grafi orientati.

La caratteristica saliente di questa classe di metodi è la classificazione di ogni network bayesiano rispetto alla propria bontà di adattamento sulla base di un *punteggio*, come ad esempio la *log-verosimiglianza*:

$$Score_{ML}\left(\mathcal{G}, \mathcal{D}\right) = \log \mathrm{L}\left(\mathcal{D} \mid \mathcal{G}\right)$$

o le sue versioni penalizzate in base al numero d di parametri, l'*Akaike Information Criterion* (AIC) ed il *Bayesian Information Criterion* (BIC):

$$\begin{aligned} Score_{AIC}\left(\mathcal{G}, \mathcal{D}\right) &= \log \mathrm{L}\left(\mathcal{D} \mid \mathcal{G}\right) - d \\ Score_{BIC}\left(\mathcal{G}, \mathcal{D}\right) &= \log \mathrm{L}\left(\mathcal{D} \mid \mathcal{G}\right) - \frac{d}{2} \log N \end{aligned}$$

che privilegiano la semplicità del modello rispetto alla perfezione della stima.

In alternativa è possibile utilizzare come punteggio la probabilità a posteriori, che per il teorema di Bayes può essere scritta come:

$$Score_{Bayes}\left(\mathcal{G}, \mathcal{D}\right) = \mathrm{P}\left(\mathcal{G} \mid \mathcal{D}\right) = \frac{\mathrm{P}\left(\mathcal{D} \mid \mathcal{G}\right) \mathrm{P}\left(\mathcal{G}\right)}{\mathrm{P}\left(\mathcal{D}\right)} \propto \mathrm{P}\left(\mathcal{D} \mid \mathcal{G}\right) \mathrm{P}\left(\mathcal{G}\right)$$

Per la probabilità a priori sullo spazio dei possibili grafi viene spesso assunta una distribuzione uniforme, permettendo di ridurre ancora il confronto tra i diversi network a quello tra le loro verosimiglianze:

$$Score_{Bayes}\left(\mathcal{G}, \mathcal{D}\right) \propto \mathrm{P}\left(\mathcal{D} \,|\, \mathcal{G}\right) \equiv \mathrm{L}\left(\mathcal{D} \,|\, \mathcal{G}\right)$$

Nel caso multinomiale la probabilità a posteriori può essere scritta in forma esplicita (Cooper and Herskovits, 1992), utilizzando la coniugata Dirichlet (DeGroot, 2004) come distribuzione a priori sullo spazio parametrico ed assumendo l'indipendenza locale e globale dei parametri (Heckerman, 1999):

$$\mathrm{P}\left(\mathcal{D} \,|\, \mathcal{G}\right) = \prod_{i=1}^{n} \left(\prod_{j=1}^{|\Pi_{X_i}|} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{|X_i|} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \right)$$

dove:

- n è il numero delle variabili.
- $|\Pi_{X_i}|$ è il numero delle possibili configurazioni di Π_{X_i}.
- $|X_i|$ è il numero di possibili stati di X_i.
- gli α_{ijk} sono i parametri della distribuzione Dirichlet relativi alla variabile i, al suo valore k ed alla configurazione j dei suoi genitori.
- gli n_{ijk} sono le numerosità osservate nelle stesse condizioni.

L'altro aspetto che caratterizza i metodi *score-based* è l'algoritmo con cui viene effettuata la ricerca del modello ottimale (che massimizza il punteggio scelto) nello spazio dei grafi. Quest'ultimo è molto vasto e cresce in modo esponenziale rispetto al numero n dei nodi; anche senza tenere conto dell'orientamento degli archi ha cardinalità $2^{\frac{1}{2}n(n-1)}$. Per questa ragione occorre utilizzare algoritmi basati su scelte euristiche, che sono computazionalmente meno pesanti di un approccio

Algoritmo *HillClimbing*

$A = \varnothing$
$\mathcal{G} = (\mathbf{U}, A)$
$score = -\infty$
do {
  $maxscore = score$
  foreach $((X, Y) : X, Y \in \mathbf{U})$ {
    foreach $(A' \in \{A \cup (X, Y), A \cup (Y, X), A - \{(X, Y), (Y, X)\}\})$ {
      $\mathcal{G}' = (\mathbf{U}, A')$
      $newscore = Score\,(\mathcal{G}', \mathcal{D})$
      if $(newscore > score)$ {
        $\mathcal{G} = \mathcal{G}'$
        $score = newscore$
      }$//then$
    }$//foreach$
  }$//foreach$
} while $(score > maxscore)$
return $\mathcal{G}$

Tabella 4.1: Algoritmo di *hill climbing* per i grafi orientati.

brute force ma la cui convergenza può essere rallentata dalla presenza di massimi locali. Questa limitazione è stata superata in letteratura mediante tecniche di *random restart* o *simulated annealing* (Bouckaert, 1995).

Tra le molte scelte possibili (*best-first*, algoritmi genetici, soluzioni specifiche per gli alberi), una molto popolare è costituita dall'*hill climbing* (Pearl, 1984). Questo tipo di procedura si muove nello spazio dei grafi modificando un arco alla volta (se l'arco non è presente lo aggiunge, mentre in caso contrario lo inverte o lo elimina), e si arresta quando nessuno di questi cambiamenti produce un incremento significativo nel punteggio di riferimento. Un altro algoritmo molto utilizzato è il *K2* (Cooper and Herskovits, 1992) e la sua variante *K2SN* (Callejon, 2001), che impongono un ordinamento alle variabili aleatorie per ridurre il numero di

possibili strutture.

4.3 Metodi *Constraint-based*

I metodi *constraint-based* fondano la ricerca della struttura del network bayesiano sulla combinazione di vincoli (*constraint*) di indipendenza condizionale, sfruttando i relativi assiomi ed opportuni indicatori statistici.

Anche in questo caso la complessità computazionale risulta un problema di primaria importanza. La cardinalità dello spazio dei grafi infatti preclude una ricerca esaustiva, esattamente come accade per i metodi *score-based*, e rende necessaria la scelta di un criterio euristico che risponda ai seguenti criteri:

- minimizzi il numero di test di indipendenza condizionale, in modo da ridurre l'incidenza degli errori di prima e seconda specie e da non amplificare inutilmente la complessità computazionale dei test stessi.
- contenga la dimensione dei condizionamenti, per avere un numero sufficiente di osservazioni per ogni combinazione delle variabili condizionanti.
- utilizzi un test che sia a sua volta semplice da calcolare; che ad esempio richieda una sola scansione dei dati o che non necessiti di mantenere contemporaneamente tutti i dati in memoria.

In letteratura sono presenti vari approcci di questo tipo. Molti di essi prendono spunto dall'algoritmo *Inductive Causation* (Verma and Pearl, 1991), che ne delinea formalmente i punti chiave; tra questi l'algoritmo *PC* (Spirtes et al., 2001), il

nodi	1	2	3	4	5	6	7	8	9	10
$\lvert\mathcal{G}\rvert$	1	3	25	543	29281	378153	$\approx 10^9$	$\approx 10^{11}$	$\approx 10^{15}$	$\approx 10^{18}$

Tabella 4.2: Cardinalità dello spazio dei grafi.

Algoritmo *Inductive Causation*

1. Per ogni coppia di nodi X e Y si cerchi un insieme di nodi $\mathbf{S}_{XY}$ che li renda condizionatamente indipendenti. Se non esiste un $\mathbf{S}_{XY}$ che soddisfi questa condizione, si colleghino X e Y con un arco non orientato.

2. Per ogni coppia di nodi X e Y non adiacenti si individuino i nodi Z adiacenti ad entrambi, e si orientino gli archi in una connessione convergente ($X \rightarrow Z \leftarrow Y$) se $Z \notin \mathbf{S}_{XY}$.

3. Si orientino iterativamente gli archi rimanenti secondo le due regole seguenti:

   (a) se X e Y non sono adiacenti ma esiste un arco orientato che porta da X a Z ($X \rightarrow Z$) ed uno non orientato tra Y e Z, si orienti quest'ultimo in direzione di Y ($Z \rightarrow Y$) in modo da formare una connessione seriale.

   (b) se esiste un cammino costituito da soli archi orientati che porta da X a Z e questi nodi sono adiacenti, si orienti l'arco che li congiunge in modo da non creare un ciclo ($X \rightarrow Z$).

4. Se esiste un arco orientato tra X e Y ($X \rightarrow Y$), si orienti ogni arco tra Y e Z in direzione di quest'ultimo ($Y \rightarrow Z$) se Z non è adiacente a X.

Tabella 4.3: Algoritmo IC (Verma and Pearl, 1991).

Grow-Shrink (Margaritis, 2003), l'*Incremental Association Markov Blanket* (Tsamardinos et al., 2003) e le loro varianti.

4.3.1 Algoritmo *Inductive Causation* (IC)

Pur non essendo applicabile in forma diretta, dato che assume che la struttura del grafo sia nota a priori, l'algoritmo *Inductive Causation* (Verma and Pearl, 1991) definisce le linee guida che sono alla base dei metodi *constraint-based*.

Le quattro regole che lo compongono delineano una procedura in due fasi:

- partendo da un grafo privo di archi, nella prima regola si individuano le relazioni di indipendenza condizionale tra le variabili e si tracciano degli archi non orientati tra quelle che non risultano d-separabili.

- si orientano gli archi presenti nel grafo ricavato precedentemente confrontando il comportamento delle possibili triplette di nodi (X, Y, Z) con quello delle tre connessioni fondamentali (la seconda regola identifica quelle convergenti, la terza e la quarta quelle seriali e divergenti), e premurandosi di non introdurre dei cicli.

Sotto l'ipotesi di isomorfismo questo algoritmo permette di individuare la classe di equivalenza di network bayesiani sottostante ai dati, che viene rappresentata con un grafo parzialmente orientato. Gli archi di cui è stato possibile stabilire l'orientamento sono *invarianti* all'interno di questa classe e rappresentano le relazioni causa-effetto comuni a tutti i modelli (*genuine cause*). I restanti, pur indicando

Algoritmo *PC*

1. Si consideri un grafo non orientato completamente connesso, e per ogni coppia di nodi si fissi $\mathbf{S}_{XY} = \varnothing$ e $k = |\mathbf{S}_{XY}| = 0$.

2. Per ogni coppia di variabili X e Y si rimuova l'arco che le congiunge se e solo se per ogni sottoinsieme $\mathbf{S}$ di dimensione k dei nodi adiacenti ad X (eventualmente escludendo Y) la correlazione parziale $\rho_{XY|\mathbf{S}}$ non è significativa.
   In questo caso si assegni $\mathbf{S}_{XY} = \mathbf{S}$.

3. Se è stato rimosso almeno un arco, si incrementi k e si ripeta la regola precedente.

4. Si orientino gli archi rimanenti secondo le regole 2, 3 e 4 dell'algoritmo *Inductve Causation*.

Tabella 4.4: Algoritmo PC (Spirtes et al., 2001).

una correlazione tra i nodi che congiungono, sono delle relazioni causa-effetto solo in potenza (*potential cause*).

La ricerca di un'applicazione efficiente della prima delle due fasi è l'obiettivo primario degli algoritmi che derivano dall'*Inductive Causation*, dato che in essa si concentra la maggior parte del carico computazionale di questo metodo. Una sua applicazione abbastanza letterale, presente nell'algoritmo *PC* (Spirtes et al., 2001), presenta vari problemi di stabilità (dato il numero elevato di test, un errore nelle verifiche di ipotesi si perpetua e si amplifica nei test successivi) e richiede una numerosità campionaria molto elevata in rapporto al numero di variabili (Korb and Nicholson, 2004).

4.3.2 Algoritmo *Grow-Shrink* (GS)

L'algoritmo *Grow-Shrink* (Margaritis, 2003) sfrutta le proprietà del *Markov blanket*, che per definizione d-separa ogni nodo dal resto del grafo, per ridurre il numero di test richiesti dalla prima parte dell'*Inductive Causation*.

L'individuazione del *Markov blanket* di ogni nodo X del grafo viene eseguita

Algoritmo *Grow-Shrink Markov Blanket*
$\mathbf{S} = \varnothing$
/* *Growing phase* */
while $(\exists Y \in \mathbf{U} - \{X\} : Y \not\perp X \mid \mathbf{S})$ {
$\quad \mathbf{S} = \mathbf{S} \cup Y$
$\}//while$
/* *Shrinking phase* */
while $(\exists Y \in \mathbf{S} : Y \perp X \mid \mathbf{S} - \{Y\})$ {
$\quad \mathbf{S} = \mathbf{S} - \{Y\}$
$\}//while$
return $\mathbf{S}$

Tabella 4.5: Algoritmo GS-MB (Margaritis, 2003) per un generico nodo X.

Algoritmo *Grow-Shrink*

1. Si individui il *Markov Blanket* $Bl(X)$ di ogni nodo $X \in \mathbf{U}$.
2. All'interno di ogni *Markov Blanket* si determinino i nodi $Y \in Bl(X)$ adiacenti ad X, sfruttando il fatto che in quel caso X e Y sono dipendenti per ogni possibile sottoinsieme di $Bl(X) - Y$ (o equivalentemente di $Bl(Y) - X$).
3. Si orientino gli archi i cui nodi individuano connessioni convergenti, secondo il criterio indicato nella seconda regola dell'algoritmo *Inductive Causation*.
4. Si individuino ricorsivamente gli archi che violano l'aciclicità del grafo, rimuovendoli dal grafo in base al numero di cicli di cui fanno parte.
5. Si reinseriscano in ordine inverso gli archi eliminati nella regola precedente, orientandoli in senso opposto
6. Se esiste un cammino tra due nodi X e Y connessi da un arco non ancora orientato, si orienti l'arco che li congiunge in modo da non creare un ciclo.

Tabella 4.6: Algoritmo GS (Margaritis, 2003).

tramite l'algoritmo *Grow-Shrink Markov Blanket*, che in una prima fase seleziona ricorsivamente in un insieme $\mathbf{S}_X$ tutti i nodi Y che in quell'iterazione non risultano condizionatamente indipendenti ($Y \not\perp X \mid \mathbf{S}_X$) per poi rimuoverli da $\mathbf{S}_X$ in base al principio inverso. La complessità di questa operazione è lineare ($o(n)$) nel numero di test rispetto al numero di variabili, quindi risulta computazionalmente accettabile anche per grandi numeri di variabili.

La determinazione degli archi a questo punto si riduce all'individuazione dei coniugi in ogni $Bl(X)$, che per definizione sono gli unici nodi al suo interno non adiacenti ad X stesso. Anche se si tratta di una ricerca esaustiva il numero di test ($o(2^{|Bl(X)|})$) è limitato dalla dimensione del *Markov blanket*, che in grafi

sufficientemente sparsi è molto contenuta.

Il successivo orientamento degli archi rilevati nella prima parte dell'algoritmo *Grow-Shrink* è sostanzialmente equivalente a quello implementato nell'*Inductive Causation*, ed è più preciso nell'assegnazione delle direzioni in alcuni casi particolari (come ad esempio i nodi con genitori multipli e tra loro connessi) grazie alla sua formulazione specifica.

Capitolo 5

Informazione reciproca e network bayesiani

I metodi di selezione *score* e *constraint-based* vengono solitamente trattati in letteratura come due classi differenti di algoritmi. Sfruttando l'informazione reciproca, i suoi rapporti con alcuni degli indicatori di indipendenza condizionale più diffusi, e la condizione di Markov che caratterizza i network bayesiani è possibile mostrare che il criterio su cui si fondano queste due classi di metodi in realtà è lo stesso.

Una volta individuato questo criterio nella stessa informazione reciproca, si mostrerà come la stima del suo livello di significatività possa essere allo stesso tempo semplificata e resa più accurata in presenza di numerosità campionarie limitate utilizzando le tecniche definite dalla statistica non parametrica.

5.1 Relazione tra metodi *score* e *contraint-based*

Si supponga di avere due network bayesiani definiti sullo stesso insieme di variabili aleatorie $\mathbf{U} = \{X_1, \ldots, X_n\}$ i cui grafi differiscono per la presenza di un solo arco $X_k \rightarrow X_l$, dando così luogo a due distribuzioni di probabilità distinte

$\mathrm{P}(\mathbf{U})$ e $\mathrm{P}'(\mathbf{U})$, e si supponga di trovarsi sotto le ipotesi descritte nel capitolo precedente.

In entrambi i casi le distribuzioni globali possono essere fattorizzate nelle distribuzioni condizionate locali sfruttando la condizione di Markov:

$$\mathrm{P}(\mathbf{U}) = \prod_{X_i \in \mathbf{U}} \mathrm{P}(X_i \,|\, \Pi_{X_i}) \qquad \mathrm{P}'(\mathbf{U}) = \prod_{X_i \in \mathbf{U}} \mathrm{P}\left(X_i \,|\, \Pi'_{X_i}\right)$$

e possono essere confrontate tramite il log-rapporto di verosimiglianza, dato che rappresentano dei modelli annidati. Sfruttando le scomposizioni delle due densità quest'ultimo può essere riscritto evidenziando le componenti relative ai singoli nodi:

$$G^2 = -2 \log \frac{\mathrm{P}(\mathbf{U})}{\mathrm{P}'(\mathbf{U})} = -2 \log \frac{\prod_{X_i \in \mathbf{U}} \mathrm{P}(X_i \,|\, \Pi_{X_i})}{\prod_{X_i \in \mathbf{U}} \mathrm{P}\left(X_i \,|\, \Pi'_{X_i}\right)}$$

e può essere notevolmente semplificato, dato che l'unica distribuzione locale che differisce nei due casi è quella relativa a X_l:

$$-2 \log \frac{\prod_{X_i \in \mathbf{U}} \mathrm{P}(X_i \,|\, \Pi_{X_i})}{\prod_{X_i \in \mathbf{U}} \mathrm{P}\left(X_i \,|\, \Pi'_{X_i}\right)} = -2 \log \frac{\mathrm{P}(X_l \,|\, \Pi_{X_l})}{\mathrm{P}\left(X_l \,|\, \Pi'_{X_l}\right)} = -2 \log \frac{\mathrm{P}(X_l \,|\, \Pi_{X_l})}{\mathrm{P}(X_l \,|\, \Pi_{X_l} \cup X_k)}$$

Il confronto globale tra due network bayesiani, i cui grafi differiscono per la presenza di un singolo arco, quindi è equivalente ad una verifica di ipotesi locale sull'indipendenza condizionale di X_l e X_k:

$$H_0 : X_l \perp X_k \,|\, \Pi_{X_l} \qquad H_1 : X_l \not\perp X_k \,|\, \Pi_{X_l}$$

dato che l'indicatore utilizzato non dipende in alcun modo dalla struttura del resto del grafo.

Questo risultato può essere dimostrato in modo analogo sfruttando l'additività del G^2, su cui si basano proprietà analoghe dei modelli grafici (Edwards, 2000) e più in generale delle tabelle di contingenza multidimensionali (Bishop (1971) e Goodman (1971)). Infatti interpretando il G^2 come la differenza tra le devianze re-

lative a $\mathrm{P}(\mathbf{U})$ e $\mathrm{P}'(\mathbf{U})$, ed applicando anche in questo caso la fattorizzazione nelle probalità locali, queste si semplificano dando luogo al test locale di indipendenza condizionale definito in precedenza:

$$\begin{aligned} G^2(\mathrm{P}(\mathbf{U}), \mathrm{P}'(\mathbf{U})) &= G^2(\mathrm{P}(\mathbf{U})) - G^2(\mathrm{P}'(\mathbf{U})) \\ &= \sum_{X_i \in \mathbf{U}} G^2(\mathrm{P}(X_i \mid \Pi_{X_i})) - \sum_{X_i \in \mathbf{U}} G^2\left(\mathrm{P}\left(X_i \mid \Pi'_{X_i}\right)\right) \\ &= G^2(\mathrm{P}(X_l \mid \Pi_{X_l})) - G^2(\mathrm{P}(X_l \mid \Pi_{X_l} \cup X_k)) \\ &= G^2(\mathrm{P}(X_l \mid \Pi_{X_l}), \mathrm{P}(X_l \mid \Pi_{X_l} \cup X_k)) \end{aligned}$$

Questa relazione è ulteriormente rafforzata nel caso in cui si utilizzi l'informazione reciproca come indicatore, dato che quest'ultima è direttamente proporzionale al log-rapporto di verosimiglianza (utilizzato nei metodi *score-based*):

$$G^2(X_l, X_k \mid \Pi_{X_l}) = 2n\mathrm{MI}(X_l, X_k \mid \Pi_{X_l})$$

ed è approssimativamente equivalente al X^2 di Pearson (utilizzato comunemente in letteratura per i metodi *constraint-based*):

$$X^2 \simeq \mathrm{MI}(X_l, X_k \mid \Pi_{X_l})$$

In questo caso quindi non solo le due classi di metodi procedono all'individuazione della struttura del grafo tramite una successione di verifiche di ipotesi locali equivalenti, ma utilizzano nella loro stima la stessa quantità e lo stesso criterio di valutazione (la minimizzazione della distanza di Kullback-Leibler tra il modello stimato e quello sottostante ai dati osservati).

Sulla base di questa considerazione è possibile stabilire un'analogia tra i metodi euristici utilizzati nei metodi *score-based*, che esplorano lo spazio dei grafi proprio tramite il cambiamento di singoli archi, ed i metodi *constraint-based*, che si basano su una successione di verifiche di ipotesi locali. Le quantità utilizzate

nella selezione del modello infatti sono le stesse; quindi ad ogni sequenza di modelli esaminati nei metodi *score-based* corrisponde un'analoga sequenza di test di indipendenza condizionale nei metodi *constraint-based*.

La principale differenza risiede nell'interpretazione dei risultati che esse forniscono. Rappresentando l'insieme dei possibili grafi come uno spazio in cui ogni arco è associato ad una dimensione (che può esprimere solamente tre valori, corrispondenti ai possibili stati dell'arco ad essa associato), le strategie di selezione caratteristiche delle due classi di metodi possono essere interpretate come segue:

- nei metodi *score-based* ogni verifica di ipotesi si traduce in un movimento parallelo ad uno degli assi dello spazio; la procedura di selezione quindi si traduce in una sequenza di spostamenti ortogonali che, a partire da un modello iniziale, conducono ad un secondo modello il cui punteggio è più elevato.

- nei metodi *constraint-based* ogni test viene effettuato senza alcun riferimento ad un modello particolare, e le sue conclusioni sono quindi valide su tutto lo spazio dei grafi. Pertanto l'accettazione dell'ipotesi nulla corrispondente al test comporta l'eliminazione di una delle dimensioni dello spazio, il cui valore si assume noto in forza dell'ipotesi di causalità, e la conseguente riduzione progressiva del numero di modelli validi. Il successivo orientamento degli archi rimanenti contribuisce a restringere ulteriormente lo spazio, individuando una specifica classe di equivalenza markoviana.

5.2 Stima dell'informazione reciproca

Alla luce del fatto che in tutti i metodi di selezione esaminati nel capitolo precedente gli effetti degli errori nelle verifiche di ipotesi sono cumulativi, la qualità della stima dell'informazione reciproca riveste un'importanza fondamentale nell'apprendimento della struttura dei network bayesiani.

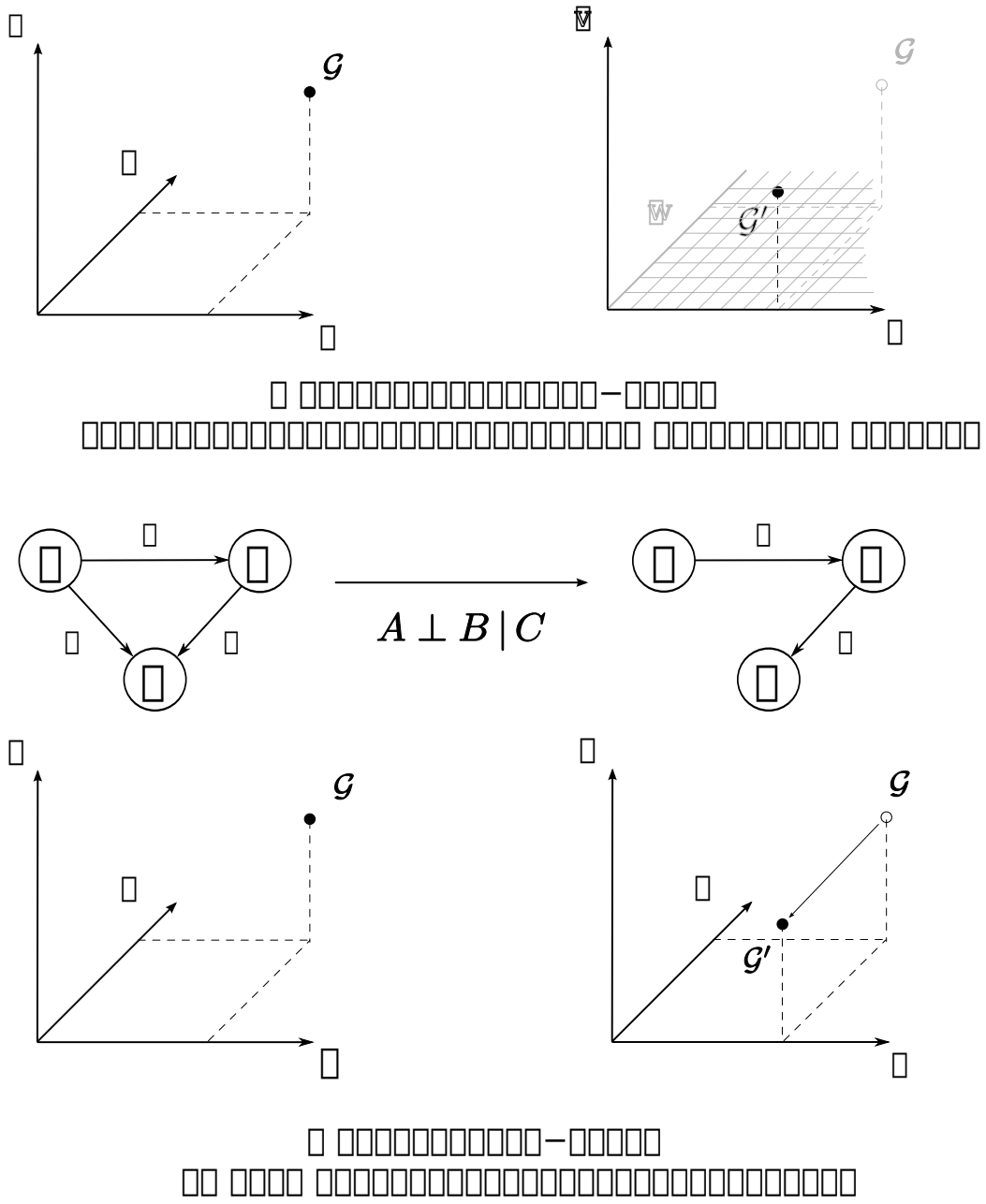


Figura 5.1: Rappresentazione grafica dei criteri di selezione dei metodi *score-based* e *constraint-based*.

L'assunzione della distribuzione asintotica χ^2 per i test di indipendenza condizionale in generale richiede l'assunzione di ipotesi distributive abbastanza forti e la disponibilità di una elevata numerosità campionaria. Quest'ultimo requisito in particolare risulta difficile da soddisfare, soprattutto a causa dalla presenza di un insieme di variabili condizionanti potenzialmente esteso. In queste condizioni infatti garantire la presenza di almeno cinque osservazioni per ogni cella (Agresti, 2002) di tutte le tabelle di contingenza marginali (corrispondenti alle possibili configurazioni delle variabili condizionanti) individuate dalle ipotesi in esame rende difficile l'utilizzo di questi metodi in molte situazioni.

Per queste ragioni risulta conveniente stimare la significatività osservata dell'informazione reciproca tramite la sua distribuzione di permutazione. Quest'ultima infatti non richiede alcuna assunzione distributiva, ed essendo definita condizionatamente ai dati osservati non ha alcuna componente asintotica che ne distorca la valutazione.

L'approccio non parametrico inoltre non richiede il calcolo esplicito della statistica test condizionata, il cui livello di significatività può essere ricavato dalla combinazione dei test associati alle singole tabelle marginali. Infatti l'ipotesi di indipendenza condizionale trattata precedentemente:

$$H_0 : X_l \perp X_k \,|\, \Pi_{X_l} \qquad\qquad H_1 : X_l \not\perp X_k \,|\, \Pi_{X_l}$$

può essere scomposta in un insieme di ipotesi parziali, ognuna delle quali corrisponde proprio ad una verifica di indipendenza non condizionata relativa ad una particolare configurazione dei genitori di X_l:

$$H_0 : \bigcap_{i=1}^{|\Pi_{X_l}|} \{X_{l_i} \perp X_{k_i}\} \qquad\qquad H_1 : \bigcup_{i=1}^{|\Pi_{X_l}|} \{X_{l_i} \not\perp X_{k_i}\}$$

Queste ipotesi, essendo equivalenti ad altrettante ipotesi di uguaglianza in distri-

n	distribuzione di permutazione	distribuzione asintotica
25	0.03995 (−0.01005)	0.13250 (+0.08250)
50	0.04250 (−0.00750)	0.09000 (+0.04000)
100	0.04600 (−0.00400)	0.07105 (+0.02105)
200	0.05157 (+0.00157)	0.06216 (+0.01216)
500	0.05052 (+0.00052)	0.05940 (+0.00940)

Tabella 5.1: Livelli osservati di significatività dell'informazione reciproca in tabelle di contingenza 4x4. I numeri tra parentesi sono gli scarti rispetto al livello teorico, assunto al 5%.

n	distribuzione di permutazione	distribuzione asintotica
200	0.04435 (−0.00565)	0.26036 (+0.21036)
500	0.04866 (−0.00134)	0.14845 (+0.09845)
1000	0.04840 (−0.00160)	0.07201 (+0.02201)
2000	0.05143 (+0.00143)	0.05715 (+0.00715)
5000	0.05057 (+0.00057)	0.05473 (+0.00473)

Tabella 5.2: Livelli osservati di significatività dell'informazione reciproca in tabelle di contingenza 10x10. I numeri tra parentesi sono gli scarti rispetto al livello teorico, assunto al 5%.

buzione, garantiscono la scambiabilità delle osservazioni sotto l'ipotesi nulla:

$$H_0 : \bigcap_{i=1}^{|\Pi_{X_l}|} \left\{ X_{l_i} X_{k_i} \stackrel{d}{=} X_{l_i} \cdot X_{k_i} \right\} \qquad H_1 : \bigcup_{i=1}^{|\Pi_{X_l}|} \left\{ X_{l_i} X_{k_i} \stackrel{d}{\neq} X_{l_i} \cdot X_{k_i} \right\}$$

e possono quindi essere studiate tramite la distribuzione di permutazione.

Uno studio comparativo dei due tipi di distribuzione, condotto simulando delle tabelle di contingenza da varie distribuzioni multinomiali sia sotto l'ipotesi nulla che sotto quella alternativa, conferma quanto detto in precedenza sulla base delle proprietà teoriche dei due approcci.

Il livello di significatività osservato rispetto alla distribuzione χ^2 risulta distorto per numerosità campionarie anche abbastanza elevate; al contrario quello osservato rispetto alla distribuzione di permutazione rimane aderente al livello teorico del

5% in tutte le configurazioni esaminate, dimostrandosi leggermente conservativo in presenza di un basso numero di osservazioni.

La potenza della statistica test, calcolata rispetto ad un livello di significatività osservato (e non teorico) del 5% per garantire un confronto equilibrato, non differisce in maniera sostanziale nei due casi. La distribuzione di permutazione sembra essere leggermente più potente dell'equivalente parametrico, specialmente in presenza di tabelle di contingenza con un numero elevato di celle, ma dato l'ordine di grandezza dello scarto che li separa ed il numero di simulazioni effettuate questo potrebbe essere dovuto alla variabilità della stima.

Questi risultati, anche se ottenuti studiando il comportamento dell'informazione reciproca semplice, possono essere estesi anche a quella condizionata, che costi-

n	distribuzione di permutazione	distribuzione asintotica	distribuzione asintotica corretta	livello corretto
25	0.0820	0.2160	0.0715	0.018
50	0.1969	0.3063	0.2023	0.020
100	0.4865	0.5365	0.4216	0.027
200	0.7533	0.7966	0.7530	0.045
500	0.9834	0.9849	0.9841	0.050

Tabella 5.3: Livelli osservati di potenza dell'informazione reciproca in tabelle di contingenza 4x4. La distribuzione asintotica corretta è tarata in modo da avere un livello di significatività osservato vicino a quello teorico sotto l'ipotesi nulla.

n	distribuzione di permutazione	distribuzione asintotica	distribuzione asintotica corretta	livello corretto
200	0.104	0.431	0.082	0.014
500	0.3556	0.4825	0.3346	0.015
1000	0.7185	0.7742	0.6783	0.025
2000	0.9831	0.9842	0.9835	0.048
5000	0.9996	0.9992	0.9995	0.050

Tabella 5.4: Livelli osservati di potenza dell'informazione reciproca in tabelle di contingenza 10x10. La distribuzione asintotica corretta è tarata in modo da avere un livello di significatività osservato vicino a quello teorico sotto l'ipotesi nulla.

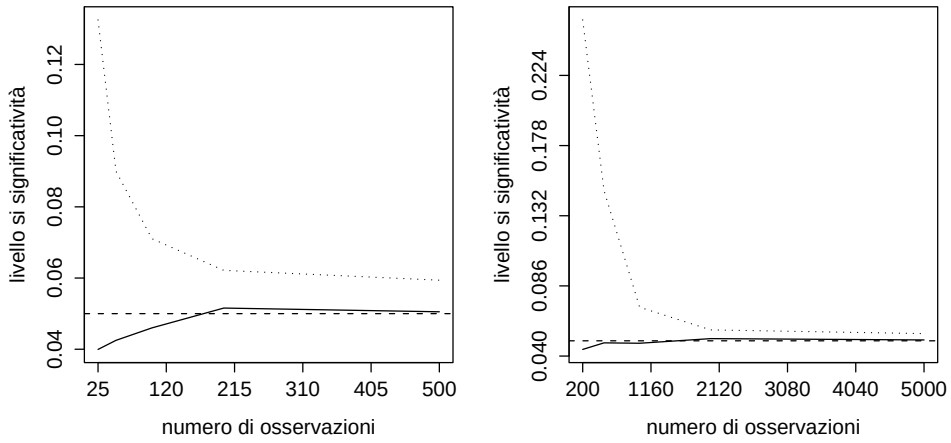


Figura 5.2: Livelli osservati di significatività dell'informazione reciproca in tabelle di contingenza 4x4 (tabella di sinistra) e 10x10 (tabella di destra). La linea punteggiata si riferisce alla distribuzione asintotica χ^2, quella continua alla distribuzione di permutazione.

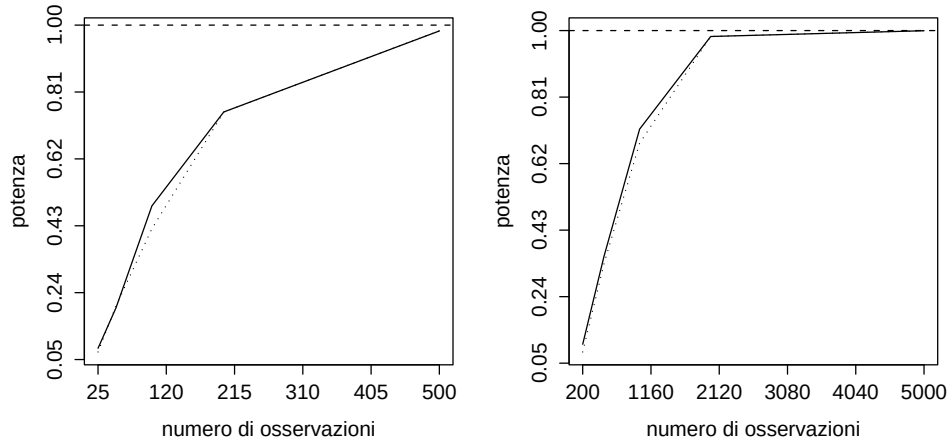


Figura 5.3: Livelli osservati di potenza dell'informazione reciproca in tabelle di contingenza 4x4 (tabella di sinistra) e 10x10 (tabella di destra). La linea punteggiata si riferisce alla distribuzione asintotica χ^2, quella continua alla distribuzione di permutazione.

tuisce la quantità d'interesse nei metodi di apprendimento dei network bayesiani. La combinazione non parametrica infatti è non distorta se lo sono le statistiche test utilizzate nella verifica delle ipotesi parziali (Pesarin, 2001), come accade per la distribuzione di permutazione.

Inoltre dato che queste ultime sono analizzate tutte con l'informazione reciproca, che essendo equivalente a meno di una costante al log-rapporto di verosimiglianza è consistente, anche la loro combinazione risulta consistente.

Appendice A

Implementazione

A.1 Ambiente di lavoro

Tutte le simulazioni effettuate in questa tesi sono state svolta sulla seguente macchina:

Processore:	Dual Athlon MP 1200
RAM:	768 Mb
Sistema Operativo:	Debian GNU/Linux, kernel 2.6.17.8
Compilatore:	gcc version 4.1.2 20061115 (Debian 4.1.1-21)
Libreria C:	2.3.6.ds1-12
R:	2.4.1

I tempi richiesti dalle routine di simulazione in queste condizioni sono di circa uno o due secondi per ogni tabella generata, a seconda del numero di permutazioni utilizzate per calcolare la significatività del test, e la memoria utilizzata è inferiore a 25 Mbyte.

A.2 Calcolo dell'informazione reciproca

```
mi = function(table, gsquare=TRUE) {

  n = sum(table)
  px = rowSums(table)/n
  py = colSums(table)/n

  if (gsquare) {

    2*sum(table * (log(table/n) - log(outer(px, py))), na.rm=TRUE)

  }#THEN
  else {

    sum(table/n * (log(table/n) - log(outer(px, py))), na.rm=TRUE)

  }#ELSE

}#MI
```

A.3 Simulazione del livello di significatività

```
# CONTINGENCY.SIM: simulazione condizionata alle marginali di
#   tabelle di contingenza a 2 entrate di dimensione arbitraria.
#
# nrow, ncol:   numero di righe e di colonne
# num:          numerosita' campionaria
# ntab:         numero di tabelle da estrarre per ogni distribuzione
# ndist:        numero di distribuzioni da simulare
# B:            numero di elementi della distribuzione di permutazione
# FUN:          statistica test

contingency.sim = function(nrow, ncol, num, ntab, ndist, B, FUN) {

  lres = vector("list", ndist)

  for (i in 1:ndist) {

    repeat {

      nx = runif(nrow,1,10)
```

```
      ny = runif(ncol,1,10)
      px = nx/sum(nx)
      py = ny/sum(ny)

      if (identical(sum(px*num), sum(py*num))) break

    }#REPEAT

    res = matrix(rep(0, 2*ntab), ntab, 2)

    for (j in 1:ntab) {

      m = matrix(rmultinom(1, num, outer(px, py)), nrow, ncol)
      s = FUN(m)
      l = sapply(unique(r2dtable(B, px * num, py * num)), FUN)
      res[j,1] = length(which(l > s))/B
      res[j,2] = pchisq(s, (nrow-1)*(ncol-1), lower.tail=FALSE)

    }#FOR

    lres[[i]] = res

  }#FOR

  lres

}#CONTINGENCY.SIM

# CONTINGENCY.ALPHA: calcolo del livello medio osservato.
#
# result:      risultato di contingency.{sim,dep}
# alpha:       livello di significativit\'a

contingency.alpha = function(result, alpha) {

  tab = t(sapply(res,
    function(x) {
      c(length(which(x[,1] < alpha))/length(x[,1]),
        length(which(x[,2] < alpha))/length(x[,2]))
    } ))
  list(table = tab, observed = colMeans(tab, na.rm=TRUE),
    summary = summary(tab))

}#CONTINGECNY.ALPHA
```

A.4 Simulazione del livello di potenza

```
# CONTINGENCY.DEP: simulazione di tabelle di contingenza a due
#   entrate con variabili dipendenti.
#
# (per i parametri si veda contingency.sim)

contingency.dep = function(nrow, ncol, num, ntab, ndist, offset, B, FUN) {

  lres = vector("list", ndist)

  for (i in 1:ndist) {

    repeat {

      p = matrix(runif(nrow*ncol,1,2), nrow, ncol)
      u = upperTriangle(p)
      upperTriangle(p) = runif(length(u), 0,1) + offset
      p = p/sum(p)

      if (identical(sum(rowSums(p * num)), sum(colSums(p * num)))) break;

    }#REPEAT

    res = matrix(rep(0, 2*ntab), ntab, 2)

    for (j in 1:ntab) {

      m = matrix(rmultinom(1, num, p), nrow, ncol)
      s = FUN(m)
      l = sapply(unique(r2dtable(B, rowSums(p * num), colSums(p * num))), FUN)
      res[j,1] = length(which(l > s))/B
      res[j,2] = pchisq(s*(2*num), (nrow-1)*(ncol-1), lower.tail=FALSE)

    }#FOR

    lres[[i]] = res

  }#FOR

  lres

}#CONTINGENCY.DEP
```

Appendice B

Problemi aperti

B.1 Trattamento di dati continui

I metodi di apprendimento della struttura dei network bayesiani, specialmente quelli *constraint-based*, in genere sono sviluppati nell'ambito multinomiale per motivi sia teorici che computazionali. Le variabili continue devono quindi essere discretizzate per poter essere utilizzate, con una conseguente perdita di informazione.

Una possibile soluzione è l'estensione di questi algoritmi al caso continuo, ad esempio utilizzando la definizione generale di famiglia esponenziale (di cui la multinomiale fa parte). Come indicatore per le verifiche di ipotesi è possibile sfruttare l'*entropia differenziale* (Cover and Thomas, 2006)

$$\mathrm{H}(X) = -\int f(x) \log f(x)\, dx$$

e l'*informazione reciproca* corrispondente:

$$\mathrm{MI}(X, Y) = \int f(x, y) \log \frac{f(x, y)}{f(x) f(y)}\, dx\, dy$$

mantenendo come criterio di ottimizzazione la minimizzazione della distanza di

Kullback-Leibler tra il modello stimato e quello sottostante ai dati.

In alternativa risulta di interesse lo studio dei criteri di discretizzazione (Margaritis, 2005), per valutare quali tra di essi producano delle categorizzazioni più fedeli in termini di entropia, riducendo così al minimo la distorsione dei dati originali.

B.2 Orientamento degli archi

I metodi *score-based* prevedono, tra i possibili criteri di esplorazione dello spazio dei grafi, l'inversione dell'orientamento di uno degli archi. La verifica di un'ipotesi di questo tipo, che ha la forma:

$$H_0 : (X, Y) \in A \qquad\qquad H_1 : (Y, X) \in A$$

non può essere portata a termine tramite gli usuali indicatori, dato che i due modelli a confronto non sono tra loro annidati. Anche se è possibile stabilire una preferenza tra le due ipotesi effettuando due verifiche separate:

$$\begin{aligned} H_{0_f} &: (X, Y) \notin A & H_{1_f} &: (X, Y) \in A \\ H_{0_b} &: (Y, X) \notin A & H_{1_b} &: (Y, X) \in A \end{aligned}$$

e stabilendo la direzione dell'arco in questione sulla base del livello di significatività osservata di queste ultime, risulta comunque d'interesse l'individuazione di un test statistico unico per le ipotesi originali.

Un ulteriore problema è l'inquadramento di un simile test nella relazione tra metodi *score* e *constraint-based* esposta in questa tesi, dato che questi ultimi stabiliscono l'orientamento degli archi in modo completamente diverso ed in funzione delle classi di equivalenza markoviane.

B.3 Proprietà della combinazione non parametrica

Molte delle proprietà della combinazione parametrica di più test discendono direttamente da quelle dei test stessi. Altre vanno invece studiate in modo specifico; ad esempio l'utilizzo di indicatori uniformemente più potenti non implica che la loro combinazione sia anch'essa uniformemente più potente.

Per questo motivo può essere utile condurre uno studio specifico sulla combinazione non parametrica dell'informazione reciproca, sia tramite la simulazione di tabelle di contingenza marginali che tramite l'analisi formale delle sue caratteristiche.

Bibliografia

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., 2002.

A. Azzalini. *Inferenza Statistica*. Springer, 2001.

Y. M. M. Bishop. Effects of collapsing multidimensional contingency tables. *Biometrics*, 27:545–562, 1971.

R. R. Bouckaert. *Bayesian Belief Networks: from Construction to Inference*. PhD thesis, University of Utrecht, 1995.

J. M. Puerta Callejon. *Métodos Locales y Distribuitos para la Constuccíon de Redes de Creencia Estáticas y Dinamicas*. PhD thesis, E.T.S. de Ingeniería Informática, Granada, 2001.

D. M. Chickering. A transformational characterization of equivalent bayesian network structures. In *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*, pages 87–98. Morgan Kaufmann Publishers Inc., 1995.

G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

T. A. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

M. H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, Inc., 2004.

D. I. Edwards. *Introduction to graphical modelling*. Springer, 2000.

L. A. Goodman. The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistics Association*, 66: 339–344, 1971.

David Heckerman. A tutorial on learning with bayesian networks. pages 301–354, 1999.

F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.

K. Korb and A. Nicholson. *Bayesian artificial intelligence*. Chapman and Hall, 2004.

S. Kullback. *Information theory and statistics*. John Wiley & Sons, Inc., 1959.

D. Margaritis. Distribution-free learning of bayesian network structure in continuous domains. pages 825–830. Proceedings of The Twentieth National Conference on Artificial Intelligence, 2005.

D. Margaritis. *Learning Bayesian Network Model Structure from Data*. PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, May 2003. Available as Technical Report CMU-CS-03-153.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 2nd edition, 1989.

C. Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 411–41. Morgan Kaufmann, 1995.

J. Pearl. *Heuristics: intelligent search strategies for computer problem solving*. Addison-Wesley, 1984.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

F. Pesarin. *Multivariate Permutation Tests with Applications in Biostatistics*. John Wiley & Sons, Inc., 2001.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, 2001.

I. Tsamardinos, C. F. Aliferis, and A. Statnikov. Algorithms for large scale markov blanket discovery. In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, pages 376–381. AAAI Press, 2003.

T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. *Uncertainty in Artificial Intelligence*, 6:255–268, 1991.

Finito di stampare il 1 marzo 2007 utilizzando LaTeX 2ε