Università degli Studi di Milano-Bicocca

Dipartimento di Statistica
Corso di Dottorato di Ricerca in Statistica-XXII Ciclo

Tesi di dottorato

Modelli ad Equazioni Strutturali e Reti Probabilistiche Bayesiane

Due approcci a confronto nello studio delle relazioni causali

Dottoranda
Rosa Falotico

Tutore di tesi
Prof.ssa Pier Alda Ferrari

Cotutore di tesi
Prof. Vincenzo Esposito Vinzi

Anno Accademico 2009–2010

Alla mia Stella,
alla mia buona Stellina Polare
che resta tale
nonostante la precessione
degli equinozi.
Grazie, sei sempre nel mio cuore.

Ringraziamenti

Come spesso si dice “Alea iacta est”.

E se questo lavoro è arrivato a conclusione devo ringraziare principalmente la Professoressa Pier Alda Ferrari, che mi ha fatto intravedere questa meravigliosa opportunità che è stato il dottorato in Statistica ed il Professor Esposito Vinzi, il quale mi ha fatto apprezzare enormemente il meraviglioso mondo dell'analisi multivariata e al quale devo tutta la mia riconoscenza per il supporto che mi ha dato nel portare a termine questo lavoro.

Tra tutte le persone verso cui ho un forte debito di riconoscenza sicuramente spiccano Lorenzo e il Professor Piero Quatto, che hanno creduto nelle mie capacità molto più di quanto non ci abbia creduto io stessa. Non so se ce l'avrei fatta senza di voi. Grazie.

Non potrò mai dimenticare le “mie allegre comari di dottorato”, Viviana e Isa, con le quali ho trascorso splendidi momenti di puro divertimento e con cui ho condiviso attimi di terrore e paranoia apocalittici, specie negli ultimi mesi. Spero di ripensarci un giorno, con il sorriso sulle labbra. Grazie ragazze. E grazie a tutti gli amici che sono passati per quell'ufficio a donare anche solo un attimo di gioia.

Grazie a Pia, il cui saggio consiglio mi ha guidata per le vie tortuose di questo impervio cammino.

Grazie alla Professoressa Fulvia Mecatti per la fiducia accordatami durante questi anni di collaborazione didattica.

Grazie i tecnici di laboratorio Riccardo ed Andrea, ai quali mi sono agganciata come una sanguisuga nei peggiori momenti di panico.

Grazie alla mia famiglia per il fatto di esistere.

Grazie a tutti. Rosa

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

I concetti di causalità e di probabilità hanno rappresentato, fin dai primi tentativi di formalizzazione, una sfida intellettuale ardita.

Per quanto riguarda la causalità, l'impulso primario all'indagine scientifica risiede nel tentativo di stabilire relazioni di causa-effetto fra fenomeni, che siano di natura fisica, sociale o psicologica. Purtroppo definire la natura di tali relazioni non è semplice, ne' dal punto di vista teorico, ne' tanto meno da quello applicativo. Nonostante sia un obiettivo comune a tutte le scienze, la formalizzazione del concetto di causalità non ha raggiunto ancora una formulazione unitaria. Inoltre rilevare empiricamente la presenza di causalità fra fenomeni, stabilire le concause, l'entità dei legami causali sono operazioni altrettanto controverse.

Evidenziare e valutare relazioni causali per un ristretto numero di fenomeni pone diverse difficoltà, ma diventa estremamente complicato quando ad agire sono concause multiple su numerosi effetti. In questo caso la statistica viene in aiuto con le potenti tecniche di analisi dei dati, che utilizzano uno strumento concettuale non meno controverso della causalità, ovvero la teoria probabilistica.

Le interpretazioni della probabilità sono state tanto divergenti tra loro da dare origine a due distinte scuole di pensiero:

- La statistica frequentista che fa riferimento alla ripetibilità degli esperimenti in condizioni uniformi.

- La statistica bayesiana che esplicita la componente soggettiva dell'assegnazione delle probabilità.

Il proposito del presente lavoro è quello di descrivere e confrontare alcuni modelli statistici comunemente impiegati per l'analisi delle relazioni causali tra fenomeni multidimensionali (appartenenti rispettivamente alla scuola frequentista e alla scuola bayesiana) allo scopo di proporre una nuova opportunità di integrazione per affrontare il problema della ricerca e della validazione delle relazioni causali esistenti fra fenomeni.

Nel primo capitolo del lavoro vengono introdotte le principali correnti interpretative dei concetti di probabilità e di causalità.

Nel secondo capitolo viene presentata, nell'ambito della statistica frequentista, una delle più diffuse medodologie di modellazione di relazioni causali: i modelli ad equazioni strutturali (SEM). L'esposizione viene completata con la descrizione di alcuni importanti metodi di stima dei parametri: il metodo SEM-ML o LISREL in rappresentanza dei metodi *covariance based* ed il metodo PLS-PM in rappresentanza dei metodi *component based*.

Il terzo capitolo tratta dell'approccio soggettivista ed è quindi dedicato ai modelli di reti probabilistiche bayesiane. In particolare viene presentata la logica di tali modelli ed alcuni algoritmi di stima dei parametri.

Lo scopo del quarto capitolo è quello di applicare due dei metodi presentati, il modello SEM con algoritmo di stima PLS-PM e le reti probabilistiche bayesiane con algoritmo di stima *Tabu order*, ad un caso applicativo tipico dell'analisi dei dati. Verrà presentato lo studio della soddisfazione e della fedeltà della clientela, ponendo particolare attenzione ai fattori che le determinano, quali aspettative, qualità e valore percepiti.

Alla fine del lavoro vengono presentate le conclusioni e quelle che si ritiene siano interessanti prospettive future della ricerca.

Capitolo 1

Causalità e Probabilità: i concetti base

Come già anticipato, di seguito si farà riferimento ai concetti di probabilità e di causalità, nonché alle loro più note interpretazioni.

La probabilità e la causalità sono due concetti strettamente legati. L'atto di assegnare una probabilità e la ricerca di relazioni di causa-effetto sono il mezzo che l'essere umano utilizza nel tentativo di imbrigliare l'incertezza e dominare la natura.

Attribuire un valore di probabilità ha lo scopo di rendere meno ignoto il futuro, stimando il grado di "certezza" degli eventi che devono ancora verificarsi. Il criterio in base al quale attribuire questo valore resta, per molti secoli, arbitrario. Risalgono all'epoca Rinascimentale le prime testimonianze di criteri razionali di attribuzione che danno carattere di scientificità alla disciplina probabilistica. Il paragrafo 1.1 contiene alcuni cenni storici e le interpretazioni di probabilità oggi maggiormente accreditate.

La causalità è il fulcro della ricerca filosofica occidentale fin dalle origini e, a tutt'oggi, costituisce il fondamento della conoscenza. La scoperta dell'esistenza di una relazione causale permette una migliore comprensione dei fenomeni e, ove possibile, ne consente la manipolazione per fini utilitaristici. Ma mentre è indubbia l'importanza della causalità, non altrettanto chiara è la sua definizione formale e a maggior ragione le metodologie da seguire per la sua rilevazione empirica: è difficile in una relazione causale, distinguere la causa dall'effetto ed epurare l'analisi dall'azione delle concause o della causa comune a fenomeni solo in apparenza in relazione. Nel paragrafo 1.2 vengono esposte le più note nozioni di causalità, gli schemi fondamentali di rappresentazione e le principali metodologie di ricerca empirica.

1.1 La Probabilità

La necessità di formalizzare la probabilità nasce dal desiderio di attribuire una misura quantitativa alla possibilità che si verifichino eventi incerti. Questa esigenza ha trovato soluzioni qualitative piuttosto arbitrarie nel corso dei secoli. I primi tentativi di un approccio quantitativo vengono attribuiti ai matematici della scuola italiana del '400. In particolare Luca Pacioli e Niccolò Tartaglia iniziano ad occuparsi della ricerca del modo più equo per ripartire la posta rimasta in palio in un gioco d'azzardo interrotto. Il problema, che all'apparenza non sembra presentare grosse difficoltà, richiede un'impostazione probabilistica per essere risolto correttamente.

Durante il secolo successivo, Girolamo Cardano scrive il *Liber de ludo aleæ* (terminato nel 1526 ma dato alle stampe solo nel 1663) e Galileo Galilei il libro *Sulla scoperta dei dadi* (pubblicato nel 1656) che segnano gli albori della disciplina.

Il fronte più avanzato della ricerca si sposta in Francia, dove la forte diffusione dei giochi d'azzardo crea, nei secoli XVI e XVII, i presupposti per una nozione di probabilità facente riferimento ad eventi ripetibili secondo schemi ben definiti (proprio come quelli dei giochi di carte o dei dadi) in cui si conoscono tutti i possibili esiti elementari, ma non è noto quale sarà l'unico esito a verificarsi realmente. Assunti tutti gli esiti ugualmente possibili, la probabilità di un evento (ovvero di un insieme di esiti elementari) viene definita come il rapporto del numero dei singoli esiti favorevoli all'evento considerato, sul numero totale di esiti possibili.

Questa visione della probabilità, comune ai giocatori razionali, diventa una elaborazione formale molto raffinata nei lavori di Pascal, di Fermat e di Huygens e nel XVIII secolo, grazie a Laplace, assume un ruolo cardine nella disciplina probabilistica. L'appellativo di "classica" per questa interpretazione deriva sia dalla primogenitura della formalizzazione, sia dal fatto che rappresenta il canone di riferimento fino al XIX secolo, quando viene sostituita da un approccio in grado di consentire una più vasta applicabilità ai casi concreti: l'interpretazione *frequantista*.

Il nuovo approccio, più naturale e allo stesso tempo più generale, si fonda sull'elevata ripetibilità delle osservazioni in condizioni uniformi e non richiede una conoscenza a priori dell'insieme degli esiti possibili. Tale nozione di probabilità viene indicata con il termine *chance* ed interpretata come frequenza relativa di lungo periodo dell'evento, ovvero il rapporto del numero di volte in cui l'evento si presenta sul numero di osservazioni effettuate, quando la rilevazione viene ripetuta sufficientemente a lungo, da cui l'appellativo di *frequentista* per questa interpretazione.

Il periodo di maggior sviluppo della teoria frequentista è intorno al XIX secolo ad opera di Venn, Galton e Von Mises, i quali ne forniscono un'intelaiatura

teorica che permette di soppiantare l'interpretazione classica, risolvendo molti dei problemi lasciati aperti da quest'ultima.

Anche l'approccio frequentista presenta dei punti critici che offrono lo spunto per una diversa interpretazione della probabilità: l'approccio *bayesiano*.

Secondo quest'ottica nuova le probabilità non sono più legate al solo evento a cui vengono attribuite ma dipendono anche dal decisore. L'applicabilità della teoria probabilistica viene estesa al caso di fenomeni rari o unici per i quali il decisore, non potendo contare su precedenti osservazioni per assegnare le probabilità, fa riferimento al *grado di fiducia* che ha relativamente al verificarsi di un evento. Tale grado di fiducia deriva principalmente da informazioni generiche e di contesto, dall'esperienza e dal metro di giudizio del *soggetto* esprimente le probabilità, da cui il nome di interpretazione *soggettivista*.

L'estensione a casi in cui siano applicabili la concezione *classica* o quella *frequentista* non pone grandi difficoltà teoriche. Sia le informazioni relative ai possibili esiti dell'esperimento, sia quelle relative alla rilevazione, nel lungo periodo ed in condizioni uniformi, dello stesso esperimento ripetuto, possono essere ricomprese nel "bagaglio culturale" del decisore.

In fondo, quello soggettivista è il più antico e comune metodo usato dall'umanità per restringere l'incertezza relativa agli eventi futuri, l'unica novità risiede nel fatto che nei primi anni del '900, soprattutto ad opera di Ramsey, de Finetti e Savage, la concezione bayesiana viene inquadrata in una struttura matematica che le fornisce validità scientifica e ne fa uno strumento fondamentale negli ambiti di studio in cui le pesanti ipotesi richieste dalle altre interpretazioni non sono verificate.

Nel XX secolo vengono proposte altre definizioni di probabilità. Sia Keynes che Jeffreys considerano la probabilità come *una relazione oggettiva fra proposizioni*, basata sulle conoscenze di carattere puramente logico ed oggettivo del decisore, mentre Popper (1934) introduce il concetto di *propensione* che si ricollega alla definizione frequentista, ma vuole esserne un'estensione valida anche per le osservazioni non ripetibili.

Sebbene sia stata tentata più volte una riunificazione delle più accreditate interpretazioni della probabilità, non si è mai giunti ad una ricomposizione organica dei vari approcci. Solo nei primi anni del Novecento, Kolmogorov, introducendo la teoria assiomatica della probabilità, ha fornito quanto meno una base matematica comune a tutta la disciplina.

Dopo questi brevi cenni storici, vengono presentate, in maniera più approfondita, quelle che sono state le tre più influenti teorie relative al concetto di probabilità. La teoria classica viene introdotta per completezza e per la sua innovatività, mentre saranno le interpretazioni bayesiana e frequentista quelle alle quali si farà riferimento nel seguito del presente lavoro.

1.1.1 La probabilità classica: il gioco d'azzardo diventa scienza

Si può affermare che la concezione classica della probabilità nasca, ed in un certo senso muoia, sul tavolo verde.

L'origine della formula di calcolo: "numero di casi favorevoli su numero di casi possibili" non è attribuibile a nessun pensatore in particolare, ma sicuramente fa parte del patrimonio culturale comune nel XVI secolo, soprattutto nell'ambiente dei giocatori d'azzardo. Probabilmente questa interpretazione risale ad epoche ben più remote, seppure non venga formalizzata in precedenza, per cui la sua estensione a casi al di fuori del gioco d'azzardo non pone particolari problemi. L'unica difficoltà pratica che tale approccio presenta riguarda i criteri di enumerazione di tutti gli esiti elementari possibili.

Il caso classico del lancio di due monete regolari, anche nella sua semplicità, presenta dei punti di conflitto. Ad una prima analisi, seguendo l'**ipotesi 1** che assume che l'ordine non sia influente, i possibili esiti dell'esperimento sono tre: 2 testa, 2 croce, 1 testa - 1 croce. Se però consideriamo l'**ipotesi 2**, secondo la quale l'ordine deve essere considerato al momento nel calcolo, i possibili esiti sono quattro: 2 testa, 2 croce, 1 testa - 1 croce, 1 croce - 1 testa. Accettando la prima ipotesi, le probabilità di ottenere 2 volte testa è pari a $\frac{1}{3}$, ovvero 0.33, mentre seguendo la seconda ipotesi la stessa probabilità è pari a $\frac{1}{4}$, ovvero 0.25.

Proprio un problema di questo tipo dà avvio alla feconda corrispondenza fra Pascal e Fermat, occasione per il primo tentativo di formalizzazione della probabilità classica. La discussione nasce su istanza del Cavalier de Méré, accanito giocatore d'azzardo, dedicatosi al calcolo delle probabilità per ovvie ragioni. Convinto della correttezza dell'**ipotesi 1** che trascura l'ordine degli eventi elementari, de Méré esegue dei calcoli probabilistici che però non trovano riscontro nei risultati ottenuti giocando d'azzardo. Per risolvere il dilemma, il cavaliere decide di consultare Pascal, il quale a sua volta intraprende uno scambio epistolare con Fermat. Dal confronto di questi due grandi intelletti nascono i concetti fondamentali della teoria classica delle probabilità, che trova compimento nella definizione di de Moivre (XVI secolo) e Laplace (XVIII secolo). In maniera formale:

Definizione 1.1 (Probabilità "Classica").
Si consideri un esperimento $\mathcal{E}$ caratterizzato da un numero finito $m < \infty$ di esiti elementari, tutti aventi la stessa possibilità *di realizzazione, e di cui un numero $n < m$ è favorevole all'evento E associato all'esperimento $\mathcal{E}$. Si definisce come probabilità di E, $P(E) = \frac{n}{m}$, il numero di eventi elementari favorevoli ad E sul numero di tutti gli eventi elementari possibili.* △

La funzione così definita possiede quelle che sono considerate proprietà fondamentali per una probabilità:

Proprietà 1.1.
a) $0 \leq P(E) \leq 1$
b) $P(\Omega) = 1$
c) $P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P\left(E_i\right)$ se E_i sono eventi a due a due incompatibili. △

Ω è l'insieme ti tutti i possibili esiti elementari ed E_i è un evento qualsiasi, sottoinsieme di Ω.

L'interpretazione classica della probabilità rappresenta senza dubbio una svolta nello sviluppo della disciplina, ma non è certo esente da critiche. Nello specifico, i problemi che pone sono di duplice natura, teorica e applicativa.

Si riferisce proprio a quest'ultimo aspetto l'affermazione iniziale del paragrafo: al di fuori dell'ambito dei giochi d'azzardo difficilmente si verificano le forti assunzioni legate all'interpretazione classica della probabilità. Sebbene le controversie relative al calcolo teorico siano state risolte con la costituzione di una struttura matematica molto solida, il campo di applicazione dei risultati resta molto ristretto. Non è realistico pensare di riuscire ad ottenere una conoscenza tanto approfondita di fenomeni economici o sociali, solo per fare un esempio, tale da consentire di definire a priori il numero di tutti i possibili esiti di un esperimento. Inoltre, se questo numero non ha cardinalità finita, l'interpretazione classica non è neppure applicabile.

Le criticità teoriche quali la richiesta di equiprobabilità degli esiti elementari sono ancora più gravi. É chiaro che chiamare in causa il concetto di *uguali possibilità* di realizzazione di due esiti distinti, nella stessa definizione di probabilità, si riduce ad una mera tautologia: come si definisce l'uguale possibilità senza fare riferimento alla probabilità?

Dati i problemi lasciati irrisolti dall'interpretazione classica, nasce l'esigenza di un nuovo approccio, in grado di consolidare le fondamenta teoriche della disciplina e di ampliarne il campo applicativo.

La via più naturale da percorrere è quella di far riferimento a concetti che si sono mostrati già molto validi, quali la replicabilità dell'esperimento in condizioni uniformi e l'utilizzo dei risultati ottenuti per quantificare le probabilità.

1.1.2 La probabilità frequentista: la forza dell'esperimento

Come già accennato parlando dei problemi del Cavalier de Méré, la pratica del gioco d'azzardo, fin dall'inizio, fa da riscontro e da termine di paragone ai risultati della teoria classica delle probabilità. In effetti de Méré comincia a porsi il problema del metodo di enumerazione di tutti i possibili esiti elementari constatando che i calcoli teorici da lui effettuati sono in contrasto con il risultato della pratica

del gioco d'azzardo, il quale altro non è se non la replica incessante di uno stesso esperimento sotto ipotesi di condizioni esterne costanti (in caso di gioco non truccato).

Se all'inizio è solo un termine di paragone, successivamente la ripetibilità stessa diventa parte integrante della definizione di probabilità. Il primato dell'esperimento e dell'osservazione sulla teoria trova realizzazione nell'approccio *frequentista* alla probabilità. Le assunzioni su cui esso si basa sono più "realistiche" dell'interpretazione classica, vengono richiesti:

- La replicabilità dell'esperimento in condizione di uniformità, restando inalterati i fattori esterni.

- La tendenza di lungo periodo alla stabilità sotto l'azione del caso.

Definizione 1.2 (Probabilità "Frequentista").
Sia n il numero di volte in cui l'esperimento $\mathcal{E}$ viene ripetuto, mantenendo inalterate le condizioni esterne, e siano i due esiti E e $\bar{E}$, mutuamente esclusivi. Quando gli esiti dell'esperimento si alternano in modo casuale (del tutto imprevedibile), e contemporaneamente le frequenze relative $f_{n,E}$ (numero di casi in cui si è verificato E su numero di esperimenti effettuati) dopo le oscillazioni iniziali, tendono a stabilizzarsi intorno ad un valore costante p_E, allora è possibile identificare con p_E la probabilità $P(E)$ dell'evento. △

Tale definizione di probabilità conserva le proprietà fondamentali 1.1 ed al tempo stesso consente nuove applicazioni: la condizione di equiprobabilità degli eventi elementari non è più necessaria ed inoltre è possibile attribuire una probabilità a qualsiasi evento legato ad un esperimento ripetibile un numero indefinito di volte (anche solo in linea teorica).

Da questa che a prima vista appare una potenzialità derivano anche le principali critiche: quanto deve essere grande n perché nella pratica si possa dire che $f_{n,E}$ approssimi sufficientemente p_E? I criteri d'arresto non forniscono certezze teoriche, consentendo ai detrattori dell'approccio frequentista di affermare che in fondo esso si basa principalmente su euristiche assurte al grado di definizioni.

Per quanto riguarda gli ambiti di applicazione, essi sono notevolmente maggiori rispetto all'impostazione classica, ma restano comunque limitati al campo sperimentale, dove le assunzioni di replicabilità ed uniformità sono più facilmente verificate. Quando la condizione di uniformità non sussiste, i risultati vengono fortemente condizionati da fattori esterni per cui il valore $f_{n,E}$ può risultare sostanzialmente diverso da p_E. Al contrario, per le discipline in cui non è assicurata la replicabilità, come in campo economico, sociale e comportamentale, nelle quali gli eventi considerati sono per lo più rari o unici, l'approccio frequentista nega la possibilità di trattamento scientifico.

Qualora si voglia effettuare ugualmente l'analisi, è necessario far ricorso a forti assunzioni, spesso non del tutto realistiche.

Allo scopo di ovviare a queste difficoltà nasce un nuovo approccio, il quale generalizza ulteriormente il concetto di probabilità: l'interpretazione bayesiana.

1.1.3 La probabilità bayesiana: la rivoluzione antropocentrica

L'interpretazione inferenziale della probabilità inizia ad essere messa in discussione per rispondere all'esigenza di trattare eventi rari o non ripetibili. L'ottica da laboratorio, indispensabile nelle interpretazioni oggettiviste, viene abbandonata a vantaggio di una totale rifondazione della disciplina. Viene proposto un approccio primitivo e rivoluzionario allo stesso tempo, che intende ripartire dalle origini, ovvero dall'analisi del metodo di attribuzione delle probabilità da parte del decisore.

Lo scopo primario della formalizzazione della probabilità è l'esigenza di eliminare l'arbitrarietà nelle scelte del soggetto. Gli approcci classico e frequentista fanno ricorso all'*oggettività* della valutazione ed assumono la probabilità come un attributo intrinseco dei fenomeni: ad ogni evento è associata una ed una sola funzione di probabilità, costante per ogni decisore. La teoria bayesiana fa ricorso ad una concezione più ampia, non legata al solo evento in sè ma al soggetto assegnante la funzione di probabilità.

L'assunzione fondamentale dietro l'approccio bayesiano è che la probabilità dipende imprescindibilmente dal soggetto. La valutazione probabilistica diventa un'azione individuale, da cui l'appellativo di *soggettivista* per tale interpretazione. In pratica l'assegnazione di probabilità bayesiana viene effettuata dal decisore, in base alle proprie conoscenze a priori di tipo statistico integrate con tutto il generico bagaglio culturale personale, con l'obbligo di seguire criteri razionali, i quali non vincolano la probabilità P(E) assegnata ad un evento E, ad essere obbligatoriamente la stessa per tutti i soggetti, anzi prevede che possa variare a seconda delle informazioni a disposizione, dell'esperienza personale e soprattutto del punto di vista proprio di ogni decisore: la probabilità è assimilabile al grado personale di confidenza del verificarsi dell'evento, basata anche su valutazione quantitativa (sperimentazione, raccolta e analisi dati) ma necessariamente ricomposta in una più generale valutazione qualitativa, sotto vincolo di razionalità, in altri termini "la misura del grado di plausibilità che un individuo assegna ad un evento incerto" (J.Bernoulli, XVIII secolo)

Un punto essenziale della teoria bayesiana è che essa si presenta come una generalizzazione delle interpretazioni precedenti. Se infatti si considerano le assunzioni retrostanti la teoria classica e frequentista come delle conoscenze a priori esplicitate, è possibile ricondurre queste due teorie a casi particolari dell'impostazione soggettivista.

L'impostazione bayesiana originaria, sviluppata da Ramsey e de Finetti, riconduce l'assegnazione di probabilità allo *scommettere* sul verificarsi dell'evento. Successivamente sono state proposte diverse reinterpretazioni. Savage, Kranz e Luce propongono una versione detta *personalistica* della probabilità che fa riferimento alla teoria dell'utilità. Invece Aucombe e Auman suggeriscono di ricondurre l'impostazione bayesiana alla più generale disciplina della *teoria delle decisioni*.

Rimandando al capitolo 3 per l'esposizione formale della teoria bayesiana, di seguito viene presentato brevemente l'approccio soggettivista nell'impostazione della scommessa.

Le assunzioni fondamentali su cui si basa l'approccio definettiano riguardano la razionalità e la coerenza del decisore, al quale è fatto esplicito divieto di effettuare scommesse a perdita o guadagno certo. Il decisore, proponendo la scommessa, deve essere disposto a scambiare il posto dello scommettitore con quello del banco.

Definizione 1.3 (Probabilità "Soggettiva"(*Impostazione definettiana*)). *Dato un evento E, si dice probabilità di E il valore $P(E) = p_E$, se un individuo è disposto a pagare (ricevere) una somma pari a $p_E \times S$ per riscuotere (pagare) un importo monetario pari a S se E si verifica e un importo pari a 0 se E non si verifica.* △

Le parentesi sottolineano proprio il fatto che la valutazione di probabilità non può variare se l'agente cambia ruolo. La probabilità diventa la quota di scommessa su un evento, pari al rapporto

$$\frac{p_E \times S}{S} = \frac{somma\ puntata}{somma\ vinta\ se\ E\ vero}.$$

Il metodo della scommessa, oltre che una definizione, fornisce uno strumento di misura della probabilità, un mezzo operativo di facile comprensione ed utilizzo. Non è neppure richiesto che la scommessa venga realmente posta in essere; perché la definizione sia valida è sufficiente che sia razionale e coerente, ovvero invertibile[1] ed esente da guadagno o perdite certe ($p_E \times S \leq S$ e $p_E \leq 1$). La probabilità così definita mantiene le proprietà 1.1 ed in particolare, nonostante l'aspetto soggettivo, può far riferimento alla teoria assiomatica di Kolmogorov (1950).

L'impostazione della scommessa riconduce al trattamento scientifico l'attribuzione di probabilità ad un qualunque evento (come può immaginare chiunque conosca la reputazione dei bookmaker inglesi), estendendo virtualmente all'infinito il campo applicativo della disciplina.

I problemi posti dall'approccio definettiano riguardano in particolare la dipendenza della funzione di probabilità anche dal grado individuale di propensione al

[1] Per scommessa invertibile qui si intende la scambiabilità tra banco e scommettitore.

rischio del soggetto, soprattutto in relazione all'entità della somma S posta in gioco. In risposta a questa critica si suggerisce di limitare S a piccoli importi per annullare tale effetto.

Le maggiori critiche riguardano però l'interpretazione bayesiana nel suo complesso e puntano direttamente al fattore soggettivo. La soggettività dell'impostazione bayesiana viene generalmente ridotta a semplice arbitrarietà, negando la validità scientifica dell'intero costrutto teorico. Quando anche ne vengano accettati gli assunti fondamentali, l'enorme sforzo computazionale richiesto dalla statistica bayesiana viene considerato un limite ben maggiore delle forti assunzioni richieste dalle teorie oggettiviste.

In risposta alle critiche teoriche, i bayesiani sostengono che gli approcci oggettivisti alla probabilità nascondono scelte arbitrarie preliminari, accettando delle assunzioni anche quando non sono vere in realtà. É molto più onesto esplicitare subito tutte le scelte arbitrarie effettuate nel corso dell'analisi in modo da controllarne coerenza e razionalità.

Per quanto riguarda le difficoltà computazionali, la statistica soggettivista ha contribuito a sviluppare e fornisce costantemente lo stimolo per nuovi algoritmi e metodologie, utili per le applicazioni bayesiane, ma sfruttati anche in ambito frequentista.

1.2 La Causalità: problemi epistemologici ed empirici

Lo studio di relazioni causali rientra fra gli obiettivi principali della ricerca scientifica. L'evidenza che ad un seme interrato oggi, nelle giuste condizioni, corrisponde una pianta nuova domani, la scoperta che alla profilassi dei medici in uscita dall'obitorio ed in entrata nella sala parto corrisponde una drastica riduzione della mortalità da parto sono solo pochi esempi dei progressi che l'umanità ha fatto grazie alla scoperta di relazioni causali.

Nonostante, o proprio in ragione dell'ampiezza del concetto di causalità e del fatto che essa rappresenti il fondamento comune a tutte le scienze, la sua definizione è teoricamente controversa. Il dibattito sulla nozione di relazione causale percorre tutta la storia della filosofia occidentale e rimane a tutt'oggi aperto perché non si è giunti ancora ad una formulazione conclusiva condivisa.

Se la definizione pone grosse difficoltà, la rilevazione empirica rappresenta la sfida principale dal punto di vista statistico. L'indizio più comunemente accettato di presunzione di causalità è la *covariazione* di due fenomeni, ma tale condizione non è sufficiente per stabilire con certezza il rapporto di causalità e non sempre aiuta a specificare quale sia la causa e quale l'effetto, lasciando irrisolto il problema

di identificare le relazioni causali effettive e distinguerle da quelle "spurie" e di accertare la direzione ed il verso della relazione stessa.

Nel paragrafo 1.2.1 vengono riportate le maggiori correnti interpretative del concetto di causalità mentre nel paragrafo 1.2.3 si discute della rilevazione empirica della causalità, facendo riferimento alle due principali correnti metodologiche: la teoria ipotetico deduttiva e la teoria induttiva.

1.2.1 La definizione di Causalità

Le posizioni relative al concetto di causalità sono molteplici e richiamano proprietà dicotomiche che non si escludono tra loro. Una definizione di causalità si caratterizza come **mentale** se interpretata come una caratteristica dello stato epistemico dell'agente che la rileva o **fisica** se interpretata come una caratteristica del mondo esterno. Inoltre essa è **oggettiva** se non può essere ricondotta ad una scelta arbitraria e **soggettiva** se due agenti differenti possono essere in disaccordo nell'identificarla. Infine è necessario distinguere definizioni di causalità che si possono applicare eslusivamente ad **eventi singoli** ed invece definizioni che sono valide anche per **aggregati**.

Tenendo conto di tali proprietà si possono distinguere diverse posizioni interpretative. Una corrente, rappresentata da Russell (1912), sostiene che la causalità è un concetto extrascientifico poiché solo le relazioni funzionali sono l'oggetto della scienza. Alcuni accettano la scientificità del costrutto causale, ma lo considerano talmente basilare da trattarlo come concetto primitivo non ulteriormente analizzabile.

La posizione esposta di seguito, considera invece la causalità composta da un insieme di concetti più semplici non implicanti relazioni causali. A seconda di quali concetti si considerano alla base della causalità si presentano diversi approcci.

Teoria del meccanicismo causale

L'approccio meccanicistico alla causalità è tra i più antichi nella storia della filosofia occidentale e nella versione contemporanea trova i suoi maggiori esponenti in Salmon (1984) e Dowe (2010).

Secondo questa impostazione ogni fenomeno può essere analizzato nelle sue componenti fisiche elementari, quindi le relazioni causali fra fenomeni non sono altro che interazioni riconducibili al possesso o trasmissione di determinate quantità fisiche invarianti, quali quantità di moto, carica elettrica, ecc. Data la stabilità delle leggi fisiche, l'esito di ciascun evento è già predeterminato in partenza. Il motivo per cui non è possibile avere una conoscenza perfetta di passato e futuro risiede solo nella limitatezza della ragione umana che non possiede i mezzi sufficienti per sostenere lo sforzo computazionale necessario a tale calcolo; utilizzando

le parole di Laplace *e altri* (1995): "lo stato attuale dell'universo è l'effetto dello stato passato e la causa dello stato futuro".

In una siffatta impostazione, le relazioni causali sono caratterizzate dall'essere fisiche ed oggettive e sebbene si riferiscano principalmente a singoli eventi, possono essere indotte da regolarità causali singole e poi generalizzate ad aggregati.

Le critiche principali a cui si espone la teoria meccanicistica riguardano in primo luogo la totale carenza di libero arbitrio ed il determinismo che presuppone, poi il livello esplicativo troppo basso a cui fa riferimento. La riduzione alla sola componente fisica della realtà mal si adatta a spiegare fenomeni comportamentali quali quelli economici e sociologici, ad esempio. Fare riferimento a relazioni troppo elementari nega l'esistenza di causalità comportamentale o come minimo, la considera troppo complicata per essere trattata scientificamente.

Per risolvere l'*empasse* è stato proposto di trattare separatamente le relazioni fisiche da quelle sociali, che restano però ancora da definire; evidentemente l'operazione non fa altro che riproporre il problema sotto una forma diversa. Inoltre il concetto di causalità è così fondamentale da richiedere necessariamente una successiva ricomposizione dei due tipi di relazione causale (fisica e sociale), aggravando lo sforzo epistemologico.

Teoria della causazione probabilistica

Un altro aspetto critico della teoria meccanicistica riguarda l'indeterminazione insita nella natura che la causazione deterministica trascura colpevolmente. Non è possibile addebitare alla sola limitatezza umana l'incapacità di descrivere perfettamente l'universo. La stessa teoria dell'atomo presuppone che l'incertezza sia insita nelle leggi fisiche. Da qui la necessità di richiamare la teoria probabilistica nella definizione delle relazioni causali.

Secondo l'approccio probabilistico alla causalità, in linea generale, il fenomeno C causa probabilisticamente il fenomeno E se C precede E e l'occorrere di C accresce le probabilità del verificarsi di E. Naturalmente la causazione non coincide necessariamente con la presenza di una relazione probabilistica, per diverse ragioni. Due eventi possono essere stocasticamente dipendenti senza il minimo indizio di causalità: anche se il prezzo del pane presenta una forte relazione statistica con il numero di piccioni in piazza Duomo, trovare una relazione causale tra i due fenomeni ha dell'impossibile. A ciò va aggiunto che il caso di variazione contemporanea di due fenomeni può dipendere da una causa comune ad entrambi che genera una relazione causale apparente (spuria).

Questo secondo punto viene ripreso nella definizione del principio delle cause comuni che prevede che se due variabili sono probabilisticamente dipendenti, allora o sono l'una causa dell'altra o hanno una causa in comune.

La definizione della relazione causale resta fisica ed oggettiva, indipendentemente dall'interpretazione della probabilità utilizzata. Inoltre si riferisce ai casi singoli ma può essere generalizzato ad aggregati.

In ogni caso, questa interpretazione pone diversi problemi relativamente alla determinazione del verso della relazione causale. Quando l'ordine temporale della manifestazione di due fenomeni non è chiaro, l'approccio probabilistico non riesce sempre a discriminare fra causa ed effetto senza l'intervento di un decisore razionale.

Teoria controfattuale della causalità

L'interpretazione controfattuale, sviluppata da Lewis (2001), riduce le relazioni causali a subgiuntivi condizionali, ovvero affermazioni sulla relazione fra due eventi del tipo: se C allora E. Tale affermazione implica che E dipende da C se e solo se:

- Se C si verifica, allora si verifica necessariamente anche E (o le sue probabilità di verificarsi aumentano drasticamente).

- Se C non si verifica, anche E non si verifica (o almeno le sue probabilità di verificarsi calano drasticamente).

Quando l'antecedente è falso il condizionale subgiuntivo prende il nome di *controfattuale*.

La causalità così definita è applicabile al caso singolo e si intende come fisica ed oggettiva, poiché la validità dell'affermazione non è in relazione con le conoscenze o lo stato del soggetto ma con le condizioni in cui si trova il mondo fisico.

L'interpretazione di Lewis per convertire un'affermazione condizionale in una relazione causale, fa appello alla semantica degli universi possibili e al concetto di similarità di queste realtà parallele. In sostanza "C causa E" è banalmente vero se non ci sono mondi possibili in cui possa verificarsi C ed è propriamente vero se, tra i mondi in cui C si verifica, quelli in cui anche E si manifesta sono più vicini ad ogni altro in cui E non si manifesta, ed è falso altrove.

I problemi posti dall'approccio controfattuale derivano proprio dalla sintassi degli universi possibili, che nella concezione di Lewis non sono solo un espediente esplicativo ma esistono realmente in una dimensione separata dalla nostra fisicamente.

Essendo tali realtà separate e quindi inconoscibili le une per le altre, si pone il problema di come sia possibile: a) verificare il manifestarsi o meno di un fenomeno, b) valutare la distanza fra due universi separati.

Teoria manipolativa della causalità

I principali esponenti della corrente manipolativa possono essere considerati Menzies e Price (1993) e l'obiettivo essenziale dell'approccio è quello di analizzare le

relazioni causali in termini di capacità di un eventuale agente di raggiungere i propri obiettivi manipolandone le cause.

Secondo questa interpretazione, C causa E se agire su C corrisponde ad una effettiva azione su E. L'azione su C è giudicata effettiva se una teoria delle decisioni razionale la prevede come un modo per agire (manipolare) su E. Menzies e Price sostengono che tale strategia è razionale se essa aumenta la probabilità (secondo l'agente) che si verifichi E. La probabilità chiamata in causa è di tipo chance per Menzies e bayesiana per Price.

La teoria manipolativa contribuisce a limitare il problema della covariazione accidentale: è possibile che il prezzo del pane sia statisticamente dipendente dal numero di piccioni in piazza Duomo, ma è altrettanto ovvio che abbattere piccioni non fa diminuire il prezzo del pane.

L'ottica manipolativa si adatta bene alla metodologia statistica dell'inferenza causale e vede la sua realizzazione nell'esperimento randomizzato. Purtroppo tale pratica, ampiamente diffusa nella statistica sperimentale trova maggiori difficoltà applicativa su dati osservazionali.

Le critiche all'approccio manipolativo si possono riassumere in due punti:

-Teoricamente si cerca di ridurre la causalità alla manipolazione, ma non si riesce a descrivere quest'ultima in termini non causali.

-Non è chiaro se possa esistere causalità ove non sia possibile, neppure teoricamente, l'intervento umano.

1.2.2 La modellizzazione della Causalità

Qualunque sia l'interpretazione della casualità adottata, è comunemente accettato che tutti i fenomeni dell'universo sono, in maggior o minor grado, interconnessi fra loro. Non c'è la possibilità che, all'interno di una stessa realtà, esistano fenomeni completamente isolati e se anche esistessero, in quanto tali risulterebbero inconoscibili all'uomo. In definitiva ogni evento può essere rappresentato come il centro di una rete di relazioni causali, di cui è di volta in volta soggetto attivo o passivo.

Queste relazioni complesse possono essere scomposte, a scopo esplicativo, in un aggregato di relazioni più semplici tutte riconducibili ad una relazione di *causazione diretta*. Il legame diretto entra poi a far parte di altre combinazioni piuttosto elementari che sono:

-*relazione indiretta*

-*relazione reciproca*

-*relazione condizionata*

-*relazione spuria*.

I suddetti schemi teorici rappresentano uno dei principali strumenti che la filosofia occidentale ha fornito per rappresentare la causalità e di conseguenza tutta la realtà.

Relazione diretta

La relazione causale *diretta* è un concetto controverso nella sua definizione ma sempre interpretabile come relazione binaria in cui la variazione del fenomeno-causa **agisce asimmetricamente** su un effetto **contiguo** nello spazio-tempo. La rappresentazione usuale del legame diretto è quella di figura 1.1, dove una freccia indica direzione e verso della relazione causale (causa → effetto)

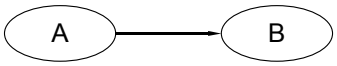

Figura 1.1: Rappresentazione grafica del legame diretto singolo.

Le tre componenti fondamentali del legame diretto singolo sono perciò:

Azione. Un legame si può definire causale se la variazione del presunto effetto è riconducibile direttamente alla variazione della presunta causa che da essa viene generata, in altri termini, la covariazione non deve essere casuale ne' tanto meno conseguenza dell'azione di un terzo fattore su entrambi.

Questa fondamentale caratteristica, che faccia riferimento alla dimensione fisica, alla manipolabilità degli eventi, alla loro costruzione logica o probabilistica, determina una prima distinzione fra ciò che è causalità e ciò che non lo è.

Asimmetria. La causazione diretta è essenzialmente una relazione binaria asimmetrica che assegna ruoli diversi ai due fenomeni implicati. Esiste un unico verso in una relazione causale diretta che non può essere invertito; questo verso si esplica nella direzione temporale della relazione: qualsiasi sia l'interpretazione adottata, la causa precede sempre l'effetto.

Tale aspetto è talmente intrinseco da rientrare anche nella definizione del concetto di tempo, il quale risulta contraddistinto dalla monodirezionalità, al contrario delle altre tre dimensioni fisiche conosciute. Sono i paradossi causali che rendono impossibile invertire la direzione temporale come invece è possibile fare con il senso di marcia nello spazio. L'asimmetria temporale non può prescindere dall'asimmetria causale e viceversa.

Contiguità. Perché una causa si possa definire diretta, è indispensabile che essa sia prossima nel *tempo* e nello *spazio* al suo effetto. Se la variazione della causa si trasmette attraverso la variazione di fenomeni intermedi, allora l'azione totale può essere scomposta in una catena di cause dirette tali che fra due fenomeni in relazione non interferisca nessuna causa terza. Solo in presenza di contiguità si può riscontrare causazione *diretta*, negli altri casi, non viene esclusa l'esistenza di un legame causale, ma devono essere presi in considerazione i fattori intermedi.

Sebbene la relazione causale diretta sia il tassello base su cui costruire tutto l'impianto di interconnessioni causali costituenti la realtà, è possibile considerare

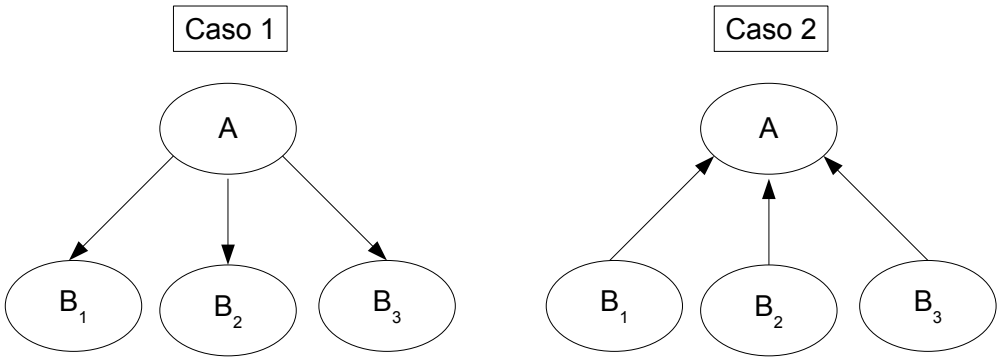

Figura 1.2: Rappresentazione grafica del legame diretto multiplo.

delle configurazioni causali composte ma ancora sufficientemente elementari di cui una prima generalizzazione sono le relazioni dirette multiple: una stessa causa agisce su più effetti (figura 1.2, Caso 1) oppure due o più cause agiscono contemporaneamente su un solo effetto (figura 1.2, Caso 2); in quest'ultima situazione il fine dell'indagine statistica è quello di rilevare l'apporto specifico di ciascuna causa all'effetto finale, ed una classica formalizzazione è data dalla regressione multipla.

Relazione indiretta o in serie

Un'ulteriore configurazione composta da prendere in considerazione è la relazione *indiretta* o in serie. Come appena spiegato l'azione diretta rappresenta la base della causazione, ma gli eventi, nella realtà, fanno parte di concatenazioni causali più ampie, in cui l'azione della causa si può trasmettere, attraverso molteplici passaggi fra le variabili mediatrici, su effetti anche molto lontani nello spazio e nel tempo[2].

Una semplice relazione indiretta è rappresentata in figura 1.3 dove l'azione di A viene *mediata* da B prima di arrivare a C.

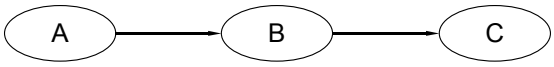

Figura 1.3: Rappresentazione grafica del legame indiretto.

L'individuazione esatta degli elementi di una serie causale è altrettanto importante che distinguere due covariazioni del tutto accidentali. È fondamentale comprendere come si trasmette l'azione principale per identificare le leve operative e gli anelli deboli e per ottenere una visione precisa dell'effetto totale.

[2]Fenomeno conosciuto, nella sua massima manifestazione, come "effetto farfalla".

Per comprendere meglio l'importanza della corretta specificazione degli elementi della serie causale si prenda in considerazione il seguente esempio. Nel '66 Philips presenta una ricerca atta a mostrare che la relazione fino ad allora considerata diretta, fra la razza e l'intelligenza umana, in realtà non è altro che l'estremità di una serie di fenomeni in cui il livello di istruzione funge da variabile mediatrice. Infatti, nei casi presi in esame, a parità di livello d'istruzione, l'intelligenza è indipendente dalla razza.

L'utilità della corretta specificazione causale, in questo caso, sta nel fatto che, alla luce dei nuovi elementi, risulta logico sostituire ambigue politiche di selezione genetica con una azione più diretta mirante ad aumentare il livello d'istruzione.

Distinguere le azioni dirette da quelle mediate permette di gestirne meglio gli effetti, ma spesso la situazione si presenta in maniera più complicata, e uno stesso fenomeno può agire per più vie. Relazioni dirette ed indirette possano combinarsi come nel semplice caso di figura 1.4, in cui la causazione si esplica oltre che in un intervento diretto, anche in un intervento mediato da un'altra concausa.

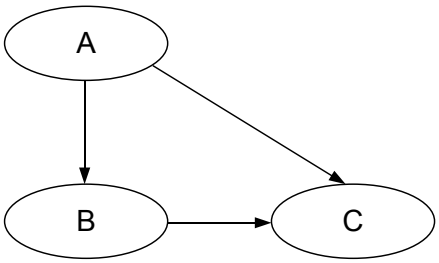

Figura 1.4: Rappresentazione grafica del legame misto (diretto ed indiretto).

In ultimo, non è detto che gli effetti siano concordi; è possibile che causa ed effetto non siano riconoscibili come tali per la presenza di una terza variabile che influisce su entrambe con segno opposto, la cui assenza nel modello maschera la relazione. L'inserimento nel modello esplicativo del fenomeno erroneamente escluso contribuisce alla corretta formalizzazione della relazione causale.

Relazione reciproca

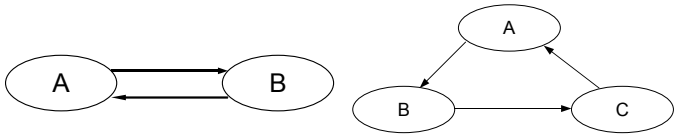

Figura 1.5: Rappresentazione grafica del legame reciproco e del legame ciclico.

Esistono coppie di fenomeni in relazione tra loro per cui non è però possibile identificare quale sia la causa e quale l'effetto poiché ad ogni azione dell'una corrisponde una retroazione o feed-back da parte dell'altra. La situazione può essere generalizzata considerando un gruppo di variabili reciprocamente concatenate per le quali ogni azione si trasmette a tutte le altre in circolo. In figura 1.5 sono rappresentate entrambe le relazioni.

Il problema generato dalla simmetria della retroazione viene il più delle volte risolto riconducendo la circolarità ad una serie causale temporale asimmetrica, in cui il feed-back viene riconvertito nell'azione in un tempo successivo. Anche questa soluzione presenta grosse difficoltà; in primo luogo non è detto che il ciclo sia stabile nel tempo, la forza delle relazioni implicate può affievolirsi per smorzarsi del tutto come può esplodere distruggendo il sistema o, in alternativa, il sistema può trovare un equilibrio stabile. In tutti i casi, stimare se e a quale equilibrio converge il sistema non è semplice, come non lo è neppure stimare la forza complessiva della relazione causale dato che essa non si esercita in un tempo finito.

La conseguenza più grave dal punto di vista teorico è che l'asimmetria temporale genera confusione fra cause ed effetto e si innesca il paradosso dell'uovo e della gallina, per cui non si riesce ad individuare la causa originale.

Relazione condizionata o interazione

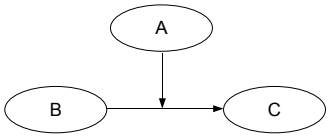

Figura 1.6: Rappresentazione grafica di interazione.

Vi sono casi in cui è lo stesso legame ad essere oggetto di azione causale. Quando la relazione fra due fenomeni varia di intensità o di segno, in dipendenza da una terza variabile avviene che quest'ultima "causa" la relazione causale stessa, nel senso che la determina. Se si considera la rappresentazione classica della situazione fornita dalla figura 1.6, si può notare come, in questo caso, la freccia termini sul legame anziché su una variabile indicando per l'appunto che esso è l'effetto. La variabile che esercita questa influenza si dice moderatrice.

In pratica è come se esistessero diverse configurazione della relazione causale, una per ogni modalità della variabile moderatrice, che determinano un cambiamento nel legame fra le variabili originarie (figura 1.7).

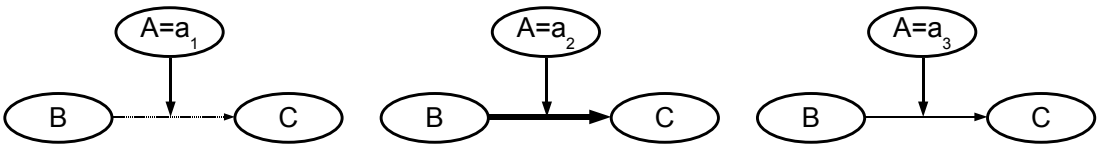

Figura 1.7: Rappresentazione grafica del l'effetto delle variazioni del fenomeno A sulla relazione causale tra B e C.

Per fare un esempio banale, è come dire che la relazione fra pubblicità e propensione all'acquisto abbia una certa intensità nell'intera popolazione, ma che essa diventi molto debole per la sola fascia ad alta istruzione e invece molto forte nel resto della popolazione, il che implica la possibilità di dividere il modello globale in due o più sottomodelli che meglio si adattano alle sottopopolazioni di riferimento.

Anche per le interazioni vale che, nel caso in cui gli effetti all'interno delle sottopopolazioni siano equivalenti ma di segno opposto, può verificarsi una situazione in cui le azioni parziali possono compensarsi creando un'apparente indipendenza a livello globale: solo l'inserimento nel modello della variabile moderatrice fa emergere la reale struttura di causalità.

Pseudo relazione causale o relazione spuria

Non sempre alla covariazione di due fenomeni corrisponde l'effettiva influenza dell'uno sull'altro. Tale situazione si verifica quando viene a mancare l'*azione* di un fenomeno sull'altro. In tal caso la causalità viene del tutto esclusa, infatti la relazione di definisce **spuria** o pseudo causale. Nella maggior parte dei casi si tratta di una mera coincidenza, di covariazione casuale, spesso però essa è determinata dall'azione contemporanea di una terza causa su entrambi i fenomeni in oggetto, come da figura 1.8.

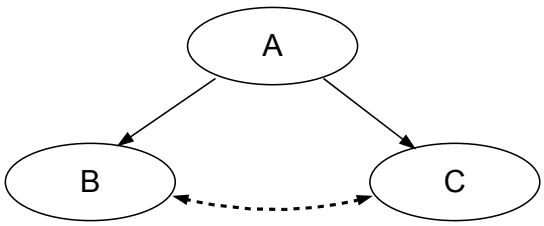

Figura 1.8: Rappresentazione grafica della relazione spuria.

Le pseudo relazioni creano gravi problemi statistici perché, mentre nei casi di relazioni indirette o condizionate un legame causale esiste, seppur mal specificato, in questo caso il legame è del tutto assente e agire sulla pseudo-causa non pro-

duce alcuna conseguenza, ne indiretta ne limitata ad una sottopopolazione della variabile pseudo-effetto.

Gli esempi nella storia della ricerca scientifica sono innumerevoli e mostrano tutti che le relazioni spurie spariscono quando nel modello viene inclusa la causa comune per i due fenomeni.

Da questa presentazione introduttiva della modellizzazione delle relazione casuali è possibile trarre subito una conclusione non marginale e cioè che la corretta identificazione di tutte le variabili implicate nella rete di causazione è indispensabile per evitare di rilevare effetti spuri o sottovalutare l'azione di alcuni fattori. Per tale motivo è fondamentale un'analisi multivariata per ottenere una visione globale e completa e controllare che i fenomeni trascurati dal modello siano davvero irrilevanti per la rete di relazioni causali analizzata.

Il problema della misura

Quanto detto fin'ora si basa sulla possibilità di rilevare e quantificare la variazione di un fenomeno in modo da metterla in relazione con la variazione dell'ambiente circostante per individuare legami di causa-effetto; la misura di tale cambiamento risulta quindi un indicatore essenziale per la ricerca delle relazioni casuali, ma non sempre le variabili implicate in legami causali sono di facile misurazione. Nella maggior parte dei casi, l'oggetto della ricerca, soprattutto in campo sociale, è rappresentato da costrutti astratti o *latenti*, come vengono comunemente definiti, di cui si ha la percezione ma che non è possibile rilevare direttamente. Tali sono ad esempio l'intelligenza umana, la ricchezza degli Stati, la salute fisica di un individuo o la salute economica di un'azienda.

Non esiste uno strumento di misura per catturare interamente queste variabili però sono stati proposti degli strumenti statistici allo scopo di sostituire la misura diretta dal fenomeno con particolari combinazioni di misure indirette effettuate sulle sue manifestazioni concrete. In pratica, ipotizzando che il concetto latente da valutare sia in relazione causale con alcuni fenomeni osservabili, questi vengono utilizzati come indicatori e diventanno gli strumenti di misura del concetto astratto.

La relazione fra variabili manifeste e variabile latente può espletarsi secondo due schemi principali:

- Le manifeste sono causate dalla latente e quindi ne *riflettono* le variazioni (schema **riflessivo**).

- Le manifeste causano la latente e quindi ne producono, ne *formano* le variazioni (schema **formativo**).

Quest'ultima situazione è riconducibile al legame di causalità diretta multipla esposto nel paragrafo 1.2.2 e rappresentato in figura 1.2-Caso2.

Lo schema riflessivo si compone invece di molteplici legami diretti latente-manifesta come in figura 1.2-Caso1; esso si presenta come un tipico caso di relazione spuria dove l'azione della latente viene scambiata per interazione causale degli indicatori.

Dalla combinazione dei due schemi precedenti è possibile ricavare un modello di misura misto (MIMIC) in cui il blocco delle variabili manifeste viene partizionato in sottogruppi riconducibili ai due schemi precedentemente illustrati.

La formalizzazione del modello causale e dei modelli di misura permette di effettuare analisi statistiche su fenomeni altrimenti trattabili solo qualitativamente e contribuisce ad accrescere il campo della ricerca sulla causalità. Essi verranno presentati più dettagliatamente in seguito, inseriti nel contesto statistico in cui si intende utilizzarli.

1.2.3 La ricerca di strutture causali

Una volta determinati gli strumenti di misura necessari e stabilito che ad ogni diverso modo di vedere la natura delle relazioni causali corrisponde un diverso metodo per rilevarle (i meccanicisti analizzano i processi fisici, i probabilisti cercano dipendenze ed indipendenze statistiche, ecc), in generale è possibile dividere il campo della ricerca della causalità in due posizioni trasversali ai vari approcci interpretativi:

-Deduttivismo-ipotetico

-Induttivismo

Le differenze tra le due opposte concezioni risiedono non tanto nelle operazioni necessarie alla ricerca quanto nell'ordine in cui effettuarle e nella loro importanza relativa. I deduttivisti ritengono necessario ipotizzare una relazione causale e solo in seguito testare se si accorda con i dati reali. Gli induttivisti utilizzano sofisticate tecniche di analisi dei dati per ricavare direttamente dai dati le ipotesi relazionali. Le omologie sono tali per cui nella pratica la stessa tecnica statistica viene utilizzata con due interpretazioni differenti.

Metodo di ricerca ipotetico deduttivo

Seguendo la concezione ipotetico deduttiva, il ricercatore deve procedere ipotizzando una relazione causale fra i fenomeni e deducendone successivamente le necessarie conseguenze. Se confrontando le deduzioni con i dati reali gli effetti della teoria sono incompatibili con la rilevazione empirica, sarà necessario abbandonare le ipotesi iniziali e procedere con una nuova formulazione.

Gli specifici test di conformità, dipendono strettamente dall'interpretazione di causalità adottata: si può rifiutare l'ipotesi in base al risultato di esperimenti fisici,

di rilevazioni statistiche o di esperimenti manipolativi. L'elemento essenziale è che il momento sperimentale sia subordinato a quello teorico.

Per Popper, lo schema della scoperta, ovvero le regole empiriche da seguire nella ricerca scientifica, sono:

a) Ipotizzare un modello causale che spieghi la relazione fra i fenomeni indagati.

b) Dedurre le conseguenze che l'ipotesi iniziale implica.

c) Raccogliere dati sperimentali o osservazionali che possano falsificare le deduzioni effettuate, proseguendo finché non dovessero emergere prove empiriche in contrasto con le ipotesi teoriche e le loro deduzioni, punti a) e b), nel qual caso la teoria deve ritenersi falsificata e si deve ripartire dal punto a).

Un'affermazione scientifica *non falsificata* non può ritenersi automaticamente *confermata*, poiché, secondo Popper, non è possibile accertare definitivamente una legge universale sulla base di rilevazioni empiriche parziali. Solo quando si è certi di aver rilevato tutte le possibili manifestazioni congiunte di due o più fenomeni, i dati raccolti danno questa informazione, altrimenti resta sempre il dubbio che esita un caso sconosciuto in contraddizione con le ipotesi.

La falsificabilità di ogni teoria è il punto fondamentale dell'approccio popperiano, infatti la formulazione stessa delle ipotesi deve comprendere la descrizione di eventi che siano in grado di mettere in crisi l'ipotesi stessa. La scientificità di una teoria risiede nella sua falsificabilità, affermazioni non testabili non possono entrare a far parte del patrimonio della conoscenza scientifica.

Il falsificazionismo empirico di Popper fornisce uno strumento potente per la salvaguardia del rigore scientifico ma al contempo presta il fianco a molte critiche a causa della sua rigidità. Un approccio più flessibile sostiene che una legge causale possa essere confermata dall'evidenza anche solo in base alla probabilità dell'ipotesi condizionale. La conferma universale viene sostituita con una misura di affidabilità: si accetta l'ipotesi causale fino al punto in cui l'evidenza la conferma.

Fortemente orientato alla fase di test delle ipotesi, Popper tratta molto superficialmente il momento creativo del processo scientifico. Seppur ammette dei meccanismi razionali nel momento della formulazione dell'ipotesi, il filosofo austriaco ne rimanda la trattazione approfondita alla psicologia. Un'altro eminente deduttivista, invece, il matematico ungherese Pòlya, pur non arrivando a definire una metodologia precisa che generi automaticamente delle ipotesi scientifiche, fornisce delle utili procedure euristiche. Ad esempio sulle conseguenze della falsificazione, Popper non si esprime, facendo supporre una tabula rasa della teoria, Pòlya al contrario propone di verificare se non sia possibile apportare degli aggiustamenti locali al costrutto teorico, riformulandolo in modo tale che, pur mantenendo coerenza logica, le sue conseguenze non siano più in contrasto con i dati empirici. Resta comunque il fatto che, generalmente, l'approccio deduttivista difficilmente approfondisce il meccanismo di generazione delle ipotesi, che invece è la caratteristica

principale del metodo induttivo.

Metodo di ricerca induttivo

La teoria induttivista segue lo schema mentale di apprendimento tipico dell'essere umano, il quale formula, più o meno razionalmente, le ipotesi sulla natura che lo circonda dopo aver accumulato una certa esperienza. Nell'ambito della ricerca scientifica, il metodo induttivo propone di determinare delle metodologie di apprendimento, principalmente statistiche, che generino o almeno supportino la formulazione delle ipotesi causali iniziali. Una volta estratta la maggior quantità possibile di informazioni dai dati, è necessario effettuare nuove rilevazioni per testare le ipotesi fatte. Nella pratica, se è antieconomico o impossibile acquisire nuovi dati, ci si accontenta di estrarre dal dataset a disposizione un sottocampione ridotto e tralasciarlo nell'analisi, per poi testarvi la bontà del modello.

Un primitivo esempio di schema induttivo è la "presentazione delle istanze dell'intelletto" di Bacon *e altri* (2000), una procedura pianificata atta all'investigazione scientifica che si compone di due parti:

a) la *tabula presentiae*

b) la *tabula absentiae in proximitatae*

Quando un fenomeno si presenta indifferentemente in contemporanea o meno ad un altro, si può affermare che essi sono statisticamente indipendenti. Perché due fenomeni siano legati da una relazione causale, il fattore causale deve essere condizione necessaria (ma non obbligatoriamente sufficiente) del suo effetto. I due fenomeni devo essere presenti contemporaneamente nella prima *tabula* e, nella seconda *tabula*, la causa non può essere mai assente quando è presente l'effetto (il riferimento alla teoria controfattuale della probabilità è evidente). Compilando le due *tabulae* con dati osservazionali o sperimentali multidimensionali, lo schema delle *istanze* fornisce un rudimentale ma efficace metodo di ricerca scientifica.

La disciplina statistica ha sviluppato tecniche sempre più raffinate di induzione dai dati grezzi, di cui l'analisi delle componenti principali e gli algoritmi di apprendimento bayesiani sono solo alcuni esempi. Purtroppo gli schemi d'apprendimento da soli non sono sufficienti ad eliminare il problema della covariazione casuale, ma richiedono l'intervento di un soggetto razionale che riesca ad escludere le ipotesi statisticamente valide ma implausibili o a restringere il campo di indagine ad un sottoinsieme di configurazioni obbligatorie, quando le conoscenze a priori dell'ambito applicativo lo impongano.

Principalmente dall'area di ricerca bayesiana, arrivano proposte di algoritmi di apprendimento che implementano la possibilità di inserire black list (liste di relazioni proibite, perché illogiche o irrazionali) e white list (relazioni avvalorate da conoscenze a priori).

Specularmente al metodo ipotetico deduttivo, l'induttivismo, assume come prioritaria la fase creativa, quindi presenta maggiori incertezze nella fase di test.

Gli induttivisti si ripropongono di estrarre la maggior quantità possibile di informazioni dai dati, di conseguenza cercano di limitare al massimo la formulazione di ipotesi a priori, ma senza assunzioni statistiche preliminari è più difficile la falsificazione. Inoltre il metodo induttivo è generalmente meno rigido sulla conferma delle ipotesi, in quanto considera i dati in accordo con la teoria come evidenze corroborative. Da qui la critica di Popper e dei deduttivisti alla generalizzazione universale delle evidenze empiriche: il fatto che il sole sia sorto negli ultimi 4.5 miliardi di anni non implica che sorgerà anche domani.

Se dal punto di vista teorico le due metodologie sono in netto contrasto, in pratica i ricercatori adottano essenzialmente un approccio misto e soprattutto variabile a seconda degli obiettivi che si ripropongono di raggiungere.

Nel momento iniziale di formulazione delle ipotesi, gli studi preliminari di contesto vengono supportati dall'analisi dei dati in modo tale che la teoria non si discosti eccessivamente dalla realtà. Successivamente si procede con la raccolta di nuove evidenze empiriche, anche applicando i modelli presentati a realtà differenti per testarne la robustezza. Maggiori sono i risultati positivi e maggiore è il credito che acquisisce la teoria all'interno del mondo scientifico, fino al momento in cui dovessero emergere evidenze empiriche contrastanti. In tal caso il ricercatore sarebbe costretto a riformulare le ipotesi o, al limite, abbandonare tutta la struttura teorica e ricominciare la ricerca dalle basi.

Bisogna ammettere che, in realtà, difficilmente i ricercatori sono disposti ad abbandonare una teoria, soprattutto se l'hanno creata e hanno passato la maggior parte della loro vita a perfezionarla. Spesso si procede alla disperata ricerca di ipotesi aggiuntive che rendono eccessivamente complesso il costrutto teorico allo scopo di metterlo al riparo dalle evidenze fattuali. Questo genere di tentativo non può resistere al *rasoio di Ockham* e prima o poi ogni teoria eccessivamente complessa viene sostituita da una formulazione più semplice, se esiste. Per usare le stesse parole del filosofo inglese: " Pluralitas non est ponenda sine necessitate"[3] o in altri termini, a parità di fattori, è sempre da preferire la spiegazione più semplice.

[3]É necessario evitare di formulare ipotesi aggiuntive non strettamente necessaire alla spiegazione.

1.3 Causalità, Modelli ad Equazioni Strutturali e Reti Probabilistiche Bayesiane

I modelli statistici con cui trattare la causalità sono numerosi e la maggior parte prescinde dalle varie interpretazioni filosofiche, sebbene tutti abbiano un orientamento particolarmente spiccato per l'uno o l'altro degli approcci alla ricerca presentati nel paragrafo 1.2.3.

Naturalmente sono rari i metodi statistici totalmente induttivi o deduttivi, ma è pur vero che ogni tecnica segue più spiccatamente una delle due correnti, a seconda degli obiettivi prefissati.

Riguardo ai modelli trattati nel prosieguo del lavoro, i modelli ad equazioni strutturali sono sicuramente più orientati all'approccio deduttivo. Essi fanno riferimento a reti di relazioni causali predeterminate, generate all'interno dell'ambito di ricerca a cui si applicano, proposte da esperti del settore.

L'obiettivo principale di questo genere di modellistica è quello di quantificare la forza dei legami ed eventualmente validarne la significatività. Nonostante ciò, esistono differenze fra tecniche di stima di tipo confermativo, finalizzate alla stabilizzazione delle basi teoriche dell'ambito di studio, un esempio delle quali è fornito dai SEM *covariance based*, e approcci più orientati all'analisi dei dati e alla predittività, come i SEM *component based*.

D'altro canto, le reti probabilistiche bayesiane si inseriscono tra le tecniche di *Machine Learning* e hanno un chiaro orientamento induttivo, in quanto tecniche di apprendimento automatico della struttura causale che utilizzano i soli dati grezzi. Anche le reti probabilistiche bayesiane assumono un orientamento deduttivo quando l'*expertise* del decisore viene integrata nella definizione della struttura causale ed il fine dell'analisi diventa la sola stima della forza dei legami causali.

Nel seguito del lavoro vengono esposti in dettaglio sia modelli ad equazioni strutturali sia le reti probabilistiche bayesiane e, traendo spunto da una applicazione concreta, vengono proposte delle prospettive di integrazione dei due modelli statistici in modo che, utilizzando le potenzialità deduttive degli uni e quelle induttive delle altre, si possa rendere più efficace la ricerca della causalità.

Capitolo 2

I Modelli ad Equazioni Strutturali

I Modelli ad Equazioni Strutturali (di seguito indicati anche con l'acronimo SEM[1]) [Bollen (1989); Kaplan (2008)]sono dei modelli di regressione multivariata, che al contrario dei più tradizionali modelli di regressione lineare prevedono la possibilità che nello stesso sistema di equazioni, ciascun fenomeno implicato nella rete di relazioni causali ricopra sia il ruolo di variabile esplicativa che di variabile risposta (da cui il nome di Modelli ad Equazioni Simultanee). Lo Structural Equation Modeling si presta dunque a modellare relazioni di causalità multipla e/o mediata, disponendo di un insieme di tecniche statistiche che, sulla base di uno schema causale stabilito, consente di stimare il segno e la forza di una rete di legami causali e verificarne la coerenza e la rilevanza empirica.

Tale approccio è, per costruzione, deduttivo in quanto lo schema su cui si basa la ricerca non viene generato autonomamente da un'analisi dei dati ma è una conseguenza di considerazioni teoriche a priori, spesso qualitative e quasi sempre suggerite da esperti del campo di applicazione. Infatti i modelli strutturali sono stati sviluppati originariamente in settori applicativi socio-psicometrici ed impiegati a scopo confermativo.

I modelli ad equazioni strutturali, possono essere considerati come la ricomposizione di due istanze distinte:

- La necessità di misurare concetti astratti.

- La necessità di valutare la forza e la significatività dei legami causali fra i predetti concetti.

Per tale ragione i SEM sono costituiti da un modello strutturale *interno* che formalizzata le ipotesi di causalità e da tanti modelli di misura *esterni* quanti sono i concetti latenti interessati dalla rete di relazioni causali.

Le due componenti del modello si sono sviluppati per lungo tempo separatamente. Il primo tentativo di costruzione di un modello di misura per una variabile

[1]Structural Equation Models.

latente nasce in ambito psicometrico e viene fatto risalire agli studi di Spearman (1904), il quale, agli inizi del XX secolo, prende in considerazione uno dei più controversi concetti psicometrici: l'intelligenza umana e la sua misura.

L'intelligenza viene comunemente percepita come una qualità positiva, un vantaggio competitivo nelle relazioni umane, ma stabilire quanto un individuo sia intelligente e se lo sia più di un altro non è semplice, come invece può essere rilevare la capacità mnemonica, logica, cognitiva, relazionale e via dicendo. L'idea di Spearman è quella di combinare opportunamente i risultati ottenuti da test specifici su singole attitudini per ottenere una valutazione globale che sarà quella assegnata al relativo concetto latente o fattore. Da questo principio prende origine l'Analisi Fattoriale.

L'approccio fattoriale alla misurazione consiste nell'analisi di un blocco di variabili osservabili, altamente covarianti fra loro, per le quali si presuppone l'esistenza di un fattore comune di cui siano manifestazione. Il richiamo allo schema di misura riflessivo è evidente: il ricercatore, scartata la possibilità di stabilire relazioni causali fra variabili fortemente legate fra loro, cerca di eliminare quello che ritiene un legame spurio introducendo una causa comune.

Il passo successivo è rappresentato dall'Analisi Fattoriale Confermativa [Thurstone (1947)], che adotta un metodo completamente deduttivo, essendo l'analisi *guidata* da un modello teorico proposto in base a conoscenze a priori ed integrata da procedure di verifica delle ipotesi, ovvero da un apparato probabilistico che permetta di valutare la congruenza fra le conseguenze teoriche del modello ipotizzato e i dati empirici rilevati.

Parallelamente allo sviluppo delle tecniche di misura, vengono proposte nuove tecniche per il trattamento dei legami causali: nei primi decenni del '900, il genetista S.Wright (1921, 1934) introduce un modello statistico per reti di relazioni causali fra variabili, dando l'avvio alla *Path Analisys*[2]. La novità rispetto alla regressione multipla sta nel fatto che la forza dei legami causali viene stimata contemporaneamente per tutte le variabili, vengono superati i problemi di covariazione delle variabili esplicative e risulta possibile calcolare anche gli effetti indiretti.

A partire dagli anni '70 del XX secolo i modelli di misura e i modelli causali vengono ricomposti in un unico schema teorico tramite l'impiego di modelli ad equazioni strutturali, con l'obiettivo di stimare simultaneamente sia i parametri dei modelli di misura sia i parametri delle equazioni di regressione multipla associate alla struttura di relazioni causali.

All'unificazione del modello non corrisponde un'unità nei metodi di stima, che da subito si dividono in due correnti aventi differenti finalità.

[2] Analisi dei tragitti (*path*), dei percorsi delle relazioni causali.

Lo statistico svedese Jöreskog propone nel 1970 [Jöreskog (1970)] il prototipo di quelli che vengono chiamati metodi di stima *covariance based*: il metodo di stima della massima verosimiglianza per modelli ad equazioni strutturali (di seguito citato anche come SEM-ML[3] o LISREL[4] dal nome del software più comunemente utilizzato per la compuutazione[Jöreskog e Sörbom (1993)]). L'obiettivo del metodo è strettamente confermativo:

- Si ipotizza uno schema di relazioni causali e di misura (formulazione ipotesi).

- Si stima la matrice di varianza/covarianza in base allo schema causale adottato (deduzione delle conseguenze teoriche).

- Si valuta se la matrice stimata sia compatibile con quella osservata (verifica di ipotesi).

- Nel caso l'ipotesi non venga falsificata, la si assume come valida fino all'evidenza di prove contrarie.

- Nel caso i dati contraddicano l'ipotesi, si ritorna al primo punto e si formula una nuova ipotesi.

L'appellativo di *covariance based* deriva dal fatto che le stime dei parametri fondamentali del modello vengono effettuate con la finalità di ottimizzare la matrice di varianza/covarianza stimata, in modo che sia il più possibile prossima a quella osservata. Le varianti di SEM *covariance based* si differenziano principalmente per la funzione di discrepanza utilizzata per valutare tale prossimità.

I modelli *covariance based* vengono impiegati comunemente per testare ipotesi sulla validità delle reti di relazioni causali ipotizzate dalla ricerca specifica del campo d'applicazione dell'analisi. Per poter utilizzare gli strumenti della teoria delle probabilità è però necessario fare delle forti ipotesi iniziali sulla distribuzione delle variabili analizzate.

Allo scopo di ovviare ai problemi di fondo dei SEM *covariance based*, a metà degli anni '70, lo statistico ed econometrico svedese Herman Wold [Wold (1975b)] propone un metodo di stima completamente diverso che si prefigge un duplice obiettivo: effettuare il minor numero possibile di assunzioni a priori sul modello (da cui l'appellativo di *soft modeling* per contrapporlo ai metodi *hard* come i SEM *covariance based*) e stimare i parametri in modo da ottenere variabili latenti[5] il più possibile rappresentative del rispettivo blocco di misura e della struttura causale di cui fanno parte.

Il nome del metodo è Partial Least Squares Path Modeling (comunemente indicato con la sigla PLS-PM) in quanto si basa su tecniche di regressione parziale già applicate nella stima di regressioni parziali multiple con il nome di regressione

[3]Structural Equation Modeling by Maximum Likelihood estimation.

[4]LInear Structural RELationships

[5]Da cui l'appellativo di *component based* per questo e per tutti i metodi che hanno come obiettivo la costruzione di variabili latenti, o *componenti* con opportune proprietà.

PLS (Tenenhaus, 1998), ma da non confondere con queste ultime che non sono strutturate per la stima di relazioni causali simultanee.

Il PLS-PM può essere considerato come una generalizzazione dell'Analisi delle Componenti Principali, che ha come obbiettivo l'individuazione e costruzione di variabili latenti il più possibile differenti fra loro ed esplicative della maggior parte possibile di variabilità degli indicatori. Nella stima PLS-PM, le variabili latenti cercate sostituiscono l'obiettivo di ortogonalità con quello di migliore rappresentazione dello schema causale del modello interno.

Come tutti i metodi *component based*, i PLS-PM sono principalmente tecniche di analisi dei dati, quindi occupano la posizione meno deduttiva nell'ambito dei metodi di stima per modelli ad equazioni strutturali. Avendo un orientamento principalmente applicativo ed in particolare predittivo, producono automaticamente, al contrario dei SEM *covariance based*, i punteggi per le variabili latenti.

D'altro canto, nonostante il forte impegno della ricerca, a tutt'oggi non sembra esistere una ben identificata funzione obiettivo massimizzata dalla stima PLS-PM, per cui non è possibile individuare un criterio di ottimizzazione globale. Dato lo scarso ricorso ad ipotesi distributive, il metodo è privo di tecniche classiche di verifica di ipotesi e per la validazione delle stime è possibile far ricorso alle sole verifiche di adattamento ai dati, utilizzando tecniche di validazione incrociata, stime bootstrap e jackknife (Efron, 1982; Efron e Tibshirani, 1993).

Nonostante il diverso approccio alla verifica delle ipotesi fondamentali del modello, entrambi questi metodi di stima rientrano nell'ambito dell'impostazione frequentista alla statistica, infatti si basano sull'assunzione di ripetibilità dell'esperimento/osservazione e sull'assunzione di omogeneità delle ripetizioni, che garantisce una funzione di verosimiglianza con opportune proprietà.

Le differenze fra le due tecniche risiede nel fatto che i metodi *covariance based*, nella validazione dei parametri e degli indici del modello, sfruttano le ipotesi distributive sulla popolazione, mentre i metodi *component based* fanno ricorso a tecniche di ricampionamento.

Nel seguito del capitolo viene fornito un inquadramento formale generale per i SEM e vengono esposte più dettagliatamente le principali tecniche di stima per i parametri, ma in primo luogo vengono esplicitate le due principali componenti del modello ad equazioni strutturali, ossia la parte causale e la parte di misura.

Rispettando le convenzioni grafiche utilizzate nell'ambito dei modelli ad equazioni strutturali, vale quanto segue:

- Le variabili manifeste (gli indicatori) vengono comunemente racchiuse da un rettangolo ed indicate con lettere latine.

- Le variabili latenti (i concetti astratti) vengono inscritte in un ellisse ed indicate con lettere greche.

- Le relazioni di causalità fra due fenomeni vengono indicate con una freccia monodirezionale (il verso della quale specifica la direzione del rapporto di causalità).

- La semplice covariazione fra variabili viene indicata da una freccia bidirezionale, spesso tratteggiata.

2.1 Il modello di misurazione

Il fulcro intorno al quale ruotano i modelli di misura è il fenomeno *latente*, un costrutto teorico avente molteplici manifestazioni ma non direttamente osservabile. Sono innumerevoli i concetti che possono essere ricondotti a tale definizione ed una prima possibilità applicativa dei modelli di misura è rappresentata dai problemi di *precisione* della misura.

Sebbene le manifestazioni di molti concetti siano direttamente osservabili, spesso gli strumenti di misura a disposizione per registrarne le variazioni non sono sufficientemente precisi e come conseguenza si ottengono stime inaffidabili.

Una soluzione al problema consiste nel considerare il fenomeno osservabile "misura rilevata" come il risultato dell'interazione del fenomeno non osservabile "misura reale" e dell'errore di misurazione. Associando alla rilevazione una variabile $\boldsymbol{x}$ (detta *manifesta* in quanto osservabile) e allo stato reale della misura una variabile $\boldsymbol{\xi}$ (detta latente, in quanto nascosta agli occhi dell'osservatore), supponendo che l'errore abbia media zero e sia incorrelato con la misura reale, si ottiene la seguente relazione:

$$\boldsymbol{x} = \lambda\boldsymbol{\xi} + \boldsymbol{\delta} \tag{2.1}$$

dove λ è il coefficiente della relazione e $\boldsymbol{\delta}$ rappresenta l'errore di misura.

Volendo generalizzare il caso precedente, è possibile affermare che per alcuni fenomeni, solitamente molto complessi, non esiste affatto uno strumento di misura adeguato. Di essi è possibile rilevare solamente alcuni aspetti, ma non si riesce a fornire una valutazione (misura) globale.

In tali situazioni l'operazione di misura può avvenire solo per mezzo di indicatori manifesti in stretta relazione causale con la variabile non osservabile. Solitamente, se non ci sono forti evidenze empiriche contrarie, si assume che la relazione sia di tipo lineare, data la semplicità e flessibilità di quest'ultima. Per quanto riguarda il verso della causalità, come anticipato a pagina 19 i modelli di misura vengono classificati in tre schemi principali:

-Schema Riflessivo

-Schema Formativo

-Schema misto di dipendenza o MIMIC.

Schema riflessivo

Quando gli indicatori rilevano fenomeni configurabili come diretta conseguenza di un concetto latente retrostante, è possibile utilizzare lo schema di misura **riflessivo**. Un esempio in tal senso è il modello di precisione della misura formalizzato con l'equazione 2.1 e riportato in figura 2.1:

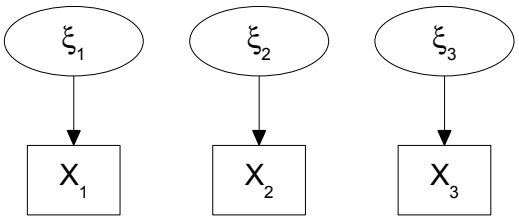

Figura 2.1: Schema riflessivo: ogni variabile manifesta $\boldsymbol{x}_i$ dipende dalla rispettiva variabile latente $\boldsymbol{\xi}_i$.

La rappresentazione algebrica di tale modello consta di tante regressioni lineari semplici quante sono le p variabili manifeste presenti e, nel caso di errori di misura a media zero e incorrelati con i concetti latenti considerati centrati nella propria media, si ottiene:

$$\begin{aligned} \boldsymbol{x}_1 &= \lambda_1 \boldsymbol{\xi}_1 + \boldsymbol{\delta}_1 \\ \boldsymbol{x}_2 &= \lambda_2 \boldsymbol{\xi}_2 + \boldsymbol{\delta}_2 \\ &\cdots \\ \boldsymbol{x}_p &= \lambda_p \boldsymbol{\xi}_p + \boldsymbol{\delta}_p. \end{aligned}$$

Utilizzando la notazione matriciale:

$$\boldsymbol{X} = \boldsymbol{\Xi}\ \boldsymbol{\Lambda}\ +\ \boldsymbol{\Delta}$$

dove $\boldsymbol{\Lambda}$ è la matrice diagonale associata al vettore dei parametri λ_i. I pesi λ_i dell'equazione di dipendenza delle variabili manifeste dalla latente vengono comunemente chiamati *loadings* per distinguerli dai pesi w_i della regressione della variabile latente sulle manifeste.

Quando la matrice $\boldsymbol{\Lambda}$ non è vincolata ad essere diagonale, la formalizzazione algebrica diventa:

$$\begin{aligned} \boldsymbol{x}_1 &= \lambda_{11} \boldsymbol{\xi}_1 + \lambda_{12} \boldsymbol{\xi}_2 + \cdots + \lambda_{1m} \boldsymbol{\xi}_m + \boldsymbol{\delta}_1 \\ \boldsymbol{x}_2 &= \lambda_{21} \boldsymbol{\xi}_1 + \lambda_{22} \boldsymbol{\xi}_2 + \cdots + \lambda_{2m} \boldsymbol{\xi}_m + \boldsymbol{\delta}_2 \\ &\cdots \\ \boldsymbol{x}_p &= \lambda_{p1} \boldsymbol{\xi}_1 + \lambda_{p2} \boldsymbol{\xi}_2 + \cdots + \lambda_{pm} \boldsymbol{\xi}_m + \boldsymbol{\delta}_p. \end{aligned}$$

dove le variabili ξ_i sono centrate nella propria media.

L'equazione matriciale resta:

$$\boldsymbol{X} = \boldsymbol{\Xi}\,\boldsymbol{\Lambda} \;+\; \boldsymbol{\Delta}, \tag{2.2}$$

questa volta però, ogni variabile osservata può essere indicatore di più concetti latenti (figura 2.2: caso A) e viceversa, ogni latente può avere più manifestazioni (figura 2.2: caso B):

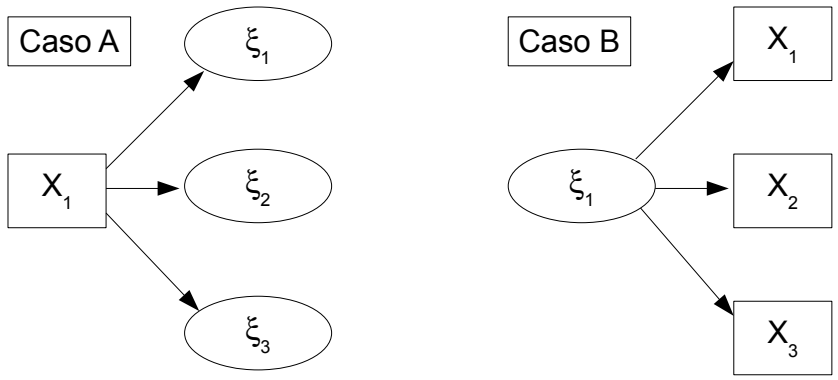

Figura 2.2: Schema riflessivo multiplo: generalizzazioni.

Mentre la configurazione del Caso B è quella classica della schema riflessivo, il Caso A pone seri problemi teorici poiché denota confusione concettuale e spesso nasconde una struttura causale globale non ben specificata.

Il modello riflessivo richiede coerenza interna fra gli indicatori di uno stesso blocco, in ragione del fatto che dipendono tutti dallo stesso fenomeno latente. Non essendo possibile adottare un criterio globale di coerenza, si fa ricorso ad indicatori di unidimensionalità quali l'Analisi delle Componenti Principali, l'indice α di Cronbach e l'indice ρ di Dillon-Goldstein.

-L'*Analisi delle Componenti Principali* è uno strumento statistico che permette di stabilire se è possibile ridurre la dimensionalità dei dati a disposizione senza eccessiva perdita di informazioni. Per poter utilizzare lo schema riflessivo, il blocco degli indicatori deve essere monodimensionale, essendo le variabili osservabili derivazione di un'unica latente. Se l'ACP rileva la presenza di un solo autovalore significativo (maggiore di 1) allora ci sono indizi di unidimensionalità, in caso contrario c'è il sospetto che uno stesso blocco misuri diversi concetti astratti.

-L'indice α di *Cronbach* può essere utilizzato per misurare l'unidimensionalità del blocco di variabili $\boldsymbol{x}_i\,(i = 1, \ldots, p)$, positivamente correlate fra loro. Nel caso di variabili standardizzate l'indice vale:

$$\alpha = \frac{\sum_{i \neq j} cor\,(\boldsymbol{x}_i, \boldsymbol{x}_j)}{p + \sum_{i \neq j} cor\,(\boldsymbol{x}_i, \boldsymbol{x}_j)} \times \frac{1}{(p-1)} \tag{2.3}$$

- Anche l'indice ρ di *Dillon-Goldstein* quantifica l'unidimensionalità all'interno di un blocco di variabili positivamente correlate fra loro, ma in aggiunta tiene in considerazione anche il modello di misura. Nel caso siano standardizzate sia le variabili manifeste $\boldsymbol{x}_i\,(i = 1, \ldots, P_q)$ che la latente $\boldsymbol{\xi}_q$ a cui afferiscono, per $\boldsymbol{X}_q = \boldsymbol{\xi}_q\boldsymbol{\lambda}' + \boldsymbol{\Delta}$, con $\lambda' = \left[\lambda_1, \lambda_2, \ldots, \lambda_{P_q}\right]$ l'indice di unidimensionalità del q-esimo blocco, ρ_q viene calcolato come:

$$\rho_q = \frac{\left(\sum_i^{P_q} \lambda_i\right)^2}{\left(\sum_i^{P_q} \lambda_i\right)^2 + \sum_i^{P_q} (1 - \lambda_i^2)} \tag{2.4}$$

Per entrambi gli indici α e ρ, un valore almeno pari a 0.7 segnala una sufficiente coerenza interna delle variabili, tenendo in considerazione però che in accordo con Chin (1998), le informazioni fornite dall'indice ρ sono da considerarsi migliori.

Schema formativo

Il modello **formativo** si adatta a situazioni in cui ciascun indicatore contribuisce autonomamente alla determinazione del concetto latente retrostante il blocco di misura (figura 2.3).

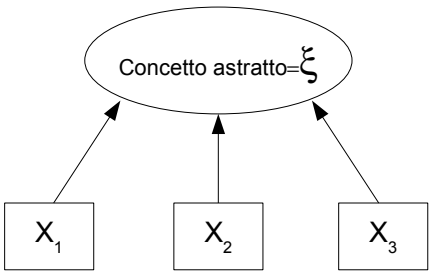

Figura 2.3: Schema formativo: la variabile latente $\boldsymbol{\xi}$ dipende dalle variabili manifeste $\boldsymbol{x}_i$.

Ipotizzando ancora relazioni causali lineari variabili centrate nella propria media ed errori a media zero e incorrelati con gli indicatori, è possibile rappresentare algebricamente lo schema formativo tramite un modello di regressione multipla:

$$\boldsymbol{\xi} = w_1\boldsymbol{x}_1 + \cdots + w_p\boldsymbol{x}_p + \boldsymbol{\delta}$$

che in forma matriciale si può riassumere come:

$$\boldsymbol{\xi} = \boldsymbol{X}\ \boldsymbol{w}\ +\ \boldsymbol{\delta} \tag{2.5}$$

dove $\boldsymbol{\xi}$ è la variabile latente, $\boldsymbol{X}$ è la matrice delle p variabili manifeste centrate nella propria media, $\boldsymbol{w}$ è il vettore dei coefficienti ed $\boldsymbol{\delta}$ è il vettore dei residui, la parte di variabile latente non spiegata dalle variabili osservate.

Il vettore $\boldsymbol{w}$ viene definito comunemente vettore dei pesi esterni[6] e rappresenta l'entità d'influenza di ciascuna variabile manifesta sulla rispettiva latente.

Utilizzando la notazione matriciale è possibile formalizzare anche il caso di più variabili latenti e i rispettivi blocchi di misura:

$$\mathbf{\Xi} = \mathbf{W}\ \mathbf{X}\ +\ \mathbf{\Delta} \tag{2.6}$$

dove $\mathbf{\Xi}$ è la matrice delle variabili latenti, $\mathbf{\Delta}$ quella dei residui e $\mathbf{W}$ rappresenta la matrice dei pesi esterni.

Esattamente all'opposto dello schema riflessivo, il modello formativo si caratterizza per la presenza di indicatori disomogenei che congiuntamente determinano un unico concetto più generale, per tale ragione è opportuno avere a disposizione variabili osservabili diversificate, che catturino il maggior numero possibile di aspetti differenti del fenomeno latente. La presenza di forti correlazioni tra le variabili del blocco esclude la possibilità di utilizzo di un modello esclusivamente formativo.

Schema MIMIC

L'ultimo modello citato si presenta come una fusione dei precedenti in quanto prevede che alcune delle variabili manifeste siano formative ed altre riflessive del fenomeno latente (figura 2.4).

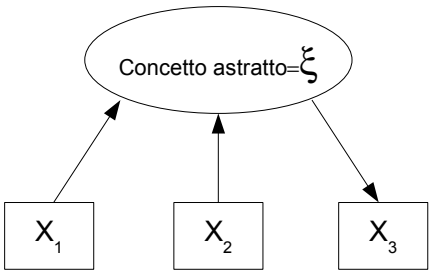

Figura 2.4: Schema MIMIC: sono presenti entrambi i tipi di relazione tra le variabili.

L'obiettivo di minimizzare gli errori dei residui in questo caso viene raggiunto tramite due tipi diversi di regressione, a seconda che si tratti della parte riflessiva o formativa del modello.

[6] In inglese *weights*.

Lo schema MIMIC permette di ottenere la stima del modello di misura ma pone altresì, seri problemi teorici in quanto necessita di un trattamento separato per i diversi indicatori, la qual cosa suggerisce la possibile presenza di errori nella scelta degli indicatori.

2.2 Il modello causale

La difficoltà dei tentativi di modellare reti causali sta nella necessità di considerare alcune variabili contemporaneamente nel ruolo di causa ed effetto.

La regressione lineare multipla assolve il compito di modellare fenomeni di causazione diretta, ma è noto che fornisce risultati distorti nel caso di variabili interdipendenti. In una situazione del genere è necessario fornire una formulazione matematica più complessa per integrare nella stima la causalità indiretta.

Utilizzando un modello tipico della *Path Analysis* (figura 2.5), viene introdotta la formalizzazione impiegata nei SEM per descrivere le relazioni causali lineari.

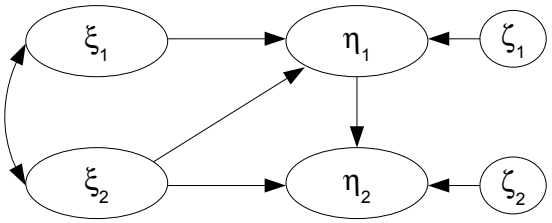

Figura 2.5: Esempio di *Path diagram*.

Vengono comunemente indicate con $\boldsymbol{\xi}_j$ le variabili esplicative esogene e con $\boldsymbol{\eta}_i$ le variabili endogene, esplicative o meno, in questa sede tutte considerate centrate nella propria media senza che ci sia perdita di generalità. Riguardo ai coefficienti, i β_i sono associati alle variabili endogene e i γ_j alle esogene. Gli ζ_i sono gli errori dell'equazione, i quali non rappresentano errori di misura ma tengono conto dell'influenza di variabili non espressamente inserite nello schema causale.

Il modello è ben specificato se le variabili escluse dallo schema causale e ricomprese negli errori esercitano una trascurabile influenza complessiva e se gli errori hanno media zero e sono incorrelati con le esplicative, altrimenti potrebbe configurarsi il caso di fenomeni rilevanti erroneamente esclusi dal modello o che la correlazione fra gli errori sia dovuta all'azione di una causa comune esterna. Entrambe le condizioni suggeriscono una revisione della struttura causale.

Esplicitando i legami causali del diagramma si hanno:

- Azioni *dirette*:

$\boldsymbol{\xi}_1 \rightarrow \boldsymbol{\eta}_1$, $\boldsymbol{\xi}_2 \rightarrow \boldsymbol{\eta}_1$, $\boldsymbol{\xi}_2 \rightarrow \boldsymbol{\eta}_2$, $\boldsymbol{\eta}_2 \rightarrow \boldsymbol{\eta}_2$.

- Azioni *indirette*:

$\boldsymbol{\xi}_1 \rightarrow\rightarrow \boldsymbol{\eta}_2$ e $\boldsymbol{\xi}_2 \rightarrow\rightarrow \boldsymbol{\eta}_2$

Inoltre $\boldsymbol{\xi}_2$ esercita un'influenza *mista* (sia diretta che mediata da $\boldsymbol{\xi}_1$) su $\boldsymbol{\eta}_2$, mentre $\boldsymbol{\xi}_1$ e $\boldsymbol{\xi}_2$ non sono in relazione causale ma presentano una correlazione.

Posto di aver considerato variabili centrate nella propria media e considerando per semplicità di notazione le relazioni riferite ad ogni singola unità del modello, vale che:

$$\begin{aligned}\eta_1 &= 0\eta_1 + 0\eta_2 + \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1\\ \eta_2 &= \beta_{21}\eta_1 + 0\eta_2 + 0\xi_1 + \gamma_{22}\xi_2 + \zeta_2\end{aligned}$$

ovvero, in forma matriciale:

$$\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \begin{bmatrix}0 & 0\\ \beta_{21} & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} + \begin{bmatrix}\gamma_{11} & \gamma_{12}\\ 0 & \gamma_{22}\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix} + \begin{bmatrix}\zeta_1\\ \zeta_2\end{bmatrix}$$

Nel caso generale di n variabili esogene ξ_i ed m variabili endogene η_j si ha che:

$$\begin{bmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_m\end{bmatrix} = \begin{bmatrix}0 & \beta_{12} & \cdots & \beta_{1m}\\ \beta_{21} & 0 & \cdots & \beta_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ \beta_{m1} & \cdots & \cdots & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_m\end{bmatrix} + \begin{bmatrix}\gamma_{11} & \cdots & \cdots & \gamma_{1n}\\ \vdots & \ddots & & \vdots\\ \vdots & & \ddots & \vdots\\ \gamma_{m1} & \cdots & \cdots & \gamma_{mn}\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\\ \vdots\\ \xi_n\end{bmatrix} + \begin{bmatrix}\zeta_1\\ \zeta_2\\ \vdots\\ \zeta_m\end{bmatrix}$$

ossia:

$$\boldsymbol{\eta} = \mathbf{B}\,\boldsymbol{\eta} \;+\; \boldsymbol{\Gamma}\,\boldsymbol{\xi} \;+\; \boldsymbol{\zeta} \tag{2.7}$$

Si noti che la matrice quadrata $\mathbf{B}$ ha la diagonale principale composta di soli 0, perché non è ammessa retroazione di una variabile su se stessa.

Per specificare completamente il modello causale è necessario definire anche alcune matrici di varianza/covarianza (o di correlazione nel caso di variabili standardizzate):

- La matrice $\boldsymbol{\Phi}$ di varianza/covarianza fra le variabili esogene ξ_j, la quale permette di tenere in considerazione nel modello eventuali relazioni fra variabili esogene non altrimenti specificate.

- La matrice $\boldsymbol{\Psi}$ di varianza/covarianza fra gli errori ζ_i delle variabili endogene η_i, la quale consente di considerare l'influenza di variabili escluse dal modello che però esercitano un'influenza trascurabile sul fenomeno.

In quest'ultimo caso, ad esempio, se esiste una variabile che agisce su η_1 ed η_2 ma non è stata inclusa nel modello, dalla stima risulterà una relazione spuria tra le due variabili, che però scompare se si include nel modello una correlazione tra gli errori ζ_1 e ζ_2. Dire che gli errori ζ_1 ed ζ_2 sono correlati equivale a dire che esiste una variabile non inclusa nel modello che agisce su η_1 ed η_2.

2.3 Modelli ad Equazioni Strutturali: schema completo e tecniche di stima

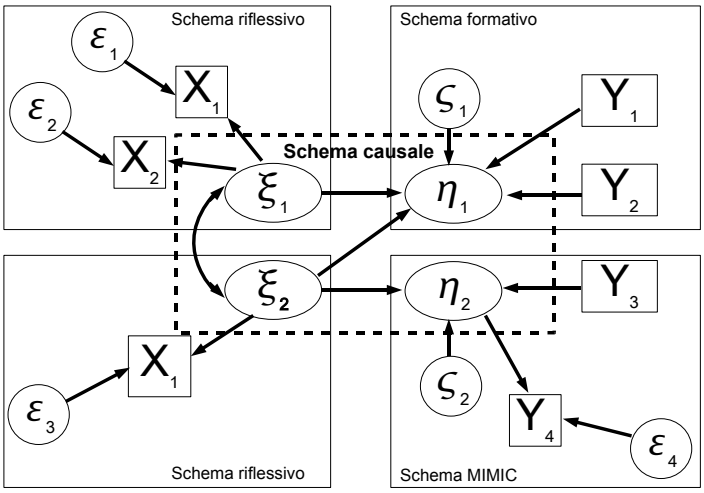

Figura 2.6: Esempio di *path diagram* di un modello ad equazioni strutturali.

Nei precedenti paragrafi è stata presentata ciascuna delle due parti di cui si compone il modello ad equazioni strutturali. In figura 2.6 viene mostrato un ipotetico schema completo che si ottiene dell'unione della parte causale del modello e degli schemi di misura.

Prima di procedere ad una formalizzazione algebrica di tale schema è bene ricordare che esistono due approcci alla stima dei parametri, che essendosi sviluppati separatamente, hanno utilizzato una notazione algebrica leggermente differente.

Nel lavoro si preferisce presentare ogni metodo con la propria notazione per mantenerla il più possibile uniforme alla letteratura, pertanto l'operazione di formalizzazione viene posticipata ai paragrafi relativi ai metodi di stima. In questo paragrafo vengo presentate le due macro aree metodologiche e nel seguito vengono esposte le principali tecniche.

Il punto su cui gli studiosi si dividono nell'applicazione dei modelli ad equazioni strutturali riguarda principalmente i metodi di stima dei parametri, i quali possono essere riassunti (usando i termini inglesi che più comunemente vengono impiegati in letteratura) in *Covariance based* e *Component based*.

Come per tutto il contesto dell'analisi multivariata, esistono due correnti di pensiero principali (rimaste per lungo tempo inconciliabili) che affrontano il problema del metodo di stima dei parametri incogniti dei Modelli ad equazioni Strutturali secondo punti di vista anche molto differenti.

Nell'approccio cosiddetto "anglosassone" si fa ampio ricorso ad ipotesi probabilistiche sulla distribuzione dei fenomeni in modo da poter impiegare il più possibile i potenti mezzi dell'inferenza statistica classica al fine di valutare l'attendibilità delle soluzioni ottenute.

L'indirizzo è chiaramente di puro deduttivismo, tant'è che la procedura di falsificazione, nel caso dei SEM è limpida: dato il modello teorico, la matrice di varianza/covarianza dei dati *dovrebbe* avere una certa configurazione (deducibile dal modello stesso). Si stimano i parametri in modo tale che le informazioni contenute nella matrice di varianza/covarianza *osservata* siano preservate, si calcola la matrice *stimata* e si valuta se è sufficientemente prossima a quella *teorica*: se le fluttuazioni possono essere imputate al caso si adotta la teoria fino a che nuovi dati non la mettano alla prova. Se si ritiene invece che la differenza fra le due non sia imputabile al caso, la teoria risulta falsificata e si deve procedere con una nuova formulazione delle ipotesi.

Poiché il fulcro principale di tutto il processo è la matrice di varianze/covarianze, tutti i metodi afferenti a tale approccio si dicono *covariance based* e le varianti della tecnica di stima riguardano la funzione di discrepanza, ovvero lo strumento per valutare la differenza fra matrice stimata e matrice teorica.

L'approccio "francese", d'altronde, è meno restrittivo, non richiedendo che sia conosciuta a priori la distribuzione iniziale dei fenomeni. La caratteristica di questi metodi, nel caso SEM, è la ricerca di stime che meglio rappresentino i costrutti latenti ipotizzati nel modello. Il nome di *component based* è dovuto appunto al fatto che l'obiettivo dell'analisi è ottenere punteggi delle variabili latenti con buone proprietà.

Le tecniche utilizzate hanno come scopo principale la costruzione di variabili latenti fortemente correlate fra di loro (coerentemente con lo schema del modello causale che le ipotizza) e rappresentative del gruppo di variabili manifeste ad esse associato.

L'indirizzo è chiaramente quello dell'analisi dei dati con finalità esplorative e la componente induttiva è molto più spiccata sebbene non predominante. Facendo uso di un minor numero di ipotesi iniziali, i metodi *component based* non possono far ricorso alle tecniche inferenziali classiche, quindi devono adottare metodi come validazione incrociata e tecniche di ricampionamento ma in compenso si avvalgono di algortmi numerici che trovano soluzioni a diversi problemi lasciati irrisolti dalle tecniche *covariance based*.

Nel seguito del capitolo saranno presentate le principali tecniche di stima dei parametri dei Modelli ad Equazioni Strutturali:

- Structural Equations Modeling by Maximum Likelihood (SEM-ML o LISREL) e cenni ai restanti metodi *covariance based* (paragrafo 2.4).

- Partial Least Squares Path Modeling (PLS-PM), rappresentativo dei metodi

component based (paragrafo 2.5).

2.4 Modelli ad Equazioni Strutturali: SEM-ML

2.4.1 La stima *Covariance Based*

L'obiettivo dei metodi *Covariance Based* è quello di ottenere la stima dei parametri liberi del modello preservando le informazioni contenute nella matrice di varianza/covarianza osservata. La soluzione proposta per questo problema è una stima che produca una matrice di varianza/covarianza riprodotta dal modello il più vicina possibile a quella osservata.

I metodi *covariance based* possono essere considerati una generalizzazione dell'Analisi Fattoriale Confermativa [Thurstone (1947)], la quale di conseguenza è un caso particolare di SEM in cui non si ipotizzano legami causali tra le variabili latenti.

Se si considera che l'analisi fattoriale confermativa è fondamentalmente un sistema di modelli di misura in cui vengono fatte delle ipotesi sui possibili costrutti latenti (fattori comuni) retrostanti le variabili manifeste, ma nessuna ipotesi sulle relazioni causali fra gli stessi, allora è possibile vedere i metodi *covariance based* come una CFA in presenza di tavole multiple legate fra loro.

Nel 1970 Jöreskog introduce e sviluppa un metodo basato sulla massima verosimiglianza per stimare i parametri di un modello che prevede, oltre ai modelli di misura, legami causali fra variabili latenti. Tale metodo prende il nome dal software LISREL®, sviluppato dallo stesso Jöreskog, che è stato il primo e per molti anni è rimasto l'unico strumento utile per effettuare l'enorme mole di calcoli necessaria per le stime ML.

Il nome LISREL è stato successivamente associato all'intera famiglia dei metodi SEM *covariance based*, che negli anni si è ingrandita per comprendere diverse tecniche di stima come ad esempio i minimi quadrati non pesati detti ULS (Unweighted Least Squares), i minimi quadrati generalizzati detti anche GLS (Generalized Least Squares), e i metodi a distribuzione asisntoticamente libera detta ADF (Asymtotically Distribution Free).

Ad ogni metodo di stima corrisponde una differente funzione (detta di discrepanza) impiegata per valutare la distanza fra la matrice di varianza/covarianza osservata e quella stimata.

In questo paragrafo viene riproposta la formulazione algebrica delle equazioni dei modelli SEM nell'ambito della stima *covariance based* e viene ricavata la matrice di varianza/covarianza predetta in relazione a esse. Vengono esplicitate le ipotesi fondamentali sui parametri, specifiche di tali modelli e per completezza

vengono presentate le diverse funzioni di discrepanza utilizzate nell'algoritmo di calcolo della stima nonché alcuni indici per la valutazione della bontà dei modelli.

2.4.2 La notazione algebrica

I Modelli ad Equazioni Strutturali nascono dalla fusione dei modelli causali e dei modelli di misura, quindi fanno uso di una notazione piuttosto complessa. Allo scopo di semplificare l'esposizione, la formalizzazione è quella relativa alla singola unità della matrice dei dati, senza che ci sia perdita di generalità. L'unico schema di misura ammesso nella stima *covariance based* è quello riflessivo, quindi solo ad esso si fa riferimento nella notazione.

Ai fini della stima *covariance based* il modello viene suddiviso in una parte esogena (relativa alle variabili latenti esplicative ed ai loro schemi di misura) e una parte endogena (riguardante tutte le variabili latenti che intervengono come dipendenti in una qualsiasi delle relazioni causali e i relativi modelli di misura). La denominazione dei parametri segue questa distinzione, per cui:

$\boldsymbol{\xi}$ rappresenta il vettore delle variabili latenti ***esogene***, ossia tali da assumere sempre e solo il ruolo di variabili esplicative all'interno del modello causale.

$\boldsymbol{\eta}$ rappresenta il vettore delle variabili latenti ***endogene***, le quali, all'interno del modello causale, ricoprono almeno in un'equazione il ruolo di variabile dipendente, ma che possono rappresentare, in altre equazioni, anche il ruolo di variabile esplicativa.

$\boldsymbol{x}$ viene impiegato per il vettore delle variabili manifeste indicatrici di variabili latenti *esogene* $\boldsymbol{\xi}$.

$\boldsymbol{y}$ viene impiegato per il vettore delle variabili manifeste indicatrici di variabili latenti *endogene* $\boldsymbol{\eta}$.

Per quanto riguarda gli errori:

$\boldsymbol{\zeta}$ è l'errore associato alla variabile latente endogena $\boldsymbol{\eta}$ e rappresenta l'errore dell'equazione: tutti gli effetti esercitati sulle $\boldsymbol{\eta}$ dalle variabili tralasciate dal modello.

$\boldsymbol{\delta}$ è il vettore degli errori di misura associato alle variabili manifeste $\boldsymbol{x}$.

$\boldsymbol{\epsilon}$ è il vettore degli errori di misura associato alle variabili manifeste $\boldsymbol{y}$.

Per rappresentare la forza delle relazioni causali espresse nel modello vengono utilizzati quattro tipi di coefficiente:

$\boldsymbol{\Lambda}_x$ è la matrice dei coefficienti λ_{ij}^x relativi alla regressione delle $\boldsymbol{x}$ sulle $\boldsymbol{\xi}$.

$\boldsymbol{\Lambda}_y$ è la matrice dei coefficienti λ_{ij}^y relativi alla regressione delle $\boldsymbol{y}$ sulle $\boldsymbol{\eta}$.

$\mathbf{B}$ è la matrice dei coefficienti β_{ij} relativi alle relazioni fra le variabili $\boldsymbol{\eta}$.

$\boldsymbol{\Gamma}$ è la matrice dei coefficienti γ_{ij} relativi alle relazioni fra la variabile $\boldsymbol{\eta}$ e la variabile $\boldsymbol{\xi}$.

Infine le covariazioni fra variabili sono rappresentate da:

$\boldsymbol{\Phi}$ la matrice delle varianze/covarianze delle variabili latenti *esogene* $\boldsymbol{\xi}$.

$\boldsymbol{\Psi}$ la matrice delle varianze/covarianze fra gli errori $\boldsymbol{\zeta}$ delle variabili latenti *endogene* $\boldsymbol{\eta}$.

$\boldsymbol{\Theta}^{\delta}$ la matrice delle varianze/covarianze fra gli errori $\boldsymbol{\delta}$ delle variabili manifeste $\boldsymbol{x}$.

$\boldsymbol{\Theta}^{\epsilon}$ la matrice delle varianze/covarianze fra gli errori $\boldsymbol{\epsilon}$ delle variabili manifeste $\boldsymbol{y}$.

2.4.3 Le equazioni fondamentali dei SEM *covariance based*

Nei paragrafi 2.2 e 2.1 sono state già introdotte le equazioni fondamentali che qui di seguito vengono riprese e adattate ai modelli ad equazioni strutturali.

Per la parte *causale* del modello vale l'equazione 2.7:

$$\boldsymbol{\eta} = \mathbf{B}\,\boldsymbol{\eta} \;+\; \boldsymbol{\Gamma}\,\boldsymbol{\xi} \;+\; \boldsymbol{\zeta}$$

Per la parte di *misura* del modello è ammesso solo lo schema riflessivo e l'equazione 2.2 viene riformulata sia per le variabili endogene del modello:

$$\boldsymbol{y} = \boldsymbol{\Lambda}_{\boldsymbol{y}}\,\boldsymbol{\eta} \;+\; \boldsymbol{\epsilon} \tag{2.8}$$

che per quelle esogene:

$$\boldsymbol{x} = \boldsymbol{\Lambda}_{\boldsymbol{x}}\,\boldsymbol{\xi} \;+\; \boldsymbol{\delta} \tag{2.9}$$

Perché il modello sia completo devono essere considerate anche le matrici che rappresentano legami tra variabili non esplicitamente inseriti nel modello causale, ovvero le matrici $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}_{\epsilon}$ e $\boldsymbol{\Theta}_{\delta}$.

Da quanto detto si desume che per una completa specificazione del modello si rende necessario stimare le quattro matrici di varianza/covarianza sopra indicate e le matrici dei coefficienti $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}_{\boldsymbol{y}}$ e $\boldsymbol{\Lambda}_{\boldsymbol{x}}$

Da tali matrici è possibile ricavare algebricamente il valore della matrice di varianza/covarianza associata al modello che d'ora in avanti sarà indicata con $\boldsymbol{\Sigma}$.

2.4.4 Le ipotesi del modello

I modelli ad equazioni strutturali per essere stimati con i metodi ML devono sottostare ad alcune ipotesi specifiche. Altre alla classica ipotesi di multinormalità degli errori, si presuppone che:

- Le variabili sono centrate (misurate in termini di scarti dalla propria media):

$$E(\boldsymbol{\eta}) = E(\boldsymbol{\zeta}) = 0 \tag{2.10}$$
$$E(\boldsymbol{\xi}) = 0 \tag{2.11}$$
$$E(\boldsymbol{y}) = E(\boldsymbol{\epsilon}) = 0 \tag{2.12}$$
$$E(\boldsymbol{x}) = E(\boldsymbol{\delta}) = 0 \tag{2.13}$$

- Le variabili dipendenti non sono correlate agli errori, in ogni singola equazione:

$$E(\boldsymbol{\xi}\boldsymbol{\zeta}') = 0 \tag{2.14}$$
$$E(\boldsymbol{\eta}\boldsymbol{\epsilon}') = 0 \tag{2.15}$$
$$E(\boldsymbol{\xi}\boldsymbol{\delta}') = 0 \tag{2.16}$$

  e fra le equazioni:

$$E(\boldsymbol{\eta}\boldsymbol{\delta}') = 0 \tag{2.17}$$
$$E(\boldsymbol{\xi}\boldsymbol{\epsilon}') = 0 \tag{2.18}$$

- Gli errori relativi ad equazioni diverse sono tra loro incorrelati:

$$E(\boldsymbol{\zeta}\boldsymbol{\epsilon}') = 0 \tag{2.19}$$
$$E(\boldsymbol{\zeta}\boldsymbol{\delta}') = 0 \tag{2.20}$$
$$E(\boldsymbol{\epsilon}\boldsymbol{\delta}') = 0 \tag{2.21}$$

- La matrice $\mathbf{B}$ è non singolare, non sono ammesse variabili latenti endogene ridondanti, e non prevede relazioni reciproche o cicliche (per esempio, $\beta_{ii} = 0$ e se $\beta_{ij} \neq 0$, allora $\beta_{ji} = 0$).

Alle ipotesi più generali del modello si aggiungono altri due ordini di ipotesi complementari che riguardano l'adeguatezza dello schema riflessivo ai dati e l'identificabilità del modello nella sua interezza.

É già stato sottolineato che il modello di misura riflessivo è l'unico che si adatta agli algoritmi di calcolo dei SEM *covariance based*, ma non sempre è aderente ai dati in posso del ricercatore. Qui si intende ribadire la necessità di verificare l'ammissibilità di tale assunzione per mezzo degli indicatori presentati nel paragrafo 2.1: Analisi Componenti Principali, α di Cronbach e ρ di Dillon-Goldstein).

Di seguito viene invece presentato il problema dell'identificazione dei parametri del modello.

L'identificabilità del modello

Presentando la questione solo in linea molto generale, il problema dell'identificabilità del modello risiede nella necessità di trovare una soluzione unica per la stima dei parametri e considera sotto quali ipotesi ciò avviene.

Nel caso specifico dei modelli ad equazioni strutturali, la condizione a cui si deve necessariamente sottostare è che il numero dei parametri liberi, ossia da stimare, non sia superiore al numero di informazioni fornite dalla matrice si varianza/covarianza osservata, rappresentate dalle singole varianze e covarianze fra le variabili manifeste, non duplicate. Infatti il numero di tali informazioni viene calcolato come:

$$\frac{P\,(P+1)}{2}$$

dove P è il numero delle variabili manifeste. Essendo $\mathbf{\Sigma}$ una matrice simmetrica, è ovvio che le informazioni in essa contenute non sono P^2 ma circa la metà (essendo i valori fuori dalla diagonale principale uguali a coppie).

Introducendo il concetto di *gradi di libertà* del modello, denominati *df* (*degrees of freedom*) e definiti come:

$$df = \frac{P\,(P+1)}{2} - t \tag{2.22}$$

dove per t intendiamo il numero di parametri liberi dello stesso, possiamo riassumere la condizione necessaria per l'identificabilità come:

$$df \geq 0$$

I modelli con un valore negativo di *df* non sono identificabili e, sebbene gli algoritmi di calcolo producono ugualmente delle soluzioni, esse conducono spesso a conseguenze inaccettabili, quali varianze negative e correlazioni maggiori di 1.

Il caso particolare di $df = 0$ viene indicato come modello *saturo*: il modello teorico si adatta perfettamente ai dati essendo utilizzate per l'operazione di stima tutte le informazioni disponibili. I parametri così ottenuti permettono di riprodurre perfettamente la matrice di varianza/covarianza, ma non consentono l'uso degli strumenti di misura della bontà di adattamento ai dati.

Per due ordini di motivi i modelli saturi non sono interessanti. In primo luogo si adattano perfettamente al campione estratto, il quale per definizione non è che una rappresentazione parziale dell'intera popolazione, soffrono quindi di specificità.

In secondo luogo, è vero che un modello complicato si adatta meglio ai dati, ma al contempo pone problemi di interpretazione e soprattutto di falsificabilità[7]:

[7] Vedi paragrafo 1.2.3.

un modello facilmente falsificabile che superi i test di adattamento ha maggiore credibilità di un modello perfettamente aderente al campione ma non testabile.

É auspicabile avere un numero di *df* tanto più elevato quanto la teoria retrostante il modello lo consente, e ciò in definitiva si risolve nell'imposizione di vincoli sui parametri, i quali posso essere fissati a priori pari ad un certo valore o vincolati ad essere in relazioni specifiche con altri parametri liberi.

Sebbene la condizione 2.4.4 sia necessaria essa non è sufficiente, come viene più ampiamente giustificato in Bollen (1989), quindi un numero positivo di gradi di libertà non garantisce l'identificabilità del modello.

2.4.5 La matrice di varianza/covarianza predetta e la sua derivazione dai parametri del modello

Poiché l'obiettivo dei metodi *covariance based* è quelllo di ottenere una matrice di varianza/covarianza predetta $\boldsymbol{\Sigma}$ con opportune proprietà, è necessario stimare i parametri del modello mettendoli in relazione con tale matrice. Di seguito vengono esposti i passaggi algebrici.

$\boldsymbol{\Sigma}$ è scomponibile in quattro sottomatrici corrispondenti alle due matrici di varianza/covarianza delle variabili manifeste $\boldsymbol{x}$ e $\boldsymbol{y}$ (rispettivamente $\boldsymbol{\Sigma}_{xx}$ e $\boldsymbol{\Sigma}_{yy}$) e alla matrice $\boldsymbol{\Sigma}_{xy}$ di covarianza tra le variabili manifeste $\boldsymbol{x}$ e $\boldsymbol{y}$ (e di conseguenza anche alla trasposta di quest'ultima $\boldsymbol{\Sigma}_{yx}$), ottenendo:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}$$

Ognuna delle matrici $\boldsymbol{\Sigma}_{ij}$ può essere ricavata dei parametri del modello ad equazioni strutturali.

Scomposizione della matrice $\boldsymbol{\Sigma}_{xx}$

Essendo $\boldsymbol{\Sigma}_{xx}$ la matrice di varianza/covarianza delle variabili $\boldsymbol{x}$ vale che:

$$\boldsymbol{\Sigma}_{xx} = E\left(\boldsymbol{x}\boldsymbol{x}'\right)$$

e valendo l'equazione 2.9

$$\boldsymbol{x} = \boldsymbol{\Lambda}_{\boldsymbol{x}}\, \boldsymbol{\xi} \; + \; \boldsymbol{\delta}$$

si può scrivere:

$$\begin{aligned}
\boldsymbol{\Sigma}_{xx} &= E\left[\left(\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\xi}+\boldsymbol{\delta}\right)\left(\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\xi}+\boldsymbol{\delta}\right)'\right] \\
&= E\left[\left(\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\xi}+\boldsymbol{\delta}\right)\left(\boldsymbol{\xi}'\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\boldsymbol{\delta}'\right)\right] \\
&= E\left[\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\xi}\boldsymbol{\xi}'\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\boldsymbol{\delta}\boldsymbol{\xi}'\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\xi}\boldsymbol{\delta}'+\boldsymbol{\delta}\boldsymbol{\delta}'\right] \\
&= \boldsymbol{\Lambda}_{\boldsymbol{x}}\left[E\left(\boldsymbol{\xi}\boldsymbol{\xi}'\right)\right]\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\left[E\left(\boldsymbol{\delta}\boldsymbol{\xi}'\right)\right]\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\boldsymbol{\Lambda}_{\boldsymbol{x}}\left[E\left(\boldsymbol{\xi}\boldsymbol{\delta}'\right)\right]+E\left(\boldsymbol{\delta}\boldsymbol{\delta}'\right) \\
&= \boldsymbol{\Lambda}_{\boldsymbol{x}}\left[\boldsymbol{\Phi}\right]\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\left[0\right]\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\boldsymbol{\Lambda}_{\boldsymbol{x}}\left[0\right]+\boldsymbol{\Theta}_{\delta}
\end{aligned}$$

essendo per definizione $E\left(\boldsymbol{\xi}\boldsymbol{\xi}'\right)=\boldsymbol{\Phi}$, $E\left(\boldsymbol{\delta}\boldsymbol{\delta}'\right)=\boldsymbol{\Theta}_{\delta}$ e per l'ipotesi 2.16 del modello SEM *covariance based* $E\left(\boldsymbol{\delta}\boldsymbol{\xi}'\right)=E\left(\boldsymbol{\xi}\boldsymbol{\delta}'\right)=0$, si ricava:

$$\boldsymbol{\Sigma}_{xx}=\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\Phi}\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\boldsymbol{\Theta}_{\delta} \tag{2.23}$$

Scomposizione della matrice $\boldsymbol{\Sigma}_{yy}$

Essendo $\boldsymbol{\Sigma}_{yy}$ la matrice di varianza/covarianza delle variabili $\boldsymbol{y}$ vale che:

$$\boldsymbol{\Sigma}_{yy}=E\left(\boldsymbol{y}\boldsymbol{y}'\right)$$

e valendo l'equazione 2.8

$$\boldsymbol{y}=\boldsymbol{\Lambda}_{\boldsymbol{y}}\,\boldsymbol{\eta}\;+\;\boldsymbol{\epsilon}$$

si può scrivere:

$$\begin{aligned}
\boldsymbol{\Sigma}_{yy} &= E\left[\left(\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta}+\boldsymbol{\epsilon}\right)\left(\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta}+\boldsymbol{\epsilon}\right)'\right] \\
&= E\left[\left(\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta}+\boldsymbol{\epsilon}\right)\left(\boldsymbol{\eta}'\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\epsilon}'\right)\right] \\
&= E\left[\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta}\boldsymbol{\eta}'\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\epsilon}\boldsymbol{\eta}'\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\eta}\boldsymbol{\epsilon}'+\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] \\
&= \boldsymbol{\Lambda}_{\boldsymbol{y}}\left[E\left(\boldsymbol{\eta}\boldsymbol{\eta}'\right)\right]\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\left[E\left(\boldsymbol{\epsilon}\boldsymbol{\eta}'\right)\right]\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\Lambda}_{\boldsymbol{y}}\left[E\left(\boldsymbol{\eta}\boldsymbol{\epsilon}'\right)\right]+E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right) \\
&= \boldsymbol{\Lambda}_{\boldsymbol{y}}\left[E\left(\boldsymbol{\eta}\boldsymbol{\eta}'\right)\right]\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\left[0\right]\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\Lambda}_{\boldsymbol{y}}\left[0\right]+\boldsymbol{\Theta}_{\epsilon}
\end{aligned}$$

perché per definizione $E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)=\boldsymbol{\Theta}_{\epsilon}$ e per l'ipotesi 2.15 del modello SEM *covariance based* $E\left(\boldsymbol{\epsilon}\boldsymbol{\eta}'\right)=E\left(\boldsymbol{\eta}\boldsymbol{\epsilon}'\right)=0$. Da ciò si ricava che:

$$\boldsymbol{\Sigma}_{yy}=\boldsymbol{\Lambda}_{\boldsymbol{y}}E\left(\boldsymbol{\eta}\boldsymbol{\eta}'\right)\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\Theta}_{\epsilon} \tag{2.24}$$

Per quanto riguarda $E\left(\boldsymbol{\eta}\boldsymbol{\eta}'\right)$ vale la 2.7:

$$\boldsymbol{\eta}=\mathbf{B}\,\boldsymbol{\eta}\;+\;\boldsymbol{\Gamma}\,\boldsymbol{\xi}\;+\;\boldsymbol{\zeta}$$

da cui

$$\begin{aligned} \boldsymbol{\eta} - \mathbf{B}\boldsymbol{\eta} &= \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \\ (I - \mathbf{B})\,\boldsymbol{\eta} &= \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \\ \boldsymbol{\eta} &= (I - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}) \end{aligned} \tag{2.25}$$

dove $(I - \mathbf{B})^{-1}$ esiste perché esiste $\mathbf{B^{-1}}$ per ipotesi.
In definitiva si può scrivere:

$$\begin{aligned} E(\boldsymbol{\eta}\boldsymbol{\eta}') &= E\left(\left[(I - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})\right]\left[(I - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})\right]'\right) \\ &= E\left(\left[(I - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})\right]\left[(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})'(I - \mathbf{B})^{-1\,\prime}\right]\right) \\ &= E\left[(I - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})'(I - \mathbf{B})^{-1\,\prime}\right] \\ &= (I - \mathbf{B})^{-1} E\left[\boldsymbol{\Gamma}\boldsymbol{\xi}\boldsymbol{\xi}'\boldsymbol{\Gamma}' + \boldsymbol{\zeta}\boldsymbol{\xi}'\boldsymbol{\Gamma}' + \boldsymbol{\Gamma}\boldsymbol{\xi}\boldsymbol{\zeta}' + \boldsymbol{\zeta}\boldsymbol{\zeta}'\right](I - \mathbf{B})^{-1\,\prime} \\ &= (I - \mathbf{B})^{-1}\left[\boldsymbol{\Gamma}\left(E(\boldsymbol{\xi}\boldsymbol{\xi}')\right)\boldsymbol{\Gamma}' + \left(E(\boldsymbol{\zeta}\boldsymbol{\xi}')\right)\boldsymbol{\Gamma}' + \boldsymbol{\Gamma}\left[E(\boldsymbol{\xi}\boldsymbol{\zeta}')\right] + E(\boldsymbol{\zeta}\boldsymbol{\zeta}')\right](I - \mathbf{B})^{-1\,\prime} \end{aligned}$$

Essendo per definizione $E(\boldsymbol{\xi}\boldsymbol{\xi}') = \boldsymbol{\Phi}$, $E(\boldsymbol{\zeta}\boldsymbol{\zeta}') = \boldsymbol{\Psi}$ e per l'ipotesi 2.14 del modello SEM *covariance based* $E(\boldsymbol{\zeta}\boldsymbol{\xi}') = E(\boldsymbol{\xi}\boldsymbol{\zeta}') = 0$, si ricava:

$$E(\boldsymbol{\eta}\boldsymbol{\eta}') = (I - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(I - \mathbf{B})^{-1\,\prime} \tag{2.26}$$

che inserita nella 2.24 dà:

$$\boldsymbol{\Sigma}_{yy} = \boldsymbol{\Lambda}_{\boldsymbol{y}}(I - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(I - \mathbf{B})^{-1\,\prime}\boldsymbol{\Lambda}_{\boldsymbol{y}}' + \boldsymbol{\Theta}_{\epsilon} \tag{2.27}$$

Scomposizione della matrice $\boldsymbol{\Sigma}_{xy}$

Essendo $\boldsymbol{\Sigma}_{xy}$ la matrice di covarianza tra le variabili $\boldsymbol{x}$ e $\boldsymbol{y}$ vale che:

$$\boldsymbol{\Sigma}_{xy} = E(\boldsymbol{x}\boldsymbol{y}')$$

e valendo le equazioni 2.8 e 2.9:

$$\boldsymbol{y} = \boldsymbol{\Lambda}_{\boldsymbol{y}}\,\boldsymbol{\eta} \;+\; \boldsymbol{\epsilon}$$

$$\boldsymbol{x} = \boldsymbol{\Lambda}_{\boldsymbol{x}}\,\boldsymbol{\xi} \;+\; \boldsymbol{\delta}$$

si può scrivere:

$$\begin{aligned}
\boldsymbol{\Sigma}_{xy} &= E\left(\boldsymbol{xy}'\right) \\
&= E\left[\left(\boldsymbol{\Lambda_x}\boldsymbol{\xi}+\boldsymbol{\delta}\right)\left(\boldsymbol{\Lambda_y}\boldsymbol{\eta}+\boldsymbol{\epsilon}\right)'\right] \\
&= E\left[\boldsymbol{\Lambda_y}\boldsymbol{\xi}\boldsymbol{\eta}'\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\Lambda_y}\boldsymbol{\xi}\boldsymbol{\epsilon}'+\boldsymbol{\epsilon}'\boldsymbol{\delta}\right] \\
&= \boldsymbol{\Lambda_x}E\left(\boldsymbol{\xi}\boldsymbol{\eta}'\right)\boldsymbol{\Lambda}_{\boldsymbol{y}}'+E\left(\boldsymbol{\delta}\boldsymbol{\eta}'\right)\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\Lambda_x}E\left(\boldsymbol{\xi}\boldsymbol{\epsilon}'\right)+E\left(\boldsymbol{\delta}\boldsymbol{\epsilon}'\right) \\
&= \boldsymbol{\Lambda_x}E\left(\boldsymbol{\xi}\boldsymbol{\eta}'\right)\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\left(0\right)\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\Lambda_x}\left(0\right)+0
\end{aligned}$$

essendo per le ipotesi 2.14, 2.15 e 2.16 del modello SEM *covariance based* $E\left(\boldsymbol{\xi}\boldsymbol{\eta}'\right)=E\left(\boldsymbol{\delta}\boldsymbol{\eta}'\right)=E\left(\boldsymbol{\xi}\boldsymbol{\epsilon}'\right)=0$, da cui ancora:

$$\boldsymbol{\Sigma}_{xy}=\boldsymbol{\Lambda_x}E\left(\boldsymbol{\xi}\boldsymbol{\eta}'\right)\boldsymbol{\Lambda}_{\boldsymbol{y}}' \tag{2.28}$$

Avendo già ricavato l'equazione 2.25:

$$\boldsymbol{\eta}=\left(I-\mathbf{B}\right)^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\xi}+\boldsymbol{\zeta}\right)$$

vale che:

$$\begin{aligned}
\boldsymbol{\Sigma}_{xy} &= \boldsymbol{\Lambda_x}E\left(\boldsymbol{\xi}\boldsymbol{\eta}'\right)\boldsymbol{\Lambda}_{\boldsymbol{y}}' \\
&= \boldsymbol{\Lambda_x}E\left(\boldsymbol{\xi}\left(I-\mathbf{B}\right)^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\xi}+\boldsymbol{\zeta}\right)'\right)\boldsymbol{\Lambda}_{\boldsymbol{y}}' \\
&= \boldsymbol{\Lambda_x}\left[E\left(\boldsymbol{\xi}\boldsymbol{\xi}'\right)\boldsymbol{\Gamma}'\left(I-\mathbf{B}\right)^{-1\ \prime}+E\left(\boldsymbol{\xi}\boldsymbol{\zeta}'\right)\left(I-\mathbf{B}\right)^{-1\ \prime}\right]\boldsymbol{\Lambda}_{\boldsymbol{y}}'
\end{aligned}$$

essendo per definizione $E\left(\boldsymbol{\xi}\boldsymbol{\xi}'\right)=\boldsymbol{\Phi}$ e per l'ipotesi 2.16 del modello SEM *covariance based* $E\left(\boldsymbol{\xi}\boldsymbol{\zeta}'\right)=0$. Infine si ottiene:

$$\boldsymbol{\Sigma}_{\boldsymbol{xy}}=\boldsymbol{\Lambda_x}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\left(I-\mathbf{B}\right)^{-1\ \prime}\boldsymbol{\Lambda}_{\boldsymbol{y}}' \tag{2.29}$$

Per riassumere nella stessa formula tutte le conclusioni precedenti, è possibile scrivere la matrice $\boldsymbol{\Sigma}$ come:

$$\begin{bmatrix}
\boldsymbol{\Sigma_{yy}}=\boldsymbol{\Lambda_y}\left(I-\mathbf{B}\right)^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'+\boldsymbol{\Psi}\right)\left(I-\mathbf{B}\right)^{-1\ \prime}\boldsymbol{\Lambda}_{\boldsymbol{y}}'+\boldsymbol{\Theta}_{\epsilon} & \boldsymbol{\Sigma}_{xy}=\boldsymbol{\Lambda_x}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\left(I-\mathbf{B}\right)^{-1\ \prime}\boldsymbol{\Lambda}_{\boldsymbol{y}}' \\
\boldsymbol{\Sigma}_{xy}=\boldsymbol{\Lambda_x}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\left(I-\mathbf{B}\right)^{-1\ \prime}\boldsymbol{\Lambda}_{\boldsymbol{y}}' & \boldsymbol{\Sigma}_{xx}=\boldsymbol{\Lambda_x}\boldsymbol{\Phi}\boldsymbol{\Lambda}_{\boldsymbol{x}}'+\boldsymbol{\Theta}_{\delta}
\end{bmatrix} \tag{2.30}$$

2.4.6 Le funzioni di discrepanza

Il metodo di stima SEM-ML si basa fondamentalmente sul confronto della matrice $\boldsymbol{\Sigma}$ ottenuta con la formula 2.30 e la matrice di varianza/covarianza osservata $\boldsymbol{S}$, per cui è necessario introdurre una funzione di discrepanza che rende possibile tale confronto.

Perché una funzione F sia utilizzabile come funzione di distanza o *discrepanza* (in questo caso per quantificare la differenza tra le due matrici $\boldsymbol{\Sigma}$ e $\boldsymbol{S}$) deve possedere le seguenti proprietà:

Proprietà 2.1.

1. F è una funzione scalare di $\boldsymbol{\Sigma}$ e $\boldsymbol{S}$.
2. F è non negativa, $F \geq 0$.
3. F è pari a 0 sse $\boldsymbol{\Sigma} = \boldsymbol{S}$.
4. F è una funzione continua in $\boldsymbol{\Sigma}$ e in $\boldsymbol{S}$. △

Una volta trovata la F più opportuna, la stima dei parametri liberi si ottiene minimizzando tale funzione.

In letteratura sono state proposte diverse funzioni di discrepanza a cui corrispondono altrettanti metodi di stima. Di seguito vengono presentate le più utilizzate.

Funzione di discrepanza nel metodo ML

Il primo e più comunemente utilizzato approccio alla stima dei parametri SEM *covariance based* è l'approccio di stima di massima verosimiglianza, che fa ricorso alla seguente funzione di discrepanza:

$$F_{ML} = \text{tr}\left(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\right) - P + \ln|\boldsymbol{S}| - \ln|\boldsymbol{\Sigma}| \tag{2.31}$$

dove P è il numero delle variabili manifeste, quindi la dimensione delle matrici $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$ di cui $|\bullet|$ è il determinante e tr la traccia.

Sotto l'ipotesi che $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$ siano definite positive, F_{ML} è scalare, continua in $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$, non-negativa e pari a 0 sse $\boldsymbol{S} = \boldsymbol{\Sigma}$ perché si verificherebbe che $\ln|\boldsymbol{S}| - \ln|\boldsymbol{\Sigma}| = 0$ e $\boldsymbol{S}\boldsymbol{\Sigma}^{-1} = \boldsymbol{S}\boldsymbol{S}^{-1} = \mathrm{I}_P$. Essendo la traccia di I_P pari alla sua dimensione si ottiene:

$$F_{ML} = P - P + 0 = 0$$

Poiché solitamente la F_{ML} non ha una forma semplice, non è possibile effettuare l'operazione di minimizzazione algebricamente. L'algoritmo di stima generalmente utilizzato è l'algoritmo EM[8].

Sotto l'ipotesi che gli indicatori abbiano distribuzione multinormale e che la matrice di varianza/covarianza campionaria abbia distribuzione di Wishart, gli stimatori ML hanno ottime proprietà asintotiche quali correttezza, consistenza, efficienza e distribuzione normale (per campioni sufficientemente grandi).

Gli stimatori ML sono inoltre invarianti rispetto a trasformazioni di scala, il che permette di utilizzare la matrice di correlazione al posto di quella di varianza/covarianza senza nuocere all'analisi.

Funzione di discrepanza nel metodo ULS

Il metodo di stima dei minimi quadrati non pesati (Unweighted Least Squares) si basa sulla minimizzazione della funzione:

$$F_{ULS} = \frac{\operatorname{tr}\left(\boldsymbol{S} - \boldsymbol{\Sigma}\right)^2}{2} \tag{2.32}$$

ovvero la somma dei quadrati delle differenze fra $\boldsymbol{S} - \boldsymbol{\Sigma}$ divisa per 2.

F_{ULS} è scalare, continua in $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$, non-negativa e pari a 0 sse $\boldsymbol{S} = \boldsymbol{\Sigma}$ poiché:

$$F_{ULS} = \frac{\operatorname{tr}\left(\boldsymbol{S} - \boldsymbol{\Sigma}\right)^2}{2} = \frac{\operatorname{tr}\left(\mathbf{0} - P^2\right)}{2} = 0$$

Anche la 2.32 non si presenta come una funzione facilmente ottimizzabile, da cui la necessità di ricorrere ad algoritmi numerici di calcolo.

Le proprietà asintotiche degli stimatori ULS sono correttezza, e consistenza (anche senza bisogno di particolari ipotesi distributive). Per quanto riguarda l'efficienza asintotica, gli stimatori ML sono preferibili anche se per grandi campioni le due stime si avvicinano molto, inoltre gli stimatori ULS non sono asintoticamente normali.

Bisogna rilevare che gli ULS non sono invarianti per trasformazioni di scala, ciò implica che le stime ottenute a partire dalla matrice di varianza /covarianza non sono riconducibili tramite passaggi algebrici a quelle ottenute con la matrice di correlazione.

[8] Expectation-maximization algorithm

Funzione di discrepanza nel metodo GLS

Il metodo GLS utilizza una generalizzazione della funzione 2.32 in cui entrambe le matrici $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$ vengono premoltiplicate per $\boldsymbol{S}^{-1}$, che funge da matrice dei pesi, in modo da ottenere:

$$F_{GLS} = \frac{\operatorname{tr}\left(\mathrm{I}_P - \boldsymbol{S}^{-1}\boldsymbol{\Sigma}\right)^2}{2} \tag{2.33}$$

Sotto l'ipotesi che $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$ siano definite positive, F_{GLS} è scalare, continua in $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$, non-negativa e pari a 0 sse $\boldsymbol{S} = \boldsymbol{\Sigma}$
poiché:

$$F_{GLS} = \frac{\operatorname{tr}\left(\mathrm{I}_P - \boldsymbol{S}^{-1}\boldsymbol{\Sigma}\right)^2}{2} = \frac{\operatorname{tr}\left(\mathrm{I}_P - \mathrm{I}_P\right)^2}{2} = \frac{\operatorname{tr}\left(\mathbf{0}_P^2\right)}{2} = 0$$

Sotto ipotesi alquanto stringenti, gli stimatori GLS possiedono asintoticameente correttezza, consistenza e normale distribuzione.

Funzione di discrepanza nel metodo ADF

Il metodo a distribuzione asintoticamente libera (Asymptotically distribution free) utilizza la funzione di discrepanza:

$$F_{ADF} = (\boldsymbol{s} - \boldsymbol{\sigma})' \boldsymbol{W}^{-1} (\boldsymbol{s} - \boldsymbol{\sigma}) \tag{2.34}$$

dove $\boldsymbol{s}$ e $\boldsymbol{\sigma}$ derivano dalla vettorizzazione degli elementi non ridondanti rispettivamente di $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$ e dove W, la matrice dei pesi, è una stima consistente della matrice di covarianza, su un ampio campione, degli elementi di $\boldsymbol{S}$ considerati come variabili aleatorie.

F_{ADF} è scalare, continua in $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$, non-negativa e pari a 0 sse $\boldsymbol{S} = \boldsymbol{\Sigma}$
poiché ne deriva che anche $\boldsymbol{s} = \boldsymbol{\sigma}$ e:

$$F_{ADF} = (\boldsymbol{s} - \boldsymbol{\sigma})' W^{-1} (\boldsymbol{s} - \boldsymbol{\sigma}) = (\mathbf{0})' W^{-1} (\mathbf{0}) = 0$$

Per campioni sufficientemente ampi, i metodi ADF possono essere considerati una generalizzazione dei metodi ML, ULS e GLS precedentemente presentati, con opportuni valori della matrice $\boldsymbol{W}$.

2.4.7 Valutazione del modello

La funzione di discrepanza non viene impiegata solo nell'algoritmo di calcolo, ma anche nell'analisi dell'adattamento del modello ai dati.

Come è stato già anticipato parlando dell'identificabilità del modello, le informazioni contenute nella matrice di varianza/covarianza campionarie servono sia per la stima dei parametri sia per la falsificazione del modello (da qui la necessità di un numero di gradi di libertà positivo).

Tali informazioni vengono sfruttate per mezzo della funzione di discrepanza. Le tecniche presentate producono stime che consentono di minimizzare la funzione di discrepanza scelta, e quindi la distanza tra le due matrici $\boldsymbol{S}$ e $\boldsymbol{\Sigma}$, ma alla fine bisogna valutare se quella distanza, la più piccola possibile dato il modello, non sia ancora troppo grande per ritenere che lo schema di relazioni causali che si va a testare sia compatibile con la struttura del campione.

I metodi per valutare la bontà del modello sono molteplici, alcuni hanno solide basi inferenziali, altri si limitano alla valutazione della *performance* dello stesso.

χ^2

La tecnica inferenziale per eccellenza di valutazione di un modello è l'utilizzo del test del χ^2 di Pearson. Se si considera ad esempio la funzione di discrepanza F_{ML}, tale funzione, moltiplicata per l'ampiezza campionaria meno uno, si distribuisce come χ^2 con numero di gradi di libertà pari ai gradi di libertà del modello, ossia:

$$(n-1)\, F_{ML} \sim \chi^2$$

sotto l'ipotesi nulla $F_{ML} = 0$ (vale a dire che il modello testato sia aderente alla realtà). Per tale ragione, quando il p-value del χ^2_{df} associato al modello assume un valore troppo piccolo, generalmente minore di 0.5, si deve rifiutare l'ipotesi nulla, ed il modello non può ritenersi adeguato ai dati campionari.

I problemi che presenta questo metodo sono relativi alla sua dipendenza dall'ampiezza campionaria che, quantunque diminuita di 1, pondera il valore della funzione di discrepanza.

In pratica, in due situazioni distinte in cui viene utilizzata la medesima funzione di discrepanza e se ne ottiene il medesimo valore, ma che differiscono per l'ampiezza del campione utilizzato, potrebbero tranquillamente verificarsi due risultati opposti del test del χ^2, dove il modello associato al campione più piccolo non viene rifiutato, mentre quello che deriva dal campione più ampio risulta inaccettabile.

Esistono diversi modi per ovviare a tale inconveniente, come ad esempio calcolare la numerosità campionaria oltre la quale il modello viene sistematicamente rifiutato e valutare se ci si trovi in questa situazione (e allora è opportuno effettuare indagini supplementari) oppure la numerosità è sufficientemente bassa da ritenere che il motivo del rifiuto risieda esclusivamente nell'inadeguatezza del modello.

In letteratura viene proposto anche un indice che deriva da un "aggiustamento" del χ^2, ossia dalla divisione di quest'ultimo per i suoi gradi di libertà. I gradi di libertà di una χ^2 corrispondono al suo valore atteso, quindi ci si aspetta che tale rapporto assuma un valore prossimo, o quanto meno non distante da 1. Generalmente, per valori minori di 2, si ritiene che il modello presentato sia sufficientemente aderente ai dati.

Gli indici che cercano di ovviare ai difetti del χ^2, non ne hanno però la stessa valenza probabilistica e sebbene generalmente utilizzati, non offrono le medesime garanzie inferenziali, in quanto ad essi non è associata alcuna distribuzione e pertanto alcun test statistico. Tra questi uno dei più interessanti è la radice dell'errore quadratico medio di approssimazione (RMSEA[9]). Introdotto da Steiger e Lind (1980) assume la seguente formulazione:

$$\text{RMSEA} = \sqrt{\frac{F^*}{df}} \tag{2.35}$$

dove F^* è la funzione di discrepanza che si otterrebbe se le tecniche di stima fossero applicate alla matrice di varianza/covarianza della popolazione.

Non essendo a conoscenza della matrice di varianza/covarianza per la popolazione, si procede ad una sua stima tramite la matrice di varianza/covarianza campionaria per ottenere:

$$\text{RMSEA}_{stimato} = \sqrt{\frac{F}{df} - \frac{1}{n-1}} \tag{2.36}$$

dove F è il valore della funzione di discrepanza tra la matrice di varianza/covarianza osservata e quella stimata.

Non esistono indicazioni probabilistiche su quale sia il valore del RMSEA che permetta di stabilire se un modello è plausibile o meno, ma, generalmente, per valori di tale indice pari o inferiori a 0.05 le ipotesi sul modello non vengono rifiutate.

2.5 Modelli ad Equazioni Strutturali: PLS Path Modeling

I parametri principali del modello stimato con metodi *component based* sono i pesi esterni e i coefficienti di *path*, ma soprattutto i punteggi dei concetti astratti. Non a caso l'obiettivo della stima per i PLS-PM è quello di ottenere delle variabili latenti

[9]Root Mean Square Error of Approximation.

il più possibile rappresentative del rispettivo blocco di misura e dello schema di relazioni causali ipotizzato, seguendo un pò quella che è l'ottica dell'Analisi delle Componenti Principali.

Non a caso il primo e più conosciuto metodo di stima *component based* è il PLS-PM, il quale viene introdotto da Wold (1975b) come algoritmo di stima iterativo avente lo scopo di generalizzare la PCA ai dati multi-tavola.

In aggiunta a ciò, il PLS-PM si presenta come un'alternativa all'orientamento deduttivo dei SEM *covariance based*. Infatti, sebbene si occupino entrambi di stimare gli stessi modelli, i due metodi hanno finalità molto differenti. Il metodo SEM-ML per esempio, ha uno scopo confermativo, quindi necessita di un numero considerevole di ipotesi iniziali sui dati perché si possa far ricorso all'inferenza probabilistica classica. Il PLS-PM invece ha chiare finalità esplorativo-predittive, per cui il ricorso alle ipotesi iniziali è limitato alla condizione di specificazione dei predittori: in ciascuna equazione che interviene nel modello ad equazioni strutturali il vettore dei residui deve avere media pari a zero e deve essere incorrelato alle variabili esplicative che intervengono nella regressione. Questa condizione garantisce buone proprietà delle stime ottenute con gli stimatori OLS.

La parsimonia delle assunzioni sui dati fa sì che il campo d'applicazione sia notevolmente ampliato e risulta possibile applicare il PLS Path Modeling a dataset con una modesta quantità di dati mancanti, con un numero di unità anche molto ristretto, nonché con indicatori formativi.

Un'alta caratteristica tipica del PLS-PM, in quanto metodo *covariance based*, è la capacità naturale di fornire automaticamente i punteggi delle variabili latenti (cosa che il LISREL non è in grado di fare se non con grosse difficoltà e sotto pesanti vincoli). Per tale ragione i PLS-PM sono molto diffusi in quegli ambiti in cui la variabile latente rappresenta un indice interpretabile, ad esempio gli indici di soddisfazione della clientela o di redditività o di solidità aziendale, e soprattutto confrontabile.

Il metodo PLS-PM sebbene abbia il merito di non richiedere forti assunzioni, spesso non del tutto aderenti alla realtà, per ora presenta il limite di non avere alla base una funzione scalare globale da ottimizzare, anche se la ricerca sta procedendo in tal senso. Inoltre la convergenza dell'algoritmo non è stata dimostrata se non per diagrammi ad uno o due blocchi [Lyttkens *e altri* (1975)].

2.5.1 La notazione, le equazioni fondamentali e le ipotesi sul modello

La simbologia utilizzata per presentare il PLS-PM cerca di ricalcare quella più comunemente usata nella recente letteratura. Poiché in quest'ambito non vengono fatte distinzioni fra variabili esogene ed endogene il modello causale viene sempli-

ficato, mentre si arricchisce la parte di misura in quanto i PLS-PM ammettono l'impiego dello schema formativo e del MIMIC in aggiunta a quello riflessivo.

Essendo state osservate P variabili su N unità, senza perdere in generalità, viene considerata la matrice $\mathbf{X}$ di tali variabili standardizzate. Supponendo che le $\boldsymbol{x}_i$ siano gli indicatori di Q variabili latenti, è possibile partizionare la matrice dei dati come segue:

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_q, \ldots, \boldsymbol{X}_Q],$$

dove $\boldsymbol{X}_q$ rappresenta il q-esimo blocco di misura composto a sua volta dalle P_q variabili $\boldsymbol{x}_{pq}$, indicizzate doppiamente sia da un indice di blocco q ($q = 1, \ldots, Q$) sia da un indice all'interno del blocco p ($p = 1, \ldots, P_q$).

Con $\boldsymbol{\Xi}$ si indica la matrice di tutte le variabili latenti, in questa sede centrate nella propria media senza perdita di generalità, $\boldsymbol{\xi}_q$ ($q = 1, \ldots, Q$), senza fare distinzione fra esogene ed endogene. I vettori ottenuti in seguito alla stima vengono comunemente indicati con $\hat{\boldsymbol{\xi}}_q$ e chiamati punteggi (scores); essi rappresentano la "quantificazione" dei relativi concetti latenti.

Inoltre $\boldsymbol{\zeta}_q$ e $\mathbf{B}$ rappresentano rispettivamente il vettore dei residui e la matrice dei coefficienti delle relazioni *causali* in cui è implicata la variabile ξ_q.

Rielaborando l'equazione 2.7, è possibile riassumere il modello causale come:

$$\boldsymbol{\xi}_q = \boldsymbol{\Xi}_q \boldsymbol{B} + \boldsymbol{\zeta}_q \tag{2.37}$$

Dove $\boldsymbol{B}$ rispetta le condizioni del modello, il quale non tratta relazioni reciproche o cicliche (per esempio, $\beta_{qq} = 0$ e se $\beta_{qq'} \neq 0$, allora $\beta_{q'q} = 0$)

In aggiunta, $\boldsymbol{\zeta}_q$ deve soddisfare la condizione di specificazione del predittore, quindi deve avere media pari a zero ed essere incorrelato con tutti i predittori di ciascuna variabile latente $\boldsymbol{\xi}_q$, ovvero nelle equazioni:

$$\boldsymbol{\xi}_q = \sum_{q'=1}^{Q} \beta_{qq'} \boldsymbol{\xi}_{q'} + \boldsymbol{\zeta}_q \tag{2.38}$$

dall'equazione 2.38 $\boldsymbol{\zeta}_q$ deve essere incorrelato con ciascuna variabile $\boldsymbol{\xi}_{q'}$ a cui corrisponde un coefficiente $\beta_{qq'}$ diverso da zero per costruzione.

Per quanto riguarda il modello di misura, nel caso *riflessivo* ogni indicatore $\boldsymbol{x}_{pq}$ è legato da una relazione diretta semplice alla latente $\boldsymbol{\xi}_q$ che, semplificando l'equazione 2.2, può essere espressa come:

$$\boldsymbol{x}_{pq} = \lambda_{pq} \boldsymbol{\xi}_q + \boldsymbol{\epsilon}_{pq} \tag{2.39}$$

dove con λ_{pq} si indica il coefficiente[10] della regressione della variabile manifesta $\boldsymbol{x}_{pq}$ sulla latente $\boldsymbol{\xi}_q$ e $\boldsymbol{\epsilon}_q$ rappresenta il vettore dei residui.

Anche in questo caso, come per il LISREL, se si impiega un modello riflessivo è necessario valutare sensatezza di tale assunzione tramite l'impiego degli indicatori presentati nel paragrafo 2.1: Analisi Componenti Principali, α di Cronbach e ρ di Dillon-Goldstein).

La condizione di *specificazione del predittore* per il modello riflessivo, che corrisponde a residui aventi media pari a zero e incorrelati con la variabile latente, viene formalizzata come:

$$E\left(\boldsymbol{x}_{pq} \mid \boldsymbol{\xi}_q\right) = \lambda_{pq}\boldsymbol{\xi}_q \tag{2.40}$$

Nel caso *formativo* invece, la variabile latente $\boldsymbol{\xi}_q$ è il risultato dell'azione causale multipla delle variabili manifeste del blocco $\boldsymbol{X}_q$, che, richiamando l'equazione 2.5, può essere espressa come:

$$\boldsymbol{\xi}_q = \boldsymbol{X}_q\boldsymbol{w}_q + \boldsymbol{\delta}_{pq} = \sum_{p=1}^{P_q} w_{pq}\boldsymbol{x}_{pq} + \boldsymbol{\delta}_{pq} \tag{2.41}$$

dove con w_{pq} si indicano i coefficienti della regressione della variabile latente $\boldsymbol{\xi}_q$ sulle manifeste $\boldsymbol{X}_q$ e con $\boldsymbol{\delta}_p q$ il termine di errore, ossia la parte di variabile latente non spiegata dagli indicatori osservati. Per tale modello è opportuno che le variabili siano correlate fra loro il meno possibile.

La condizione di *specificazione del predittore*, nel caso formativo, assume la forma:

$$E\left(\boldsymbol{\xi}_q \mid \boldsymbol{x}_{pq}\right) = \sum_{p=1}^{P_q} w_{pq}\boldsymbol{x}_{pq} \tag{2.42}$$

e prevede anche in questo caso che i residui $\boldsymbol{\delta}_{pq}$ abbiano media zero e siano incorrelati con i predittori $\boldsymbol{x}_{pq}$.

Nel caso di un modello causale MIMIC, il blocco $\boldsymbol{X}_q$ viene suddiviso in sottogruppi di variabili ed ognuno viene formalizzato secondo il relativo schema di misura, valendo per ognuno dei sottomodelli di misura vale la condizione di specificazione del predittore nella formulazione opportuna.

L'ultima indicazione sulla notazione riguarda le stime parziali delle variabili latenti che vengono indicate con ν_q quando sono il risultato della stima *esterna* e con ϑ_q nel caso di stima *interna*.

[10] Loading.

2.5.2 L'algoritmo di stima

Per i metodi *component based* i parametri fondamentali del modello sono rappresentati dai punteggi $\boldsymbol{\xi}_q$ delle variabili latenti, dai pesi w_{pq} da cui essi derivano e dai coefficienti $\beta_{qq'}$ che quantificano la forza dei legami causali fra le $\boldsymbol{\xi}_q$ e le $\boldsymbol{\xi}_{q'}$. La procedura che permette di ottenerne la stima PLS-PM *alterna* la stima ottenuta utilizzando i modelli di misura alla stima ottenuta sfruttando le relazioni causali, finché entrambe non convergono ad uno stesso valore. Purtroppo la convergenza del metodo è stata formalmente dimostrata solo per modelli ad uno o due blocchi, benché in pratica sia sempre assicurata.

Proprio dall'alternarsi asincrono della stima delle due parti del modello (quella causale e quella relativa ai modelli di misura) deriva l'attributo di *partial* del metodo.

Indicata con $\boldsymbol{\nu}_q$ la stima *esterna* (ottenuta grazie ai modelli di misura) dei punteggi della variabile latente $\boldsymbol{\xi}_q$, tale stima viene ricavata dalla standardizzazione della combinazione lineare delle relative variabili manifeste, utilizzando i pesi *esterni* w_{pq}:

$$\boldsymbol{\nu}_q \propto \sum_{p=1}^{P_q} w_{pq}\boldsymbol{x}_{pq} = \boldsymbol{X}_q\boldsymbol{w}_q \tag{2.43}$$

dove il simbolo $\propto$ indica che il lato sinistro dell'equazione corrisponde alla standardizzazione del lato destro.

All'avvio della procedura i pesi w_{pq} vengono scelti arbitrariamente (è comune la scelta di pesi pari al primo autovalore nella PCA) per essere successivamente aggiornati alla luce dei risultati ottenuti dalla stima interna.

Utilizzando i punteggi ottenuti con la stima esterna risulta poi possibile effettuare la stima *interna* del vettore dei punteggi di $\boldsymbol{\xi}_q$, indicata con $\boldsymbol{\vartheta}_q$, la quale considera invece i legami causali fra $\boldsymbol{\xi}_q$ e le variabili latenti ad essa connesse.

Sia la matrice binaria $\boldsymbol{C}$, quadrata di ordine Q, tale che il generico elemento $c_{qq'}$ sia pari ad uno se $\boldsymbol{\xi}_q$ è legato a $\boldsymbol{\xi}_{q'}$ e sia pari a zero altrimenti, allora tale matrice può essere utilizzata come indicatrice di relazione e la stima interna può essere espressa come:

$$\boldsymbol{\vartheta}_q \propto \sum_{q'=1}^{Q} c_{qq'}e_{qq'}\boldsymbol{\nu}_q \tag{2.44}$$

dove il simbolo $\propto$ indica ancora che il lato sinistro dell'equazione corrisponde alla standardizzazione del lato destro e dove i coefficienti $e_{qq'}$ vengono denominati pesi *interni* e possono essere calcolati seguendo diverse proposte:

- Schema del **centroide**. Lo schema di stima interna originale introdotto da Wold prevede che i pesi $e_{qq'}$ siano posti pari al segno della correlazione fra la variabile $\boldsymbol{\nu}_q$ e la variabile $\boldsymbol{\nu}'_q$. Tale scelta velocizza notevolmente i calcoli, ma, essendo preso in considerazione solo il segno della relazione, quando questa è prossima a zero, pone seri problemi di stabilità e può cambiare anche per piccole fluttuazioni. Nonostante ciò, nella generalità dei casi lo schema non presenta grosse difficoltà pratiche.

- Schema **fattoriale**. Per ovviare ai problemi di stabilità posti dalla scelta precedente, Lohmöller (1989) propone di assegnare ai pesi $e_{qq'}$ un valore pari alla correlazione fra la variabile $\boldsymbol{\nu}_q$ e la variabile $\boldsymbol{\nu}_{q'}$.

-Schema **strutturale** (detto anche schema *path weighting*). Un approccio più esplorativo, quando non si hanno informazioni a priori attendibili, consiste nel suddividere le variabili latenti $\boldsymbol{\xi}_{q'}$ in due gruppi: variabili latenti esplicative di $\boldsymbol{\xi}_q$ e variabili latenti spiegate da $\boldsymbol{\xi}_q$. Per i coefficienti delle variabili spiegate si pongono gli $e_{qq'}$ pari alla correlazione fra $\boldsymbol{\nu}_q$ e $\boldsymbol{\nu}_{q'}$. Per quanto riguarda le esplicative invece, gli $e_{qq'}$ sono posti pari al coefficiente di regressione relativo a $\boldsymbol{\nu}'_q$ nella regressione multipla di $\boldsymbol{\nu}_q$ su tutte le $\boldsymbol{\nu}'_q$ legate alle esplicative della variabile $\boldsymbol{\xi}_q$.

I risultati non vengono significativamente influenzati dalla scelta di questo schema, ma danno un importante risultato teorico, perché è possibile mettere in relazione i PLS-PM con i metodi dell'analisi multi-tavola.

Dalle stime interne dei punteggi è possibile ricavare dei nuovi valori per i pesi esterni, che tengano conto della rete di relazioni causali. I metodi di aggiornamento dei pesi sono diversi, i principali dei quali sono:

-**Modo A**. Ogni peso interno w_{pq} è posto pari al coefficiente della regressione semplice di $\boldsymbol{x}_{pq}$, la p-esima variabile manifesta del q-esimo blocco, sul vettore $\boldsymbol{\vartheta}_q$ della stima interna della q-esima variabile latente che, nel caso di variabili manifeste standardizzate, risulta pari alla covarianza fra la manifesta e la latente, ovvero:

$$w_{pq} = cov\left(\boldsymbol{x}_{pq}\boldsymbol{\vartheta}_q\right) \tag{2.45}$$

-**Modo B**. Il vettore $\boldsymbol{w}_q$ dei pesi esterni è posto pari al vettore dei coefficienti della regressione multipla del vettore della stima interna $\boldsymbol{\vartheta}_q$, sulle sue variabili manifeste centrate $\boldsymbol{X}_q$:

$$\boldsymbol{w}_q = \left(\boldsymbol{X}'_q\boldsymbol{X}_q\right)^{-1}\boldsymbol{X}'_q\boldsymbol{\vartheta}_q \tag{2.46}$$

La scelta dello schema di aggiornamento è strettamente legato alla natura del modello di misura e alla variabile analizzata.

Il Modo A risulta, per costruzione, più appropriato al caso di indicatori legati alla latente da uno schema riflessivo. Viceversa, il Modo B si adatta notevolmente meglio agli schemi formativi. Va aggiunto che il Modo A è indicato per variabili

endogene al modello causale, invece il Modo B risulta preferibile per variabili esogene.

La regressione multipla utilizzata nel modo B potrebbe essere affetta da collinearità tra i predittori. Per ovviare a tale problema viene sempre più di frequente adottata, al posto della stima OLS nel calcolo dei pesi esterni, la regressione PLS prendendo in considerazione tutte le componenti.

Per determinare se un blocco sia riflessivo o meno si fa comunemente ricorso agli indicatori presentati nel paragrafo 2.1 quali gli autovalori della PCA, l'α di Crombach ed il ρ di Dillon-Goldstein, ma in alternativa esiste un ulteriore sviluppo, più esplorativo, per l'aggiornamento dei pesi esterni, implementato nel cosiddetto Modo PLS [[Esposito Vinzi (2008, 2009); Esposito Vinzi e Russolillo (2010)]].

-**Modo PLS**. In questo caso si utilizza una regressione PLS su ciascun blocco e in base ai risultati ottenuti si determina quale modello di misura utilizzare; $\boldsymbol{w}_q$ risulta essere il vettore dei coefficienti della regressione di $\boldsymbol{\vartheta}_q$ su $\boldsymbol{X}_q$.

Se l'algoritmo PLS-PM converge su una regressione PLS ad una sola componente, la stima dei coefficienti con il Modo PLS coincide con quella ottenuta con il Modo A, e si può ritenere che lo schema di misura sia riflessivo, mentre se converge su una regressione PLS a P_q componenti, la stima ottenuta con il Modo PLS coincide con quella del Modo B e suggerisce la presenza di uno schema formativo.

Nel caso in cui l'algoritmo converga su una regressione PLS ad un numero di componenti h con $1 < h < P_q$, si profila uno schema MIMIC, dove ogni sub-blocco di variabili sottolinea un'aspetto diverso della variabile latente, rappresentato da una sotto-dimensione.

Quando viene raggiunta la convergenza, ovvero quando stima interna e stima esterna coincidono ($\boldsymbol{\nu}_q = \boldsymbol{\vartheta}_q$), vengono ricavati i pesi esterni finali w_{pq} ed è possibile calcolare la stima standardizzata dei punteggi delle variabili latenti:

$$\hat{\boldsymbol{\xi}}_q \propto \sum_{p=1}^{P_q} w_{pq}\boldsymbol{x}_{pq} = \boldsymbol{X}_q w_q \tag{2.47}$$

L'ultimo passaggio consiste nello stimare la matrice dei coefficienti β_{pq} del modello causale, facendo ricorso alla regressione multipla OLS tra i punteggi delle variabili latenti, seguendo il diagramma delle relazioni causali:

$$\boldsymbol{\beta}_j = \left(\hat{\boldsymbol{\Xi}}'\hat{\boldsymbol{\Xi}}\right)^{-1}\hat{\boldsymbol{\Xi}}'\hat{\boldsymbol{\xi}}_j \tag{2.48}$$

Anche in questa regressione, allo scopo di ridurre la variabilità della stima, è possibile sostituire la regressione OLS con la regressione PLS nei casi in cui è presente una forte multicollinearità fra le variabili latenti.

2.5.3 I metodi per la valutazione del modello

Purtroppo, come già anticipato, non è stata ancora identificata una funzione globale ottimizzata dalle stime PLS-PM, quindi non è possibile utilizzare una funzione di adattamento globale per valutare la bontà del modello. Nonostante ciò, essendo il PLS-PM una tecnica orientata alla predizione, è possibile indirizzare la validazione verso la valutazione della capacità predittiva del modello stimato, considerandone sia entrambe le parti di cui si compone separatamente sia tutta la struttura nella sua interezza.

Valutazione del modello di misura

La validità dello schema di misura impiegato viene controllata nella fase preparatoria dell'analisi ma, a posteriori, è opportuno controllare la qualità del modello alla luce del risultato della stima dei parametri. Lo strumento utilizzato in questo caso è l'indice di comunalità, che per le variabili qui trattate vale:

$$Comunalità_q = \frac{1}{P_q}\sum_{p=1}^{P_q} \text{cor}^2\left(\boldsymbol{x}_{pq}, \hat{\boldsymbol{\xi}}_q\right) = \frac{1}{P_q}\sum_{p=1}^{P_q} \hat{\lambda}_{pq} \tag{2.49}$$

Questo indice misura quanto della variabilità dell'indicatore osservato $\boldsymbol{x}_{pq}$ viene spiegato dalla variabile latente $\boldsymbol{\xi}_q$, in modo da offrire un'indicazione di quanto bene le variabili manifeste descrivono la relativa variabile latente. Si può notare che la comunalità non è nient'altro che la media delle correlazioni al quadrato fra le variabili manifeste e la propria latente.

La comunalità ha molto in comune con l'indice AVE[11] [Fornell e Larcker (1981)] dato da:

$$AVE_q = \frac{\sum_{p=1}^{P_q} \hat{\lambda}_{pq}}{\sum_{q=1}^{P_q} \text{var}\left(\boldsymbol{x}_{pq}\right)} \tag{2.50}$$

Infatti in caso di variabili manifeste standardizzate, $\text{var}\left(\boldsymbol{x}_{pq}\right) = 1$, AVE e comunalità coincidono.

É possibile effettuare una valutazione complessiva dei modelli di misura analizzando la comunalità media:

$$\overline{Comunalità} = \frac{1}{P}\sum_{q=1}^{Q} P_q \cdot Comunalità_q \tag{2.51}$$

ovvero la media pesata delle comunalità per ogni blocco, i cui pesi sono dati dal numero di variabili manifeste di ogni blocco. Per quanto già detto, la comunalità

[11] Average Variance Extracted.

media non è altro che la media del quadrato di tutte le correlazioni tra gli indicatori e la propria variabile latente:

$$\overline{Comunalità} = \frac{1}{P}\sum_{q=1}^{Q} P_q \left[\frac{1}{P_q}\sum_{p=1}^{P_q} \mathrm{cor}^2\left(\boldsymbol{x}_{pq}, \boldsymbol{\xi}_q\right)\right] = \frac{1}{P}\sum_{q=1}^{Q}\sum_{p=1}^{P_q} \mathrm{cor}^2\left(\boldsymbol{x}_{pq}, \boldsymbol{\xi}_q\right) \quad (2.52)$$

Valutazione del modello causale

La capacità predittiva di ogni singola equazione può essere valutata con l'indice R^2, ma ciò non è sufficiente per valutare il modello causale nel suo complesso, perché le equazioni strutturali vengono considerate singolarmente ed inoltre viene del tutto ignorato il modello di misura. Per questo motivo viene introdotto l'indice di ridondanza:

$$Ridondanza_q = Comunalità_q \times R^2\left(\boldsymbol{\xi}_q, \{\text{le } \boldsymbol{\xi}_{q'} \text{ esplicative di } \boldsymbol{\xi}_q\}\right) \quad (2.53)$$

che misura la parte di variabilità degli indicatori $\boldsymbol{X}_q$ del q-esimo blocco spiegata dai predittori latenti $\boldsymbol{\xi}_{q'}$, esplicativi di $\boldsymbol{\xi}_q$.

Una misura globale della qualità del modello causale viene perciò ricavato come media delle ridondanze:

$$\overline{Ridondanza} = \frac{1}{Q}\sum_{q=1}^{Q} Ridondanza_q \quad (2.54)$$

Valutazione globale del modello

Per valutare la bontà complessiva del modello è stato introdotto l'indice di GoF[12] [Tenenhaus *e altri* (2004)] che è una media geometrica della comunalità media e dell'R^2 medio pari a

$$\overline{R^2} = \frac{1}{Q}\sum_{q=1}^{Q} R^2\left(\boldsymbol{\xi}_q, \{\text{le } \boldsymbol{\xi}_{q'} \text{ esplicative di } \boldsymbol{\xi}_q\}\right) \quad (2.55)$$

da cui:

$$GoF = \sqrt{\overline{Comunalità} \times \overline{R^2}} \quad (2.56)$$

[12] Goodness of Fit.

Tenendo conto delle equazioni 2.52 e 2.55 il GoF può essere riscritto come:

$$GoF = \sqrt{\frac{\sum_{q=1}^{Q}\sum_{p=1}^{P_q} \mathrm{cor}^2\left(\boldsymbol{x}_{pq}, \boldsymbol{\xi}_q\right)}{P} \times \frac{\sum_{q=1}^{Q} R^2\left(\boldsymbol{\xi}_q, \{\text{le } \boldsymbol{\xi}_{q'} \text{ esplicative di } \boldsymbol{\xi}_q\}\right)}{Q}} \tag{2.57}$$

Una versione normalizzata del Gof [Tenenhaus *e altri* (2004)] si ottiene rapportando tale indice al suo massimo, ovvero al prodotto tra il massimo della comunalità media e il massimo dell'$\overline{R^2}$, per ottenere:

$$GoF_{norm} = \sqrt{\frac{1}{P}\sum_{q=1}^{Q}\frac{\sum_{p=1}^{P_q} \mathrm{cor}^2\left(\boldsymbol{x}_{pq}, \boldsymbol{\xi}_q\right)}{\lambda_q} \times \frac{1}{Q}\sum_{q=1}^{Q}\frac{R^2\left(\boldsymbol{\xi}_q, \{\text{le } \boldsymbol{\xi}_{q'} \text{ esplicative di } \boldsymbol{\xi}_q\}\right)}{\rho_q^2}} \tag{2.58}$$

L'indice GoF_{norm} varia tra 0 e 1 ed è tanto più alto quanto migliore è il modello. Solitamente si ritiene che un valore dell'indice almeno pari a 0.9 sia un buon risultato.

Validazione del modello

La parsimonia delle ipotesi sulla distribuzione delle variabili implicate nell'analisi consente una notevole estensione dell'applicabilità del modello, d'altro canto limita fortemente la possibilità di utilizzare i risultai dell'inferenza classica. Per tale ragione è necessario valutare la significatività dei parametri tramite tecniche di validazione incrociata, rappresentate principalmente da stime *jackknife* e *bootstrap* [Efron (1982); Efron e Tibshirani (1993)].

A maggior ragione non è possibile fare appello alla teoria probabilistica per validare gli indici del modello. Anche in questo caso l'unica alternativa è la validazione incrociata tramite blindfolding [Tenenhaus *e altri* (2005)]. Se gli indici di comunalità ottenuti con validazione incrociata per ogni bloccco sono tutti positivi, la loro media può essere utilizzate per misurare la qualità complessiva della misura delle variabili latenti.

Allo stesso modo, la media degli indici di ridondanza con validazione incrociata può essere usata per una valutazione complessiva del modello.

Per garantire la validità predittiva delle stime PLS-PM è necessario validare le relazioni del modello, anche in questo caso principalmente con tecniche di validazione incrociata jackknife e bootstrap, ed in particolare è necessario:

- che ogni λ_{pq} sia significativamente diversa da zero, ovvero che ogni indicatore riflessivo sia correlato alla respettiva variabile latente;

- che w_{pq} sia significativamente diversa da zero, ovvero che ogni variabile latente sia significativamente influenzata da tutti gli indicatori del proprio blocco di misura;

- che $\beta_{qq'}$ sia significativamente diversa da zero, ovvero che ogni predittore $\boldsymbol{\xi}_{q'}$ eserciti un'azione significativa sulle relative variabili latenti spiegate. In questo caso specifico, se lo schema causale viene modellato in un contesto OLS, è possibile far riferimento al test t di Student, perchè questo è robusto dalla deviazione dall'ipotesi di normalità;

- che $R^2_q = R^2\left(\boldsymbol{\xi}_q, \{\text{le } \boldsymbol{\xi}_{q'} \text{ esplicative di } \boldsymbol{\xi}_q\}\right)$ sia significativamente diversa da zero, ovvero che ogni $\boldsymbol{\xi}_q$ sia spiegata in maniera significativa dai propri predittori latenti $\boldsymbol{\xi}_{q'}$;

- che $\text{cor}^2\left(\boldsymbol{\xi}_q, \boldsymbol{\xi}_{q'}\right)$ sia significativamente diversa da uno, ovvero che, sebbene i concetti astratti implicati nel modello siano tra loro naturalmente correlati secondo lo schema causale, ciascuna variabile latente misura un concetto significativamente diverso da quello misurato dalle altre. (**Validità Discriminante** del PLS-PM)

- che AVE_q e $\text{AVE}_{q'}$ siano significativamente maggiori di $\text{cor}^2\left(\boldsymbol{\xi}_q, \boldsymbol{\xi}_{q'}\right)$, ovvero che ogni variabile latente $\boldsymbol{\xi}_q$ sia legata al proprio blocco di misura in maniera significativamente maggiore che ad ogni altra variabile latente $\boldsymbol{\xi}_{q'}$ rappresentante un diverso blocco di indicatori. (**Validità Convergente** del PLS-PM)

- che $\text{cor}^2\left(\boldsymbol{x}_{pq}, \boldsymbol{\xi}_q\right)$ sia significativamente maggiore di $\text{cor}^2\left(\boldsymbol{x}_{pq}, \boldsymbol{\xi}_{q'}\right)$, ovvero che ogni indicatore sia legato al concetto che deve spiegare in maniera significativamente più forte rispetto ad ogni altro concetto latente. (**Validità Monofattoriale** del PLS-PM)

Quando, in base ai risultati della validazione incrociata, qualcuna di queste ipotesi deve essere rigettata, allora è necessario rivedere la scelta degli indicatori e la specificazione del modello.

2.5.4 L'eterogeneità latente e REBUS-PLS

Normalmente i modelli ad equazioni strutturali assumono che vi sia omogeneità tra le unità osservate e che il modello globale sia rappresentativo di tutti i dati. Quest'ipotesi non è sempre verificata e quando esistono dei sotto-gruppi di dati omogenei al loro interno ma distinti dagli altri gruppi, il modello andrebbe costruito e stimato separatamente per ognuno di essi. In tal caso si presenta la classica situazione di *relazione condizionata*, introdotta nella relativa sezione del paragrafo 1.2.2.

L'eterogeneità dei dati può essere manifesta, se esiste e si conosce la variabile di raggruppamento che determina il comportamento diverso dei soggetti, oppure latente quando il criterio di raggruppamento non è conosciuto.

Nel caso di manifesta eterogeneità è sufficiente considerare la variabile di raggruppamento come fattore moderatore e i submodelli ottenuti saranno migliori del modello globale.

Purtroppo, nella realtà raramente si conosce la fonte dell'eterogeneità, quindi è necessario utilizzare tecniche statistiche orientate a far emergere le classi latenti.

L'approccio classico prevede che si effettui un clustering a priori sui dati grezzi, ma in tal modo non si tiene conto del modello di relazioni (causali e di misura) che caratterizzano i dati stessi, ma nel PLS-PM sono state introdotte tecniche integrate che innestano la ricerca di eterogeneità latente direttamente nell'algoritmo di stima PLS, ovvero prendendo in considerazione la struttura del modello. In particolare in questa sezione viene presentato l'algoritmo REBUS-PLS[13] [Trinchera (2007); Esposito Vinzi *e altri* (2008)].

L'obiettivo del REBUS-PLS è quello di catturare l'eterogeneità latente nei dati, sia al livello di modello di misura sia al livello di modello causale. A tale scopo viene introdotta una nuova definizione di distanza fra unità e modello, in cui viene presa in considerazione la *performance* del modello in termini di residui relativi al modello di misura ed al modello strutturale, per tutte le variabili latenti a disposizione.

Le ipotesi richieste sono le stesse del PLS-PM classico, quindi non sono necessarie ipotesi distributive, però è ammesso il solo schema di misura riflessivo.

La distanza implicata nell'algoritmo REBUS-PLS è definita come la somma di residui al quadrato; ad essa si fa di seguito riferimento come misura di vicinanza (CM[14]).

La CM viene costruita prendendo in considerazione l'indice di GoF, equazione 2.56 e basandosi sui residui del modello delle comunalità (regressione delle variabili manifeste sulle rispettive latenti) e del modello strutturale (regressione delle endogene latenti sulle rispettive esplicative latenti)per ottenere:

$$CM_{ig} = \sqrt{\frac{\sum_{q=1}^{Q}\sum_{p=1}^{P_q}\left[\frac{e^2_{ipqg}}{Comunalità\left(\hat{\boldsymbol{\xi}}_{qg},\boldsymbol{x}_{pq}\right)}\right]}{\sum_{i=1}^{N}\sum_{q=1}^{Q}\sum_{p=1}^{P_q}\left[\frac{e^2_{ipqg}}{Comunalità\left(\hat{\boldsymbol{\xi}}_{qg},\boldsymbol{x}_{pq}\right)}\right]} \times \frac{\sum_{q^*=1}^{Q^*}\left[\frac{f^2_{iq^*g}}{R^2\left(\hat{\boldsymbol{\xi}}_{q^*},\left\{\text{le }\hat{\boldsymbol{\xi}}_q \text{ esplicative di }\hat{\boldsymbol{\xi}}_{q*}\right\}\right)}\right]}{\sum_{i=1}^{N}\sum_{q^*=1}^{Q^*}\left[\frac{f^2_{iq^*g}}{R^2\left(\hat{\boldsymbol{\xi}}_{q^*},\left\{\text{le }\hat{\boldsymbol{\xi}}_q \text{ esplicative di }\hat{\boldsymbol{\xi}}_{q*}\right\}\right)}\right]}} \tag{2.59}$$

dove:

- $Comunalità\left(\hat{\boldsymbol{\xi}}_{qg},\boldsymbol{x}_{pq}\right)$ è l'indice di comunalità per la p-esima variabile manifesta del q-esimo blocco nella g-esima classe latente.

[13] REsponse Based Unit Segmentation in PLS-PM.

[14] Closeness Measure.

- e_{ipqg} è il residuo del modello di misura per la i-esima unità nella g-esima classe latente, corrispondente alla p-esima variabile manifesta del q-esimo blocco (*residuo di comunalità*).

- f_{iq^*g} è il residuo del modello causale per la i-esima unità nella g-esima classe latente, corrispondente al q^*-esimo blocco endogeno.

- n_g è il numero di unità appartenenti alla g-esima classe latente.

- m_g è il numero di dimensioni. Poiché si assume che i modelli di misura siano tutti riflessivi, la dimensione di ogni blocco è sempre uno (esiste un'unica variabile latente).

Il residuo di misura (o residuo di comunalità) della i-esima unità nella g-esima classe latente è ottenuto come:

$$e_{ipqg} = x_{ipq} - \hat{x}_{ipqg} \tag{2.60}$$

dove $\hat{x}_{ipqg} = \lambda_{pqg}\hat{\xi}_{iqg}$, con λ_{pqg} pari al *loading* associato alla p-esima variabile del q-esimo blocco nella g-esima classe latente, e $\hat{\xi}_{iqg}$ è pari al punteggio della q-esima variabile latente per la i-esima unità, calcolato usando i pesi esterni stimati per la g-esima classe latente:

$$\hat{\xi}_{iqg} = \sum_{p=1}^{P_q} w_{pqg} x_{ipq} \tag{2.61}$$

dove w_{pqg} è il peso esterno della relazione tra la p-esima variabile manifesta del q-esimo blocco e la relativa variabile latente, all'interno della g-esima classe latente.

I pesi esterni vengono ottenuti applicando il PLS-PM sulle unità appartenenti alla g-esima classe latente, in altre parole i residui di comunalità sono i residui della regressione semplice di ogni indicatore sulla corrispondente variabile latente.

I residui strutturali sono i residui della regressione multipla delle variabili latenti endogene sulle proprie variabili latenti esplicative:

$$f_{iq^*g} = \hat{\xi}_{iq^*g} - y_{iq^*g} \tag{2.62}$$

dove $y_{iq^*g} = \sum_{q=1}^{Q\ su\ q^*} \beta_{qq^*g}\xi_{iqg}$ e β_{qq^*g} è il coefficiente causale che lega la q-esima variabile latente esplicativa alla q^*-esima variabile endogena della g-esima classe latente.

Se due modelli mostrano uguali coefficienti di *path* ma differiscono per uno o più pesi esterni nei blocchi esogeni, il REBUS-PLS è in grado di identificare la fonte dell'eterogeneità. Inoltre, data la struttura del GoF, il REBUS-PLS è naturalmente predisposto ad individuare sottomodelli con un maggiore potere predittivo.

La stabilità della composizione delle classi viene considerata come criterio per stabilire il raggiungimento della convergenza, sebbene non ci siano prove formali al

riguardo. Inoltre l'identificazione delle classi latenti basate unicamente sul modello non fornisce chiari criteri interpretativi, in tal senso delle variabili esterne socio-demografiche possono aiutare a definire i segmenti.

Algoritmo 1 REBUS-PLS

1: Stima PLS-PM del modello globale e valutazione della bontà del modello;
2: Calcolo dei residui di comunalità e strutturali di tutte le unità del modello, secondo le equzioni 2.60 e 2.62;
3: Classificazione gerarchica sui residui calcolati al punto 2;
4: Scelta del numero di classi (G) in base al dendrogramma ottenuto al *punto* 3;
5: Assegnazione delle unità ad ogni classe in base ai risultati della cluster analysis;

6: Stima dei G modelli locali (uno per classe);
7: Calcolo della prossimità CM per ogni unità rispetto ad ogni modello locale in base all'equazione 2.59;
8: Assegnazione di ogni unità al modello locale più prossimo;
9: **SE** si giunge alla stabilità della composizione delle classi ALLORA andare al *passo* 10; *ALTRIMENTI* tornare al *passo* 6;
10: Descrizione delle classi ottenute in accordo con le differenze fra modelli locali.

2.6 Modelli ad Equazioni Strutturali: un confronto fra i metodi *covariance* e *component based* presentati

Entrambi i metodi di stima analizzati nei paragrafi precedenti presentano punti di forza e punti di debolezza.

Il SEM-ML, come tutti i metodi *covariance based*, fa riferimento ad un ben identificato criterio di ottimizzazione globale e può attingere a migliori strumenti di misura della qualità del modello teorico, anche in ragione del fatto che il fine ultimo dell'approccio è quello di testare la struttura causale e di misura ipotizzate alla luce dei dati raccolti.

Va aggiunto che i metodi *covariance based* consentono l'imposizione di vincoli sui parametri e il trattamento di modelli non ricorsivi, mentre nessuna di tali opzioni è praticabile con la stima PLS-PM. Inoltre, poiché le stime PLS-PM non ottimizzano uno specifico criterio globale, la misura della qualità del modello interno è sottostimata, mentre la qualità del modello esterno è sovrastimata.

D'altro canto i modelli LISREL spesso presentano difficoltà di identificazione (sconosciuti nel PLS-PM) e problemi di non convergenza dell'algoritmo mentre il PLS-PM, benché non si disponga di una dimostrazione generale, converge praticamente sempre.

Tra i casi più interessanti in cui vengono a cadere le ipotesi necessarie per l'applicabilità del SEM-ML vi sono sia la situazione di database incompleti, sia quella di database *orizzontali*[15], in cui si ha a disposizione solo un piccolo campione ed il numero di variabili manifeste è nettamente maggiore del numero delle osservazioni.

Il caso dei dati mancanti può essere risolto solo a priori dal SEM-ML, tramite l'imputazione preventiva dei dati. Al contrario esistono varianti del PLS-PM che consentono di implementare il trattamento dei missing nell'algoritmo di stima per rendere più efficiente la procedura [Lohmöller (1989)].

Il problema creato dai database orizzontali è che vengono a cadere le ipotesi di applicabilità della regressione OLS, e non è possibile procedere alla stima con i metodi LISREL. Invece, sostituendo la regressione OLS con la regressione PLS, il PLS-PM resta applicabile e dà risultati abbastanza robusti.

Per quanto riguarda la quantificazione dei concetti astratti, i metodi *component based*, essendo orientati alla predittività, sono naturalmente predisposti alla produzione dei punteggi delle variabili latenti, che invece dai metodi *covariance based* non sono ben definiti e possono essere ricavati solo al prezzo di calcoli complicati e ipotesi aggiuntive pesanti.

[15]Detti anche *landscape* dal termine comunemente utilizzato in grafica per definire la disposizione orizzontale della tabella dei dati.

Capitolo 3

Le Reti Probabilistiche Bayesiane

Le Reti Probabilistiche Bayesiane sono uno strumento statistico che si pone lo scopo di analizzare un insieme di dati multivariato, ma al contrario dei modelli ad equazioni strutturali, non necessitano di uno schema causale prestabilito.

I SEM permettono di stimare, modificare o rigettare il modello, ma sicuramente non consentono di proporne uno originale, che derivi dall'analisi dei dati senza la supervisione di alcuna conoscenza a priori.

La potenza delle reti bayesiane risiede invece nel fatto che, oltre alla possibilità di inserire eventuali conoscenze a priori nel processo di estrazione delle informazioni dai dati, è integrata nel modello la possibilità di ricavare delle strutture causali che si basino esclusivamente sulle informazioni contenute nella rilevazione, senza far ricorso ad ipotesi di base sulla configurazione del modello.

Alla luce di questa considerazione, l'induzione diretta della struttura causale si rivela utile non solo nel caso di studi preliminari, ma anche successivamente, quando esiste già una idea di base sulla configurazione delle relazioni, nel caso in cui alcuni meccanismi rilevanti fossero sfuggiti al ricercatore.

Naturalmente, anche nel caso delle reti bayesiane è sempre presente il problema di distinguere le relazioni pure dalla mera covariazione casuale, che si verifica quando esistono relazioni numeriche fra i dati che non corrispondono a relazioni causali nella realtà, ma tale approccio fornisce mezzi di indagine probabilistica molto avanzati per sviluppare un approccio esplorativo abbastanza robusto.

La grande forza di tali strumenti statistici deriva soprattutto dal fatto che essi nascono dalla fusione di proprietà probabilistiche e algoritmi di intelligenza artificiale, che consentono l'estrapolazione di grandi quantità di informazioni in maniera ottimale e la conversione di queste in potere predittivo.

Le reti bayesiane devono il loro nome al fatto che integrano la rappresentazione derivante dalla teoria dei grafi ad un approccio probabilistico, ad indirizzo soggettivo, in cui le relazioni fra le variabili casuali oggetto di analisi vengono modellate in base al sistema delle loro interdipendenze, attraverso una struttura detta grafo

aciclico orientato (DAG) e una distribuzione congiunta fattorizzata in distribuzioni locali che rispecchiano il sistema di interdipendenze, soprattutto l'indipendenza di ciascuna variabile dal resto del sistema, date le variabili che su di essa influiscono direttamente.

La nascita delle reti probabilistiche bayesiane è abbastanza recente. Le prime proposte in merito hanno avuto luogo negli anni '80 in Danimarca, ad opera di un gruppo di ricercatori dell'università di Aalborg e negli Stati Uniti, all'università di UCLA (University of California, Los Angeles) grazie alle ricerche di J.Pearl.

Lo stimolo alla ricerca è nato dalla necessità di integrare l'incertezza negli algoritmi di decisione dei *sistemi esperti* per renderli più realistici e man mano si è sviluppato nella direzione della ricerca della causalità.

Il sistema esperto rientra dell'ambito di ricerca dell'intelligenza artificiale ed è un programma che cerca di riprodurre il comportamento decisionale di esseri umani portatori di un determinato bagaglio conoscitivo in importanti campi della conoscenza, in modo da rendere sempre più automatizzata l'attività decisionale.

L'essere umano però presenta ancora un vantaggio decisivo: è in grado di gestire "naturalmente" l'incertezza e riesce a prendere decisioni anche in assenza di dati completi, in base alla propria esperienza personale. Proprio per tale ragione si è reso necessario indagare sulla possibilità di integrare questa capacità anche nei sistemi esperti e l'approccio statistico più indicato è sembrato da subito quello bayesiano, in cui le conoscenze a priori del decisore sono formalizzate e valutate esplicitamente come fonte di conoscenza.

Le reti probabilistiche bayesiane sono uno strumento principalmente induttivo, in quanto riescono a ricavare conoscenza dai dati grezzi, che siano esperimenti o osservazioni. Le ipotesi a priori possono essere prese in considerazione ma non sono necessarie per il compimento del processo di apprendimento. Infatti, una volta stabilite le variabili oggetto di analisi, il sistema necessita che siano soddisfatte poche condizioni iniziali perché sia possibile effettuare un'analisi che inferisca la struttura causale tramite la costruzione di un grafo aciclico e la struttura probabilistica tramite la determinazione delle probabilità condizionate in cui è possibile fattorizzare la probabilità congiunta delle variabili.

La struttura grafica in rete, data la sua complessità, si adatta meglio a variabili discrete, a numero di stati finiti (il caso che verrà trattato nel presente lavoro); nonostante ciò il modello è stato esteso, non senza difficoltà computazionali, anche al caso di variabili continue. Resta il fatto che le reti probabilistiche bayesiane consentono una più facile implementazione ed una migliore interpretabilità quando le modalità delle variabili analizzate si presentano in numero piuttosto limitato.

Nel seguito del capitolo verranno presentati alcuni basilari richiami formali all'approccio probabilistico bayesiano, dopo del quale saranno esposti la struttura formale delle reti probabilistiche bayesiane nella loro parte grafica e probabilistica,

l'integrazione delle due parti ed infine i metodi di inferenza di entrambe dai dati grezzi.

3.1 Cenni di Inferenza Bayesiana

Come accennato nell'introduzione del capitolo, in questa sede si farà riferimento esclusivamente a variabili aleatorie discrete, la cui distribuzione però, può essere indicizzata da parametri continui. Già nel capitolo 1.1 è stato rilevato come l'interpretazione bayesiana della probabilità, benché soggettiva, venga tranquillamente formalizzata per mezzo della teoria assiomatica di Kolmogorov.

La differenza sostanziale fra approccio soggettivo ed oggettivo riguarda l'utilizzo dell'evidenza empirica a scopo inferenziale ed il trattamento dei parametri da stimare nel modello statistico.

Mentre l'approccio frequentista basa l'inferenza sulla ripetibilità degli eventi analizzati, l'approccio bayesiano fa principalmente ricorso al teorema di Bayes per ricavare nuove informazioni dai dati:

Teorema 3.1 (di Bayes).
Sia $(H_i)_{i\geq 1}$ *una partizione dell'evento certo* Ω *tale che:*

i) $\bigcup_{i=1}^{\infty} H_i = \Omega$

ii) $H_i \cap H_j = \varnothing$, $i \neq j$

iii) $P(H_i) > 0$, $i = 1, \ldots, \infty$.

Sia $E \subseteq \Omega$ *un evento tale che* $P(E) > 0$*, allora, per* $i = 1, \ldots, \infty$*:*

$$P(H_i|E) = \frac{P(E|H_i)\ P(H_i)}{\sum_{j=1}^{\infty} P(E|H_j)\ P(H_j)} \tag{3.1}$$

△

Questa formula assume un ruolo fondamentale nell'interpretazione soggettivista della probabilità perché consente l'*aggiornamento* della fiducia che si aveva nel verificarsi di una determinata ipotesi H_i (in questo approccio identificata con la probabilità assegnata all'ipotesi stessa) in conseguenza del verificarsi dell'evidenza E.

Allo scopo di distinguere il grado di fiducia pre-sperimentale da quello post, si effettua la seguente distinzione:

Definizione 3.1 (Probabilità a priori, Probabilità a posteriori).
Dati gli eventi $H, E \subseteq \Omega$

- si definisce ***probabilità a priori*** *la probabilità che viene attribuita al verificarsi di* H *prima di sapere che si è verificato l'evento* E*, seguendo l'approccio*

bayesiano, tenendo conto delle caratteristiche cognitive del decisore (esperienza, modo di pensare, ecc.)

- si definisce ***probabilità a posteriori*** *la probabilità assegnata ad H, una volta che sia noto E, ovvero l'aggiornamento della probabilità a priori alla luce della nuova evidenza E.* △

Un altro concetto fondamentale richiamato nella formula di Bayes è quello della probabilità dell'evidenza in base all'ipotesi.

Definizione 3.2 (Verosimiglianza).
Siano $H, E \subseteq \Omega$. Si definisce ***verosimiglianza*** *di H dato B, la probabilità condizionata che si verifichi B, se è vera H:*

$$P(E|H) \qquad \triangle$$

In altri termini tale concetto definisce la probabilità che ha l'evento E di verificarsi quando è vera l'ipotesi H.

3.1.1 Le Probabilità Condizionate

Per caratterizzare le relazioni esistenti fra variabili casuali assegnate esistono diversi modelli, uno dei quali (i modelli ad equazioni strutturali) è stato ampiamente discusso del capitolo 2. Un grande difetto dei SEM risiede nel fatto che essi si basano principalmente su modelli lineari. Le relazioni lineari sono spesso delle approssimazioni, più o meno buone, delle relazioni realmente esistenti che, altrimenti, possono essere difficili da formalizzare e trattare analiticamente.

I metodi bayesiani, invece, utilizzano le relazioni di indipendenza condizionale per la modellazione.

A tal fine è opportuno definire i concetti di indipendenza e indipendenza condizionale.

Si consideri il vettore di variabili casuali $\underline{X} = (X_1, X_2, \ldots, X_d)$ definito sullo spazio degli stati $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$, dove $\mathcal{X}_j = \left\{x_1^{(1)}, x_2^{(2)}, \ldots, x_d^{(k_j)}\right\}$ è lo spazio degli stati di X_j per $j = 1, ..., d$.

La funzione di probabilità congiunta $p_{X_1,X_2,...,X_d}$ contiene tutte le informazioni probabilistiche sul vettore casuale $\underline{X}$, ma se si hanno conoscenze a priori sulle relazioni intercorrenti fra le variabili casuali in oggetto, è possibile semplificare tale funzione, decomponendola in fattori più semplici. Il caso di massima decomponibilità si ha in presenza di indipendenza stocastica che implica che la distribuzione congiunta può essere fattorizzata nel prodotto delle sue marginali.

Definizione 3.3 (Indipendenza stocastica).
L'insieme di variabili casuali $\{X_1, X_2, \ldots, X_d\}$ è detto congiuntamente (stocasticamente) indipendente se per ogni vettore casuale $X_{i_1}, X_{i_2}, \ldots, X_{i_m}$, con $i_j \neq i_k \; \forall j \neq k$, vale che:

$$p_{X_{i_1},X_{i_2},\ldots,X_{i_m}} = \prod_{j=1}^{m} p_{X_{i_j}} \qquad \triangle$$

L'assunzione di indipendenza stocastica è molto utile per semplificare la distribuzione congiunta delle variabili, ma al contempo è anche molto forte. Esiste però anche un concetto più debole di dipendenza che permette ugualmente la fattorizzazione.

Definizione 3.4 (Indipendenza condizionale).
Siano $\underline{X}, \underline{Y}$ e $\underline{Z}$ tre vettori casuali aventi probabilità congiunta pari a $P_{\underline{X},\underline{Y},\underline{Z}}$. Siano $\mathcal{X}_X$, $\mathcal{X}_Y$ e $\mathcal{X}_Z$ i rispettivi spazi degli stati.

Si dice che i vettori X ed Y sono condizionatamente indipendenti dato Z, se per ogni $(\underline{x}, \underline{y}, \underline{z}) \in \mathcal{X}_X \times \mathcal{X}_Y \times \mathcal{X}_Z$,

$$p_{\underline{X},\underline{Y},\underline{Z}}(\underline{x}, \underline{y}, \underline{z}) = p_{\underline{X}|\underline{Z}}(\underline{x} \mid \underline{z})\ p_{\underline{Y}|\underline{Z}}(\underline{y} \mid \underline{z})\ p_{\underline{Z}}(\underline{z})$$

Per indicare tale relazione si scrive che: $\underline{X} \perp \underline{Y} \mid \underline{Z}$. $\triangle$

In particolare, dato $V = \{X_1, X_2, \ldots, X_d\}$ un insieme di variabili casuali, per cui vale che $A,\ B, C \subset V$, la notazione $A \perp B \mid C$ indica che ciascuna variabile casuale inclusa in A è indipendente da ogni variabile casuale inclusa in B una volta che sia stata instanziata ogni variabile di C.

Per caratterizzare ulteriormente l'indipendenza condizionale ci si avvale dei seguenti teoremi:

Teorema 3.2.
Le seguenti proposizioni sono tutte equivalenti a $\underline{X} \perp \underline{Y} \mid \underline{Z}$:

1) Per ogni $(\underline{x}, \underline{y}, \underline{z}) \in \mathcal{X}_X \times \mathcal{X}_Y \times \mathcal{X}_Z$ tale che $p_{\underline{Y}|\underline{Z}}(\underline{y} \mid \underline{z}) > 0$ e $p_{\underline{Z}}(\underline{z}) > 0$ vale che:

$$p_{\underline{X}|\underline{Y},\underline{Z}}(\underline{x} \mid \underline{y}, \underline{z}) = p_{\underline{X}|\underline{Z}}(\underline{x} \mid \underline{z})\ p_{\underline{Y}|\underline{Z}}(\underline{y} \mid \underline{z}).$$

2) Quando esiste una funzione $a : \mathcal{X}_X \times \mathcal{X}_Y \to [0, 1]$ tale che per ogni $(\underline{x}, \underline{y}, \underline{z}) \in \mathcal{X}_X \times \mathcal{X}_Y \times \mathcal{X}_Z$ e vale che sia $p_{\underline{Y}|\underline{Z}}(\underline{y} \mid \underline{z}) > 0$ sia $p_{\underline{Y}|\underline{Z}}(\underline{y} \mid \underline{z}) > 0$, allora:

$$p_{\underline{X}|\underline{Y},\underline{Z}}(\underline{x} \mid \underline{y}, \underline{z}) = a(\underline{x}, \underline{z}).$$

3) Quando esistono due funzioni $a : \mathcal{X}_X \times \mathcal{X}_Z \to \mathbb{R}$ e $b : \mathcal{X}_Y \times \mathcal{X}_Z \to \mathbb{R}$ tali che per ogni $(\underline{x}, \underline{y}, \underline{z}) \in \mathcal{X}_X \times \mathcal{X}_Y \times \mathcal{X}_Z$ e vale che $p_{\underline{Z}}(\underline{z}) > 0$, allora:

$$p_{\underline{X},\underline{Y}|\underline{Z}}(\underline{x}, \underline{y} \mid \underline{z}) = a(\underline{x}, \underline{z})\ b(\underline{y}, \underline{z}).$$

4) Per ogni $(\underline{x}, \underline{y}, \underline{z}) \in \mathcal{X}_X \times \mathcal{X}_Y \times \mathcal{X}_Z$ tale che $p_{\underline{Z}}(\underline{z}) > 0$, vale che:

$$p_{\underline{X},\,\underline{Y},\,\underline{Z}}\left(\underline{x},\ \underline{y},\ \underline{z}\right) = \frac{p_{\underline{X},\,\underline{Z}}\left(\underline{x},\ \underline{z}\right)\ p_{\underline{Y},\,\underline{Z}}\left(\underline{y} \mid \underline{z}\right)}{p_{\underline{Z}}\left(\underline{z}\right)}$$

5) Quando esistono due funzioni $a : \mathcal{X}_X \times \mathcal{X}_Z \to \mathbb{R}$ *e* $b : \mathcal{X}_Y \times \mathcal{X}_Z \to \mathbb{R}$ *allora:*

$$p_{\underline{X},\,\underline{Y},\,\underline{Z}}\left(\underline{x},\ \underline{y},\ \underline{z}\right) = a\left(\underline{x},\ \underline{z}\right)\ b\left(\underline{y},\ \underline{z}\right). \qquad \triangle$$

Per la dimostrazione si veda Koski e Noble (2009).

Teorema 3.3 (Proprietà della Indipendenza Condizionale).
Per $\underline{X}$, $\underline{Y}$ *e* $\underline{Z}$ *(sottoinsiemi distinti di V) vale che:*

(Proprietà di SIMMETRIA)

$$\underline{X} \perp \underline{Y} \mid \underline{Z} \Leftrightarrow \underline{Y} \perp \underline{X} \mid \underline{Z}.$$

inoltre, se $\underline{X} \perp \underline{Y} \mid \underline{Z}$ *allora:*

(Proprietà di DECOMPOSIZIONE)

$$\underline{X} \perp (\underline{Y} \cup \underline{W}) \mid \underline{Z} \Rightarrow \underline{X} \perp \underline{Y} \mid \underline{Z} \wedge \underline{X} \perp \underline{W} \mid \underline{Z}.$$

(Proprietà di UNIONE DEBOLE)

$$\underline{X} \perp (\underline{Y} \cup \underline{W}) \mid \underline{Z} \Rightarrow \underline{X} \perp \underline{Y} \mid (\underline{Z} \cup \underline{W})$$

(Proprietà di CONTRAZIONE)

$$\underline{X} \perp \underline{Y} \mid \underline{Z} \wedge \underline{X} \perp \underline{W} \mid (\underline{Z} \cup \underline{Y}) \Rightarrow \underline{X} \perp (\underline{Y} \cup \underline{W}) \perp \underline{Z}.$$

Inoltre, se $p_{\underline{X},\,\underline{Y},\,\underline{Z}}\left(\underline{x},\ \underline{y},\ \underline{z}\right)$ *è una funzione strettamente positiva, vale che*

(Proprietà di INTERSEZIONE)

$$\underline{X} \perp \underline{Y} \mid (\underline{Z} \cup \underline{W}) \wedge \underline{X} \perp \underline{W} \mid (\underline{Z} \cup \underline{Y}) \Rightarrow \underline{X} \perp (\underline{Y} \cup \underline{W}) \perp \underline{Z}. \qquad \triangle$$

Per la dimostrazione si veda Pearl (1988).

Sotto opportune condizioni tali relazioni di dipendenza ed indipendenza statistica possono essere riassunte tramite una rappresentazione grafica, quale il grafo aciclico orientato, che di seguito viene presentata.

3.1.2 L'aggiornamento delle probabilità a priori

Il teorema di Bayes è uno strumento di aggiornamento delle probabilità a priori in relazione alle nuove evidenze, ponendo la probabilità a posteriori proporzionale al prodotto di probabilità a priori e verosimiglianza.

Se le ipotesi considerate sono modellate da una variabile casuale discreta Y, come pure le evidenze empiriche X, caratterizzate da funzione di probabilità rispettivamente pari a p_Y p_X, e se $p_{Y|X}$ e $p_{X|Y}$ sono le distribuzioni condizionate di Y dato X e viceversa, allora è possibile riscrivere il teorema di Bayes in questi termini:
per ogni y ed x vale che:

$$p_{Y|X}(y|x) = \frac{p_Y(y)\, p_{X|Y}(x|y)}{p_X(x)}$$

Se le variabili casuali Y ed X sono entrambe continue con funzione di densità pari a π_Y e π_X, vale che, per $X = x$:

$$\pi_{Y|X}(y|x) = \frac{\pi_Y(y)\, \pi_{X|Y}(x|y)}{\pi_X(x)}$$

Infine, nel caso si abbia a che fare con una variabile X discreta con funzione di probabilità pari a p_X ed una variabile Θ continua, avente funzione di densità π_Θ con spazio degli stati $\widetilde{\Theta}$ e tali che $p_{X|\Theta}(x|\theta)$ sia la funzione di probabilità condizionata di X dato Θ, allora la formula di Bayes diventa:

$$\pi_{\Theta|X}(\theta|x) = \frac{\pi_\Theta(\theta)\, p_{X|\Theta}(x|\theta)}{\int_{\widetilde{\Theta}} \pi_\Theta(\theta)\; p_{X|\Theta}(x|\theta)} = \frac{\pi_\Theta(\theta)\, p_{X|\Theta}(x|\theta)}{p_X(x)} \tag{3.2}$$

Quest'ultimo caso costituisce un esempio fondamentale dell'approccio bayesiano all'inferenza statistica. Se si prende in considerazione una variabile casuale discreta X, di cui si conosce la distribuzione a meno di un parametro θ, secondo l'approccio bayesiano, è possibile modellare l'incertezza su tale parametro rappresentandolo con una variabile casuale continua Θ che assume valori nello spazio degli stati $\widetilde{\Theta}$ costituito da tutti i valori ammissibili per il parametro cercato.

A questo punto la probabilità a priori dell'ipotesi può essere rappresentata tramite la funzione di densità a priori $\pi_\Theta(\theta)$, mentre le probabilità a posteriori che ottengo in seguito al verificarsi di $X = x$ (evidenza empirica) possono essere sintetizzate in $\pi_{\Theta|X}(\theta|x)$.

Un interessante utilizzo dell'inferenza bayesiana è dato dalla sua estensione in ambito predittivo. Si consideri un campionamento casuale semplice di ampiezza N dalla variabile casuale discreta X, il vettore casuale $\underline{X}_{(N)} = (X_1, \ldots, X_N)$, associato alle possibili realizzazioni campionarie è costituito da variabili casuali indipendenti ed identicamente distribuite. La singola realizzazione rappresentata dal vettore $\underline{x}_{(N)} = (x_1, \ldots, x_N)$ può essere considerato un'evidenza da utilizzare per aggiornare le ipotesi sulla distribuzione di X_{n+1}, ovvero calcolare la distribuzione condizionata di X_{n+1} dato che si sono verificate le evidenze $\underline{X}_{(N)} = \underline{x}_{(N)}$.

Poiché le variabili in questione sono tutte indipendenti, indicando con

$$\underline{X}_{(N+1)} = (X_1, \ldots, X_N, X_{N+1})$$

il vettore casuale complessivo e con

$$\underline{x}_{(N+1)} = \left(x_1, \ldots, x_N, x_{(N+1)}\right)$$

il vettore delle evidenze, è possibile ricavare la distribuzione condizionata

$$p_{X_{N+1}|\underline{X}_{(N)}}\left(x_{N+1}|\underline{x}_{(N)}\right)$$

che descrive la connessione tra $\underline{x}_{(N)}$ ed $\underline{x}_{N+1}$:

$$p_{X_{N+1}|\underline{X}_{(N)}}\left(x_{N+1}|\underline{x}_{(N)}\right) = \frac{p_{X_{(N+1)}}\left(\underline{x}_{(N+1)}\right)}{p_{X_{(N)}}\left(\underline{x}_{(N)}\right)} \tag{3.3}$$

Come è stato già anticipato, è possibile ipotizzare di conoscere la distribuzione di X a meno di un parametro che viene associato alla variabile casuale Θ e scomporre la distribuzione congiunta di $\underline{X}_{(N)}$ come segue:

$$p_{X_{(N)}}\left(\underline{x}_{(N)}|\theta\right) = \prod_{i=1}^{N} q_{X_i}\left(x_i|\theta\right) \tag{3.4}$$

dove q_{X_i} è la funzione di verosimiglianza di θ dato x_i. A questo punto si ottiene che:

$$p_{X_{(N)}}\left(\underline{x}_{(N)}\right) = \int_{\tilde{\Theta}} \prod_{i=1}^{N} q_{X_i}\left(x_i|\theta\right) d\theta \tag{3.5}$$

Definizione 3.5 (Distribuzione Predittiva a Priori).
Si definisce come ***distribuzione predittiva a priori*** *la distribuzione $p_{X_{(N)}}$ per la collezione di variabili casuali $X_{(N)}$ (per le quali $x_{(N)}$ risulta essere una realizzazione).* △

De Finetti (1937) dimostra che se $x_{(N)}$ è infinitamente scambiabile e appartiene ad uno spazio con opportune proprietà, la struttura che si ricava per $p_{X_{(N)}}$ dall'equazione 3.5 è la sola possibile.

Se il risultato ottenuto con l'equazione 3.5 viene integrato con l'equazione 3.3 si ottiene che:

$$\begin{aligned} p_{X_{N+1}|\underline{X}_{(N)}}\left(x_{N+1}|\underline{x}_{(N)}\right) &= \frac{\int_{\widetilde{\Theta}} \prod_{i=1}^{N+1} q_{X_i}\left(x_i|\theta\right) \pi_{\Theta}\left(\theta\right) d\theta}{\int_{\widetilde{\Theta}} \prod_{i=1}^{N} q_{X_i}\left(x_i|\theta\right) \pi_{\Theta}\left(\theta\right) d\theta} \\ &= \int_{\widetilde{\Theta}} q_{X_{N+1}}\left(x_{N+1}|\theta\right) \frac{\prod_{i=1}^{N} q_{X_i}\left(x_i|\theta\right) \pi_{\Theta}\left(\theta\right)}{\int_{\widetilde{\Theta}} \prod_{i=1}^{N} q_{X_i}\left(x_i|\theta\right) \pi_{\Theta}\left(\theta\right) d\theta} d\theta \end{aligned} \tag{3.6}$$

La densità di probabilità condizionale di Θ dato $\underline{X}_{(N)} = \underline{x}_{(N)}$ si può ancora ottenere con la formula di Bayes:

$$\pi_{\Theta|\underline{X}_{(N)}}\left(\theta|\underline{x}_{(N)}\right) = \frac{\prod_{i=1}^{N} q_{X_i}\left(x_i|\theta\right) \pi_{\Theta}\left(\theta\right)}{\int_{\widetilde{\Theta}} \prod_{i=1}^{N} q_{X_i}\left(x_i|\theta\right) \pi_{\Theta}\left(\theta\right) d\theta} \tag{3.7}$$

da cui segue direttamente che:

$$p_{X_{N+1}|\underline{X}_{(N)}}\left(x_{N+1}|\underline{x}_{(N)}\right) = \int_{\widetilde{\Theta}} q_{X_{N+1}}\left(x_{N+1}|\theta\right) \pi_{\Theta|\underline{X}_{(N)}}\left(\theta|\underline{x}_{(N)}\right). \tag{3.8}$$

che consente di calcolare le probabilità per una nuova osservazione. Questa impostazione è strettamente legata con il concetto di causalità, soprattutto nell'imposta previsiva. Seguendo questa linea di pensiero risulta quindi possibile effettuare previsioni sul futuro, ammettendo che questo dipenda dal passato.

Inferenza Bayesiana e Distribuzioni Coniugate

Come è stato già anticipato è possibile costruire un modello statistico per l'inferenza bayesiana su una determinata caratteristica di una popolazione assegnandole una variabile casuale X che assume valore nello spazio degli stati $\widetilde{X}$ e alla quale è associata una distribuzione indicizzata dal parametro $\theta \in \widetilde{\Theta}$.

Se si assume di effettuare un campionamento bernulliano dalla variabile casuale X, si ottiene un vettore casuale $\underline{X} = X_1, \ldots, X_N$ avente distribuzione $p_{\underline{X}_N}\left(\underline{x}_N\right)$ derivante dal fatto che, subordinatamente alla conoscenza di θ, le variabili casuali associate a ciascuna realizzazione sono indipendenti ed identicamente distribuite.

Effettuare inferenza sul parametro θ, nell'ottica bayesiana, consiste nell'aggiornare le conoscenze a priori incorporate nella distribuzione iniziale $p_\Theta(\theta)$ in base all'evidenza $\underline{X}_N$, utilizzando ancora il teorema di Bayes nella versione dell'equazione 3.2, si ottiene che:

$$\pi_{\Theta|X}(\theta|x) = \frac{\pi_\Theta(\theta)\prod_{i=1}^{N} p_{X|\Theta}(x_i|\theta)}{\int_{\widetilde{\Theta}} \prod_{i=1}^{N} \pi_\Theta(\theta)\, p_{X|\Theta}(x_i|\theta) d\theta} \tag{3.9}$$

La soggettività dell'approccio è racchiusa nella scelta della distribuzione a priori da attribuire al parametro θ.

Le prime applicazioni dell'inferenza bayesiana hanno fatto riferimento soprattutto a distribuzioni *non informative*, ovvero distribuzioni che partono dalla constatazione di mancanza di informazioni a priori e che quindi tendono a dare il peso massimo all'informazione apportata dalle evidenze empiriche.

Tale scelta non prende in considerazione solo la mancanza di informazioni, ma, soprattutto all'inizio, è stata favorita anche dal fatto che molte delle distribuzioni a priori non informative consentono un aggiornamento che non implica un'eccessiva difficoltà computazionale.

L'integrale al denominatore dell'equazione 3.9 è spesso di difficile risoluzione analitica e l'esistenza di formule di calcolo facilitate ha costituito per lungo tempo l'unico metodo per utilizzare l'inferenza bayesiana, finché non sono stati introdotti metodi di ricampionamento e metodi iterativi, quali le Catene di Markov Monte Carlo (MCMC)[1].

Esiste un'altra particolare classe di distribuzioni a priori dette *distribuzioni a priori coniugate al modello* che si caratterizzano per il fatto che se la distribuzione iniziale appartiene a tale classe, anche la distribuzione finale vi appartiene e l'aggiornamento della fiducia si riduce alla modifica dei parametri della distribuzione a priori.

Una famiglia di distribuzioni coniugate impiegata spesso nell'inferenza bayesiana in ragione delle sue peculiari caratteristiche è la famiglia di distribuzioni di Dirichlet.

Definizione 3.6 (Distribuzione di Dirichlet).
La distribuzione di Dirichlet, indicata cone $Dir(a_1, \ldots, a_k)$ *è definita come:*

$$\pi(\theta_1, \ldots, \theta_k) = \begin{cases} \dfrac{\Gamma(a_1 + \ldots + a_k)}{\prod_{j=1}^{k} \Gamma(a_k)} \left(\prod_{j=1}^{k} \theta_j^{a_j - 1} \right) & \theta_j \geq 0, \sum_{j=1}^{k} \theta_j = 1 \\ 0 & altrove \end{cases} \tag{3.10}$$

△

[1] Monte Carlo Markov Chain.

Non solo tale distribuzione è la generalizzazione multivariata della *Beta*, la quale si ottiene per $k = 2$, ma soprattutto possiede la proprietà di essere una coniugata del modello multinomiale.

Proprietà 3.1.
Sia $\underline{X}$ un vettore casuale caratterizzato da una distribuzione multinomiale di vettore di parametri $\underline{\theta} = (\theta_1, ..., \theta_k)$ a cui sia a sua volta associata una variabile aleatoria Θ che si distribuisce come una $Dir\,(a_1, \ldots, a_k)$

Alla luce di un'evidenza empirica $\underline{x}_k$ la distribuzione a posteriori del vettore dei parametri $\underline{\Theta}$ è ancora una Dirichlet, ma con parametri aggiornati nel seguente modo:

$$p_{\Theta|X} \sim Dir\,(a_1 + x_1, \ldots, a_k + x_k) \tag{3.11}$$

$\triangle$

Per la dimostrazione di veda Lidstone (1920).

3.1.3 Sufficienza Bayesiana

Le statistiche sufficienti sono uno strumento essenziale nell'inferenza e sono di grande utilità nell'ambito delle reti bayesiane per la loro relazione con l'indipendenza condizionale.

Si torni a considerare il vettore $\underline{X} = X_1, \ldots, X_n$ di variabili casuali e si determini una rilevazione di ampiezza N in modo da ottenere la matrice aleatoria $\boldsymbol{X}$ di dimensioni $N \times n$ in cui le righe rappresentano vettori aleatori discreti indipendenti e identicamente distribuiti.

Sia $\underline{\Theta}$ una variabile aleatoria continua che rappresenta lo sconosciuto vettore dei parametri.

Si supponga che, condizionatamente a $\underline{\Theta} = \underline{\theta}$, la matrice $\boldsymbol{X}$ abbia distribuzione di probabilità condizionale pari a $p_{\boldsymbol{X}|\underline{\Theta}}\,(x|\theta)$.

Si supponga inoltre che $\underline{\Theta}$ abbia densità a priori pari a $\pi_{\underline{\Theta}}\,(\theta)$.

Sia infine t una *statistica* di $\boldsymbol{X}$ tale che $t = t\,(\boldsymbol{X})$, dove per statistica si intende sia una funzione (o una serie di algoritmi statistici) definita su un campione casuale, dove la funzione stessa è indipendente dalla distribuzione del campione sia la sua applicazione ad una realizzazione, avendo cura di indicate con la lettera maiuscola T la variabile casuale e con la minuscola t la realizzazione.

Un particolare tipo di statistica che possiede opportune proprietà inferenziali, è rappresentato dalle statistiche sufficienti. La definizione di sufficienza, in ambito bayesiano, è la seguente:

Definizione 3.7 (Sufficienza bayesiana).
Sia T una statistica definita come $T = t(\boldsymbol{X})$, tale che per ogni possibile densità a priori $\pi_{\underline{\Theta}}$ esiste una funzione ϕ tale che:

$$\pi_{\underline{\Theta}|\boldsymbol{X}}(\underline{\theta}|\boldsymbol{x}) = \frac{p_{\boldsymbol{X}|\underline{\Theta}}(x|\theta)\ \pi_{\underline{\Theta}}(\theta)}{p_{\boldsymbol{X}}(\boldsymbol{x})} = \phi(\underline{\theta}, t(\boldsymbol{x})) \tag{3.12}$$

allora si dice che T è una statistica bayesiana sufficiente *per $\underline{\Theta}$.* △

La caratteristica principale di una statistica sufficiente è che essa contiene tutte le informazioni rilevanti per l'inferenza sui parametri poichè la distribuzione a posteriori dipende da $\boldsymbol{X}$ solo attraverso T. Per tale ragione essa viene denominata anche *riassunto esaustivo*.

Un'importante conseguenza della sufficienza statistica è che l'indipendenza condizionale di $\boldsymbol{X}$ e $\underline{\Theta}$, dato $t(\boldsymbol{X})$ implica la sufficienza statistica.

Teorema 3.4.
Data la funzione t, sia $T = t(\boldsymbol{X})$. Se vale che:

$$\boldsymbol{X} \perp \underline{\Theta}|T, \tag{3.13}$$

allora $T = t(\boldsymbol{X})$ è una statistica bayesiana sufficiente per $\underline{\Theta}$. △

Per la dimostrazione vedere Koski e Noble (2009).

Bisogna rilevare che il viceversa vale solo se le famiglie delle misure di probabilità hanno spazio dei parametri di dimensione finita.

Sufficienza predittiva

È interessante considerare l'estensione della sufficienza bayesiana all'ambito predittivo.

Si prenda in considerazione il vettore aleatorio discreto $\underline{X}$ e si consideri $\underline{Y}$ indifferentemente vettore o variabile aleatoria discreta a sua volta. Il problema di predire $\underline{Y}$, conoscendo $\underline{X}$ è stato già affrontato, ma in questa sede si vuole estendere il caso alla possibilità di effettuare la predizione in base ad una funzione t di $\underline{X}$ senza utilizzare l'intero campione. La risposta al problema viene data dal seguente teorema.

Teorema 3.5.

$$\underline{X} \perp (\underline{Y}, \underline{\Theta})\,|T \Leftrightarrow \begin{cases} \underline{X} \perp \underline{\Theta}|T \\ \\ \underline{X} \perp \underline{Y}|\,(\underline{\Theta}, T) \end{cases} \tag{3.14}$$

△

da cui si ricava che sotto opportune condizioni $(\underline{Y}, T)$ è sufficiente in senso bayesiano per $\underline{\Theta}$. Inoltre vale che:

Teorema 3.6.
Sia t una funzione e sia $T = t(\boldsymbol{X})$.
Se $\underline{X}, \underline{Y}, \underline{\Theta}, T$ soddisfano $\underline{X} \perp (\underline{Y}, \underline{\Theta}) \,|T$, allora:

$$\pi_{\underline{\Theta}|\underline{X},\underline{Y},T}\left(\underline{\theta}|\underline{x}, \underline{y}, t\right) = \pi_{\underline{\Theta}|\underline{Y},T}\left(\underline{\theta}|\underline{y}, t\right) \tag{3.15}$$

△

Per la dimostrazione di entrambi i teoremi si veda (Koski e Noble, 2009).

Questa equazione si rivelerà utile per le reti probabilistiche bayesiane.

3.2 Reti Probabilistiche Bayesiane

Dopo aver brevemente introdotto gli spetti fondamentali dell'inferenza bayesiana vengono presentate le reti bayesiane, le quali integrano tale approccio statistico con la teoria dei grafi. Infatti le reti probabilistiche bayesiane si avvalgono di due strumenti distinti: la distribuzione congiunta delle variabili oggetto di analisi (o per essere più precisi la sua scomposizione in distribuzioni di indipendenza condizionale) e una rappresentazione grafica delle relazioni tra di esse intercorrenti.

Le reti bayesiane non sono nate con lo scopo di modellare relazioni causali ma la loro struttura si è dimostrata molto utile in tal senso e ben presto esse hanno assunto un ruolo molto importante nell'ambito della modellazione della causalità per due ragioni fondamentali:

1) Le relazioni di dipendenza ed indipendenza probabilistica, che come si è visto rappresentano un importante approccio all'interpretazione causale, vengono ben modellate da queste reti che formalizzano relazioni probabilistiche senza bisogno di assumere la linearità delle stesse.

2) I sistemi di apprendimento delle reti sono in grado, sotto condizioni piuttosto generali, di ricavare la struttura grafica e le relazioni di dipendenza direttamente dai dati grezzi, anche senza alcuna conoscenza a priori dell'ambito di ricerca.

Queste caratteristiche hanno fatto delle reti bayesiame degli strumenti molto utilizzati nel campo dell'intelligenza artificiale, ma in generale in tutti i campi in cui si applicano le teorie decisionali.

Al fine di mostrare come queste riescano ad ottenere dei risultati tanto interessanti ne vengono presentati i meccanismi statistici sottostanti. Per quanto riguarda poi le tecniche di apprendimento, vengono presentati solo alcuni esempi tra gli innumerevoli algoritmi che vengono continuamente sviluppati per migliorare le prestazioni in fase di apprendimento delle reti.

3.3 Il Modello Grafico

Le reti probabilistiche bayesiane si compongono di due strumenti atti a rappresentare le relazioni di indipendenza condizionale. Uno di questi è lo strumento grafico che viene introdotto nel presente paragrafo insieme alle sue caratteristiche fondamentali.

Definizione 3.8 (Grafo, Grafo Semplice).
Un Grafo $\mathcal{G} = (V, E)$ *consiste di un insieme finito di nodi* V *e un insieme* E *di archi, dove ogni arco è contenuto in* $V \times V$*. L'insieme degli archi consiste di coppie ordinate di nodi.*

Sia $V = v_1, \ldots, v_n$ *con* $i = 1, \ldots, n$*. Si dice che un grafo è* semplice *se* E *non contiene archi della forma* $(v_j,\ v_j)$ *(un ciclo dal nodo a se stesso) e ciascun arco* $(v_j,\ v_k) \in E$ *appare solo una volta. In altre parole non sono permessi archi multipli.*

Per ogni coppia di nodi distinti $v_j,\ v_k \in E$*, la coppia ordinata* $(v_j,\ v_k) \in E$ *sse esiste un arco orientato diretto da* v_j *a* v_k*. Un arco non orientato viene indicato con* $\langle v_j,\ v_k \rangle$*.*

In termini di archi orientati:
$\langle v_j,\ v_k \rangle \in E \Leftrightarrow (v_j,\ v_k) \in E \ \wedge\ (v_k,\ v_j) \in E$*.*

Per un Grafo Semplice che contiene sia archi orientati che non orientati, l'insieme di tutti gli archi E *può essere scomposto come* $E = I \cup D$ *con* $I \cap D = \varnothing$*, dove* I *e* D *sono definiti come:*
$\langle v_j,\ v_k \rangle \in I \Leftrightarrow (v_j,\ v_k) \in E \ \wedge\ (v_k,\ v_j) \in E$*.*
$(v_j,\ v_k) \in D \Leftrightarrow (v_j,\ v_k) \in E \ \wedge\ (v_k,\ v_j) \notin E$*.* △

Nel prosieguo con il termine grafo si farà riferimento esclusivamente al grafo semplice.

La coppia $\langle v_j,\ v_k \rangle \in I$ viene rappresentata con un arco semplice congiungente v_j e v_i (v_j—v_i).

La coppia $(v_j,\ v_k) \in D$ viene rappresentata con una freccia che parte da v_j e arriva a v_i ($v_j \rightarrow v_i$).

Per quanto riguarda i nodi che compongono un grafo, è possibile definire alcune relazioni tipiche.

Definizione 3.9 (nodi Adiacenti, Genitori, Figli, Vicini).
Sia il grafo $\mathcal{G} = (V, E)$ *talle che* $V = v_1, \ldots, v_n$ *con* $i = 1, \ldots, n$ *e l'insieme degli archi* $E = I \cup D$*, con* D*, l'insieme degli archi orientati, definito da:*
$(v_j,\ v_k) \in D \Leftrightarrow (v_j,\ v_k) \in E \ \wedge\ (v_k,\ v_j) \notin E$*.*
E con I*, l'insieme degli archi non orientati, definito da:*
$\langle v_j,\ v_k \rangle \in I \Leftrightarrow (v_j,\ v_k) \in E \ \wedge\ (v_k,\ v_j) \in E$*.*
Siano $v_k,\ v_j \in \mathcal{G}$*.*

Se $(v_k,\ v_j) \in E \vee (v_j,\ v_k) \in E$, *allora i nodi* v_k *e* v_j *vengono detti* ***adiacenti****.*
Se $(v_k,\ v_j) \in D$, *allora* v_k *viene detto* ***genitore*** *di* v_j *e* v_j *viene detto* ***figlio*** *di* v_k*.*
Per ogni nodo $v \in V$*,*

*l'****insieme dei nodi adiacenti*** *è definito come:*
$Ad\,(v) = \{v_i \in V : (v_k,\ v_j) \in E \vee (v_j,\ v_k) \in E\}$*;*

*l'****insieme dei genitori*** *è definito come:*
$\Pi\,(v) = \{v_i \in V : (v_i,\ v) \in D\}$*;*

*l'****insieme dei figli*** *è definito come:*
$Ch\,(v) = \{v_i \in V : (v,\ v_i) \in D\}$*;*

*l'****insieme dei vicini diretti*** *è definito come:*
$N_{(d)}\,(v) = \Pi\,(v) \cup Ch\,(v)$*;*

*l'****insieme dei vicini indiretti*** *è definito come:*
$N_{(i)}\,(v) = \{v_i \in V : \langle v,\ v_i \rangle\}$*;*
Per ogni sottoinsieme di nodi $A \supseteq V$

*l'****insieme dei genitori*** *di* A *è definito come:*
$\Pi\,(A) = \bigcup_{v \in A} \{v_i \in V \backslash A : (v_i,\ v) \in D\}$*;*

*l'****insieme dei figli*** *di* A *è definito come:*
$Ch\,(v) = \bigcup_{v \in A} \{v_i \in V \backslash A : (v,\ v_i) \in D\}$*;*

*l'****insieme dei vicini diretti*** *di* A *è definito come:*
$N_{(d)}\,(v) = \Pi\,(A) \cup Ch\,(A)$*;*

*l'****insieme dei vicini indiretti*** *di* A *è definito come:*
$N_{(i)}\,(A) = \bigcup_{v \in A} \{v_i \in V \backslash A : \langle v,\ v_i \rangle\}$*;*

*l'****insieme dei vicini*** *di un nodo* v *è definito come:*
$N\,(v) = N_{(d)}\,(v) \cup N_{(i)}\,(v)$*;*

per indicare che v_i *è un vicino di* v *si scrive che* $v_i \sim v$ $\triangle$

Definizione 3.10 (Coniugi).
Dato il grafo $\mathcal{G} = (V, E)$, dove $V = v_1, \ldots, v_n$ con $i = 1, \ldots, n$ e l'insieme degli archi $E = I \cup D$, con D, l'insieme degli archi orientati e con I, l'insieme degli archi non orientati.

Siano $v_k,\ v_j \in \mathcal{G}$.

*Se esiste almeno un $v_i \in \mathcal{G}$ t.c. $v_k \in \Pi(v_i) \wedge v_j \in \Pi(v_i)$, allora i nodi v_k e v_j vengono detti **coniugi**.*

*L'**insieme dei coniugi** di un nodo viene indicato con $Cn(v)$* △

Due nodi coniugi sono tali se hanno almeno un nodo figlio in comune

Definizione 3.11 (Grafo Orientato, Non Orientato).
*Se tutti gli archi di un grafo sono orientati, allora il grafo sarà definito **orientato**. Se tutti gli archi di un grafo sono non orientati, allora il grafo sarà definito **non orientato**.*

La versione non orientata di un grafo $\mathcal{G}$ si ottiene rimpiazzando tutti gli archi orientati di $\mathcal{G}$ con i corrispondenti archi non orientati. Tale versione viene denominata $\tilde{\mathcal{G}}$ △

Definizione 3.12 (Tragitto (Trail), Grafo Connesso).
Dato il grafo $\mathcal{G} = (V, E)$, dove $V = v_1, \ldots, v_n$ con $i = 1, \ldots, n$ e l'insieme degli archi $E = I \cup D$ e $I \cap D = \varnothing$, con D, l'insieme degli archi orientati e con I, l'insieme degli archi non orientati.

*Si definisce **tragitto** (trail) tra due nodi $v_i, v_j \in V$ come una collezione di nodi $\tau = (\tau_1, \ldots, \tau_m)$ con $\tau_i \in V\ \forall i = 1, \ldots, m$ e con $\tau_1 = v_i$ e $\tau_m = v_j$ e tali che $\forall i = 1, \ldots, m-1$, $\tau_i \sim \tau_{i+1}$.*

Tutto ciò equivale a dire che $\forall i = 1, \ldots, m-1$, $\boldsymbol{o}(\tau_i, \tau_{i+1}) \in D$, $\boldsymbol{o}(\tau_i + 1, \tau_i) \in D$, $\boldsymbol{o}\langle \tau_i, \tau_{i+1} \rangle \in I$, ovvero che (τ_i, τ_{i+1}) sono adiacenti $\forall i = 1, \ldots, m-1$.

*Si dice che un grafo $\mathcal{G} = (V, E)$ è **connesso** se per ogni coppia di nodi $v_i, v_j \in V$ esiste almeno un tragitto che li collega.* △

Definizione 3.13 (Traiettoria (Path), Traiettoria orientata).
Dato il grafo $\mathcal{G} = (V, E)$, dove $V = v_1, \ldots, v_n$ con $i = 1, \ldots, n$ e l'insieme degli archi $E = I \cup D$ e $I \cap D = \varnothing$, con D, l'insieme degli archi orientati e con I, l'insieme degli archi non orientati.

*Si definisce **traiettoria** (path) di lunghezza m dal nodo $v_i \in V$ al nodo , $v_j \in V$ la sequenza di nodi distinti $\tau = (\tau_0, \ldots, \tau_m)$ con $\tau_i \in V\ \forall i = 1, \ldots, m$ e con $\tau_0 = v_i$ e $\tau_m = v_j$ e tali che $\forall i = 1, \ldots, m-1$, $\tau_{i-1}, \tau_i \in E$.*

Tutto ciò equivale a dire che $\forall i = 1, \ldots, m-1$, $\boldsymbol{o}(\tau_{i-1}, \tau_i) \in D$, $\boldsymbol{o}\langle \tau_i, \tau_{i+1} \rangle \in I$,

*La traiettoria è detta **orientata** se $(\tau_{i-1}, \tau_i) \in D\ \forall i = 1, \ldots, m$ ovvero non ci sono archi non orientati in una traiettoria orientata.* △

Dalla precedente definizione segue che un *tragitto* in $\mathcal{G}$ è una sequenza di nodi che formano una *traiettoria* nella versione non orientata $\tilde{\mathcal{G}}$.

Definizione 3.14 (Discendenti, Antenati).
Dato il grafo $\mathcal{G} = (V, E)$, dove $V = v_1, \ldots, v_n$ con $i = 1, \ldots, n$ e l'insieme degli archi $E = I \cup D$, con D, l'insieme degli archi orientati e con I, l'insieme degli archi non orientati;

*il nodo v_i è **discendente** del nodo v_j sse esiste una traiettoria orientata da v_j verso v_i;*

*l'**insieme dei discendenti** del nodo v_j, indicato con $D(v_j)$, viene definito come:*
$D(v_j) =$
$= \{v_i \in D : \exists \tau = (\tau_0, \ldots, \tau_k)\, t.c. \tau_0 = v_i, \tau_m = v_j, (\tau_h, \tau_{h+1}) \in D, h = 0, \ldots, k\}$

*il nodo v_k è **antenato** del nodo v_j sse esiste una traiettoria orientata da v_k verso v_j.*

*l'**insieme degli antenati** del nodo v_j, indicato con $A(v_j)$, viene definito come:*
$A(v_j) =$
$= \{v_k \in D : \exists \tau = (\tau_0, \ldots, \tau_k)\ \ t.c.\ \tau_0 = v_k, \tau_m = v_j, (\tau_h, \tau_{h+1}) \in D, h = 0, \ldots, k\}$

in entrambi i casi sono ammessi esclusivamente archi orientati. △

Definizione 3.15 (Ciclo).
*Dato il grafo $\mathcal{G} = (V, E)$, si definisce $m-$**ciclo** in $\mathcal{G}$ una sequenza di nodi distinti $\tau_0, \ldots, \tau_{m-1}$ tali che $\tau_0, \ldots, \tau_{m-1}, \tau_0$ costituisca una traiettoria.* △

Questa definizione è fondamentale per la caratterizzazione dei grafi impiegati nelle reti bayesiane ovvero per i Grafi Aciclici Orientati (DAG)[2].

Definizione 3.16 (Grafi Aciclici Orientati (DAG)).
Dato il grafo $\mathcal{G} = (V, E)$, si dice che $\mathcal{G}$ è un Grafo Aciclico Orientato (DAG) se:

- ogni arco del grafo è orientato, ovvero $\mathcal{G}$ è un grafo semplice t.c. per ogni coppia di nodi $v, u \in V \times V$, vale che $(v, u) \in E \Rightarrow (u, v) \notin E$

- per ogni $v \in V$ non esiste alcun insieme di nodi distinti $\tau_1, \ldots, \tau_m$ tale che $v \neq \tau_i$ per ogni $i = 1, \ldots, m$ e $v, \tau_1, \ldots, \tau_m, v$ formi una traiettoria (n altre parole non esiste un $m-$ciclo in $\mathcal{G}$ per ogni $m \geq 1$). △

Esistono alcune configurazioni di base di un DAG (Jensen e Nielsen, 2007) che saranno molto utili nel seguito :

Definizione 3.17 (Connessione Seriale, Divergente, Convergente).
Dati tre nodi $a, b, c \in V$ si dice che essi formano una:

[2]L'acronimo DAG deriva dalla denominazione in lingua inglese Directed Acyclic Graph e viene di seguito impiegato per indicare tale particolare tipologia di grafo.

-***Connessione seriale*** *(*chain*): quando esiste una traiettoria orientata che li unisce:* $(a \to b \to c)$*. Il nodo b prende il nome di nodo di serie.*

-***Connessione divergente*** *(*fork*): quando a e c sono i figli del nodo b detto nodo* divergente*:* $(a \leftarrow b \to c)$*.*

-***Connessione convergente*** *(*collider*): quando a e c sono i genitori del nodo b detto nodo* convergente*:* $(a \to b \leftarrow c)$*.* △

Definizione 3.18 (Albero, Foresta, Foglia, Radice).
Un ***albero*** *è un* $\mathcal{G} = (V, E)$ connesso *e tale che non esiste alcun tragitto tra* v *e* v *per ogni coppia di nodi* $v, u \in V$ *con* $v \neq u$*, esiste un unico tragitto.*

Un nodo senza genitori viene detto ***radice****.*

Un nodo senza figli viene detto ***foglia****.*

Una foresta è un grafo dove tutti i suoi componenti connessi sono alberi. △

É possibile utilizzare i concetti precedentemente presentati per caratterizzare le relazioni esistenti fra variabili casuali, utilizzando le stesse come nodi e indicando con le diverse configurazioni di archi a disposizione le relazioni di interdipendenza tra i nodi. Nel seguente paragrafo verrà presentato come un grafo possa rappresentare una struttura di dipendenze condizionate.

3.4 La distribuzione congiunta e le indipendenze condizionali

I nodi che compongono un DAG possono essere associati a variabili casuali o, più in generale, ad un vettore aleatorio $\underline{X} = (X_1, \ldots, X_n)$ in modo tale che l'insieme dei nodi sia $V = \{X_1, \ldots, X_n\}$ e l'insieme E degli archi rappresenti un modello delle relazioni esistenti tra le v.c. stesse, ovvero tale che se esiste una relazione causale diretta fra X_i che agisce su X_j essa può essere rappresentata da un arco orientato da X_i a X_j $(X_i \to X_j)$ per cui $(X_i, X_j) \in E$.

Tutto quanto detto sulle relazioni fra nodi nel paragrafo precedente viene esteso a relazioni fra variabili per cui, ad esempio, $\Pi(X_i) = \{X_j : (X_j, X_i) \in E\}$.

Quando i nodi hanno una chiara ed univoca indicizzazione all'interno dell'insieme V, notazioni quali $Ad(X_i)$, $\Pi(X_i)$ e $Ch(X_i)$ possono essere sostituite con le più concise Ad_i, Π_i e Ch_i.

Nel caso oggetto di studio le variabili considerate sono tutte discrete con un numero finito di stati (modalità). Ogni variabile si trova in uno solo di questi stati, che può essere noto o no.

Data la definizione di probabilità e indipendenza condizionale la funzione $p_{X_1,\ldots,X_n}$ può essere riscritta come:

$$p_{X_1,\ldots,X_n} = p_{X_1}\, p_{X_2|X_1}\, p_{X_3|X_1,X_n}\, \cdots\, p_{X_n|X_1,\ldots,X_{n-1}}. \tag{3.16}$$

Più in generale vale che per ogni ordinamento σ di $(1, \ldots, n)$

$$p_{X_1,\ldots,X_n} = p_{X_{\sigma(1)}}\, p_{X_{\sigma(2)}|X_{\sigma(1)}}\, p_{X_{\sigma(3)}|X_{\sigma(1)},X_{\sigma(n)}}\, \cdots\, p_{X_{\sigma(n)}|X_{\sigma(1)},\ldots,X_{\sigma(n-1)}}. \tag{3.17}$$

Questo tipo di scomposizione della distribuzione congiunta di probabilità viene definita fattorizzazione. Un DAG può essere impiegato per indicare come ottenere una fattorizzazione semplificata in quanto evidenzia indipendenze condizionali che favoriscono la semplificazioni.

Definizione 3.19 (Fattorizzazione lungo un Grafico Aciclico Orientato). *Una distribuzione di probabilità $p_{X_1,\ldots,X_n}$ sulle variabili $X_1, \ldots, X_n$ si dice fattorizzare lungo un DAG $\mathcal{G}$ se esiste un ordinamento $X_{\sigma(1)}, \ldots, X_{\sigma(n)}$ delle variabili tale che:*

- $\Pi(\sigma(1)) = \Pi_{\sigma(1)} = \varnothing$; *Ovvero $X_{\sigma(1)}$ non ha* genitori.
- *per ogni j vale che* $\Pi(\sigma(j)) = \Pi_{\sigma(j)} \subset \left\{X_{\sigma(1)}, \ldots, X_{\sigma(j-1)}\right\}$
- $p_{X_{\sigma(j-1)}|X_{\sigma(1)},\ldots,X_{\sigma(j-1)}} = p_{X_{\sigma(j-1)}\,|\,\Pi_{\sigma(j-1)}}$ △

Per ogni ordinamento delle variabili esiste uno specifico DAG che indica il modo più opportuno di fattorizzare la distribuzione di probabilità e quali variabili possono essere escluse dal condizionamento.

Riprendendo le tre configurazioni presentate nella definizione 3.17, esse possono essere riviste alla luce dell'indipendenza condizionale per grafi di tre variabili.

Connessioni seriale. Considerate le variabili (X_1, X_2, X_3), dove X_1 agisce su X_2 che a sua volta agisce su X_3 si presenta il caso di una connessione seriale.

Se lo stato di X_2 è sconosciuto, le informazioni relative a X_1 influenzano la distribuzione di probabilità di X_2 che a sua volta influenza la distribuzione di X_3. Allo stesso modo, restando sempre ignoto X_2, eventuali informazioni relative a X_3 influenzano la distribuzione di probabilità di X_1 attraverso X_2.

Ovviamente, se lo stato di X_2 è noto, le *comunicazioni* tra X_1 e X_3 vengono interrotte e X_1 e X_3 diventano indipendenti, *separati*. Tale configurazione indica che la funzione di probabilità congiunta di (X_1, X_2, X_3) può essere fattorizzata come:

$$p_{X_1,X_2,X_3} = p_{X_1}\, p_{X_2|X_1}\, p_{X_3|X_2}. \tag{3.18}$$

Se X_2 è noto allora

$$p_{X_1,X_2,X_3}(.,x_2,.) = p_{X_1}(.)\; p_{X_2|X_1}(x_2|.)\; p_{X_3|X_2}(.|x_2)\,. \tag{3.19}$$

Il che significa, facendo riferimento alla proprietà 5) del teorema 3.2, che X_1 e X_3 sono condizionalmente indipendenti dato X_2.

Connessione divergente. In questo caso l'informazione si trasmette tra tutti i *figli* di X_1 a meno che quest'ultimo non sia noto, nel qual caso i figli vengono *separati* e diventano indipendenti. Anche questa configurazione permette di

semplificare la distribuzione congiunta per ottenere:

$$p_{X_1,X_2,X_3} = p_{X_1}\, p_{X_2|X_1}\, p_{X_3|X_1}. \tag{3.20}$$

Nel caso di X_1 noto si ottiene:

$$p_{X_2,X_3|X_1}(.,.|x_1) = p_{X_2|X_1}(.|x_1)\; p_{X_3|X_1}(.|x_1). \tag{3.21}$$

Facendo riferimento alla proprietà 3) del teorema 3.2 questa configurazione implica che X_1 e X_3 sono condizionalmente indipendenti dato X_2.

Connessione convergente. Questa configurazione è leggermente più complessa perché se X_1 è noto, X_2 e X_3 non sono condizionalmente indipendenti, poiché, visto che entrambi causano X_1, se questo è noto, l'occorrenza dell'uno fa diminuire le probabilità che occorra l'altro.

La fattorizzazione della distribuzione congiunta che si ottiene è la seguente:

$$p_{X_1,X_2,X_3} = p_{X_2}\, p_{X_3}\, p_{X_1|X_2,X_3}. \tag{3.22}$$

Si può notare però che, in questo caso X_1 ed X_3, sono indipendenti se X_1 è ignoto, infatti per ogni $\left(x_2^{(i)}, x_3^{(j)}\right) \in \mathcal{X}_2 \times \mathcal{X}_3$, dove $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ sono gli spazi degli stati per X_1, X_2, X_3, vale rispettivamente che:

$$\begin{aligned} p_{X_2,X_3}\left(x_2^{(i)}, x_3^{(j)}\right) &= \sum_{y\in\mathcal{X}_1} p_{X_2}\left(x_2^{(i)}\right)\, p_{X_3}\left(x_3^{(j)}\right)\, p_{X_1|X_2,X_3}\left(y|x_2^{(i)}, x_3^{(j)}\right) \\ &= p_{X_2}\left(x_2^{(i)}\right)\, p_{X_3}\left(x_3^{(j)}\right) \sum_{y\in\mathcal{X}_1} p_{X_1|X_2,X_3}\left(y|x_2^{(i)}, x_3^{(j)}\right) \\ &= p_{X_2}\left(x_2^{(i)}\right)\, p_{X_3}\left(x_3^{(j)}\right). \end{aligned}$$

Per una connessione seriale o divergente la condizione di blocco delle informazioni richiede che il nodo rispettivamente di serie o di divergenza sia instanziato. L'apertura nel caso della connessione convergente vale per ogni informazione su ogni variabile *discendente*.

Le informazioni in un grafico vengono trasmesse attraverso i tragitti attivi definiti come segue:

Definizione 3.20 (Tragitto S-attivo).
Sia $\mathcal{G} = (V, E)$ un grafo orientato aciclico. Sia $S \subset V$ e siano $X, Y \in V \backslash S$.

Un tragitto τ tra X e Y è detto S-attivo *se:*

- tutti i nodi di connessione convergente in τ sono in S o hanno un discendente in S.

- tutti gli altri nodi non appartengono ad S. △

Definizione 3.21 (Tragitto bloccato).
Un tragitto che non sia S-attivo *viene definito come bloccato da S.* △

A questo punto è possibile introdurre una definizione fondamentale per le reti bayesiane:

Definizione 3.22 (D-separazione).
Sia $\mathcal{G} = (V, E)$ un grafo orientato aciclico dove $V = \{X_1, \ldots, X_n\}$ è una collezione di variabili aleatorie.

Sia $S \subset V$ tale che tutte le variabili appartenenti ad S siano instanziate e tutte le variabili appartenenti a $V \backslash S$ siano non instanziate.

Due distinte variabili X_i e X_j che non appartengono ad S sono dette ***d-separate*** *da S se tutti i tragitti tra X_i e X_j sono bloccati da S.*

Siano C e D due insiemi di variabili. Se ogni tragitto da ciascuna variabile di C verso ogni variabile di D è bloccato da S allora C e D sono ***d-separate*** *da S.*

L'insieme S blocca ogni possibile traiettoria da C a D. Tale condizione si indica con:

$$C \perp D \,||_{\mathcal{G}}\, S \qquad \triangle$$

Ritornando alle configurazioni della definizione 3.17 si può notare come nelle connessioni seriali e in quelle divergenti gli estremi siano d-separati dal nodo centrale di serie o di divergenza, mentre nelle connessioni convergenti si ha d-separazione solo se il nodo centrale di convergenza non è instanziato.

La lettera d di d-separazione indica separazione orientata (*directed*) e in termini comuni può essere interpretata come *irrilevanza*. Se l'insieme C e D sono d-separati da S significa che se S è noto, nuove informazioni riguardo C sono irrilevanti per la conoscenza della distribuzione di D e viceversa.

Definizione 3.23 (D-connessione).
Se due variabili X ed Y non sono d-separate si dicono d-connesse. △

Per individuare se gli insiemi di variabili C e D sono d-separati da S è possibile impiegare la procedura dell'algoritmo 2.

Perché C e D siano d-separati è necessario che tutti i tragitti che li collegano siano inattivi. Un solo tragitto attivo impedisce la d-separazione.

Risulta evidente che una variabile X risulta d-separata dal resto del grafo dall'insieme dei suoi genitori Π, dei suoi figli Ch e dei suoi coniugi Cn. Tale insieme prende il nome di Markov Blanket:

Definizione 3.24 (Markov Blanket).
Il Markov Blanket di una variabile X_i (indicato con $BL(X_i)$) è definito come l'insieme composto dai suoi genitori Π_i, dai suoi figli Ch_i e dai suoi coniugi Cn_i

$$BL(X_i) = (\Pi_i \cup Ch_i \cup Cn_i) \tag{3.23}$$

Algoritmo 2 Algoritmo di ricerca di d-separazione

Dati: grafo $\mathcal{G} = (V, E)$, $C, D, S \subset V$;
Passo 1: trovare tutti i tragitti che collegano ogni variabile di C con ciascuna variabile di D;
Passo 2: controllare ogni tragitto fino a che ne venga reperito uno *attivo*:
Passo 2.1: Se vengono trovati un nodo seriale o un nodo divergente del tragitto che appartengano ad S allora il tragitto non è attivo.
Passo 2.2: Se nel tragitto viene trovato un nodo convergente, allora verificare che ognuno dei suoi discendenti sia in S. Se non si verifica mai tale situazione allora il tragitto non è attivo.
Passo 2.3: In tutti gli altri casi il tragitto è attivo.
Passo 3: Se viene trovato un tragitto attivo C e D non sono d-separati, se nessun tragitto attivo viene trovato C e D sono d-separati.

Se $BL(X_i)$ è minimale, ovvero nessun suo sottoinsieme è ancora un Markov Blanket, S è detto Markov Boundary $B_I(X)$ △

Quanto esposto nel paragrafo 3.4.2 consente di sostenere che due grafi aciclici orientati sono Markov-equivalenti se possiedono esattamente le stesse proprietà di d-separazione. Verma e Pearl (1990) hanno dimostrato che alcune caratteristiche grafiche rappresentano condizione necessaria e sufficiente per determinare le classi di equivalenza markoviana, esse sono le *immoralità* e lo *scheletro*:

Definizione 3.25 (Immoralità).
Sia $\mathcal{G} = (V, E)$ un grafo, sia $E = D \cup I$ e $D \cap I = \varnothing$.
Si definisce ***immoralità*** *in un grafo la tripletta a, b, c tale che $(a, b) \in D$ e $(c, b) \in D$ ma $(a, c) \notin D$, $(c, a) \notin D$e $\langle a, c \rangle \notin I$,*

Definizione 3.26 (Scheletro).
Lo ***scheletro*** *di un grafo $\mathcal{G} = (V, E)$ è il grafo non orientato ottenuto rendendo non orientato il grafo.* △

Il seguente teorema mostra come ottenere le classi di equivalenza markoviana.

Teorema 3.7.
Due grafi aciclici orientati sono Markov-equivalenti sse hanno lo stesso scheletro e le stesse immoralità. △

Per la dimostrazione si veda (Koski e Noble, 2009).

3.4.1 Grafo e distribuzione congiunta: Reti Probabilistiche Bayesiane

Al fine di presentare la definizione di rete probabilistica bayesiana è necessario introdurre il concetto di **potenziale.**

La notazione utilizzata prevede che con $V = \{X_1, \ldots, X_n\}$ venga indicata una collezione di variabili causali X_j, ognuna delle quali caratterizzata dallo spazio degli stati $\mathcal{X}_j = (x_j(1), \ldots, x_j(k_j))$ per $j = 1, \ldots, n$.

Sia $\mathcal{X} = \bigtimes_{j=1}^{n} \mathcal{X}_j$ lo spazio degli stati per $\underline{X}$.

Sia $\widetilde{V} = \{1, \ldots, n\}$ l'insieme degli indici per le variabili.

Per $D \subset \widetilde{V}$, con $D = \{j_1, \ldots, j_m\}$, sia

$$\mathcal{X}_D = \bigtimes_{j \in D} \mathcal{X}_j$$

e sia $\underline{X}_D = (X_{j1}, \ldots, X_{jm})$.

Sia $\underline{x} \in \mathcal{X}$ un generico elemento di $\mathcal{X}$ e sia $\underline{x}_D = (x_{j1}, \ldots, x_{jm}) \in \mathcal{X}_D$ con $\underline{x} = (x_1, \ldots, x_m) \in \mathcal{X}$.

Siano inoltre $W \subset V$ e sia $\widetilde{W}$ un insieme di indicizzazione per W. La notazione $\mathcal{X}_W$ sarà usata per indicare $\mathcal{X}_{\widetilde{W}}$, la notazione $\underline{X}_W$ per indicare $\underline{X}_{\widetilde{W}}$ e $\underline{x}_W$ per indicare $\underline{x}_{\widetilde{W}}$.

Definizione 3.27 (Potenziale).
Un potenziale ϕ su un dominio $\mathcal{X}_D$ è definito come una funzione non negativa $\phi : \mathcal{X}_D \to \mathbb{R}_+$.

Lo spazio $\mathcal{X}_D$ è definito come il dominio del potenziale.

Se il dominio è lo spazio degli stati del vettore casuale $\underline{X}_D$, allora il vettore casuale $\underline{X}_D$ viene anche indicato come il dominio del potenziale. △

In queste impostazioni un potenziale su un dominio $\mathcal{X}_D$ ha $\prod_{j \in D} k_j$ entrate.

Per $W \subset V$ il dominio di un potenziale $\mathcal{X}_W$ può essere ugualmente identificato con con la collezione di variabili casuali W.

Definizione 3.28 (Operazioni su potenziali aventi lo stesso dominio).
Due potenziali ϕ e ϕ' definiti sullo stesso dominio $\mathcal{X}_D$ possono essere addizionati, moltiplicati e divisi fra loro.

L'operazione di ***addizione*** *fra potenziali, indicata con $\phi + \phi'$ corrisponde alla somma per coordinate ovvero per ogni $\underline{x}_D \in \mathcal{X}_D$ vale che $(\phi + \phi')(\underline{x}_D) = \phi(\underline{x}_D) + \phi'(\underline{x}_D)$.*

L'operazione di ***moltiplicazione*** *fra potenziali indicata con $\phi.\phi'$ corrisponde alla moltiplicazione per coordinate ed è definita moltiplicando ogni entrata di ciascuna configurazione. per ottenere il nuovo potenziale $\phi.\phi'$.*

L'operazione di ***divisione*** *fra potenziali indicata con ϕ/ϕ' corrisponde alla divisione, ove consentita (denominatore del rapporto è diverso da zero) per coordinate è definita moltiplicando ogni entrata nella configurazione per ottenere il nuovo potenziale $\phi.\phi'$.* △

Nel caso di potenziali su domini diversi è possibile estendere il dominio degli stessi in modo da ottenerne uno comune.

Definizione 3.29 (Estensione del dominio).
Sia il potenziale ϕ definito sul dominio $\mathcal{X}_D$, dove $D \subset \widetilde{W} \subseteq \widetilde{V}$.

Sia il potenziale ϕ definito sul dominio $\mathcal{X}_D$ viene esteso al dominio definito sul dominio $\mathcal{X}_{\widetilde{W}}$ nel seguente modo:

- Per ogni $x_{\widetilde{W}} \in \mathcal{X}_{\widetilde{W}}$ $\phi\left(\underline{x}_{\widetilde{W}}\right) = \phi\left(\underline{x}_D\right)$.

dove $\underline{x}_D$ è la proiezione di $x_{\widetilde{W}}$ su $\mathcal{X}_D$. △

A questo punto è possibile estendere le operazioni della definizione 3.28 anche a potenziali dominio differente:

Definizione 3.30.
Le operazioni su potenziali ϕ e ϕ' aventi domini differenti $\mathcal{X}_{D_1}$ e $\mathcal{X}_{D_2}$ vengono definite estendendo entrambi i domini a $\mathcal{X}_{D_1 \cup D_2}$ e su di essi applicando la definizione 3.28. △

Marginalizzazione di un potenziale

L'ultima e più importante operazione introdotta su potenziali è la marginalizzazione, ovvero la somma sullo spazio degli stati $\mathcal{X}_{X_i}$.

Definizione 3.31 (Marginalizzazione di Potenziale).
Sia $V = \{X_1, \ldots, X_n\}$ un insieme di n casuali indicizzate dall'insieme $\widetilde{V} = \{1, \ldots, n\}$

Sia $U \subseteq W \subseteq V$ e sia ϕ un potenziale definito su $\mathcal{X}_W$.

L'espressione $\sum_{\mathcal{X}_{W \setminus U}} \phi$ (o anche $\sum_{W \setminus U} \phi$ o la più sintetica $phi^{\downarrow U}$) denota il risultato della marginalizzazione (o somma marginale) di ϕ su $\mathcal{X}_U$ e viene definita, per $\underline{x}_U \in \mathcal{X}_U$, come:

$$\left(\sum_{W \setminus U} \phi\right)(\underline{x}_U) = \left(\sum_{\underline{z} \in \mathcal{X}_W \setminus \mathcal{X}_U} \phi\right)(\underline{z}, \underline{x}_U)$$

dove $\underline{z} \in \mathcal{X}_W$ è la proiezione di $(\underline{z}, \underline{x}_U)\,\mathcal{X}$ su $\mathcal{X}_W$ e $\underline{x}_U \in \mathcal{X}_U$ la proiezione di $(\underline{z}, \underline{x}_U) \in \mathcal{X}$ su $\mathcal{X}_U$. △

Per l'operazione di marginalizzazione valgono le seguenti proprietà:

Proprietà 3.2.
-Legge commutativa: per ogni coppia di insiemi di variabili $U, W \subset V$, vale che $\left(\phi^{\downarrow U}\right)^{\downarrow W} = \left(\phi^{\downarrow W}\right)^{\downarrow U}$.
-Legge distributiva: se $\mathcal{X}_{D_1}$ è il dominio di ϕ_1 e $D_1 \subseteq \widetilde{V}$, allora $(\phi_1 \, \phi_2)^{\downarrow D_1} = \phi_1 \, (\phi_2)^{\downarrow D_1}$. △

È possibile vedere la distribuzione congiunta di tre variabili X_1, X_2, X_3 definite rispettivamente su $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$ come un potenziale a tre vie e a questo punto la distribuzione di probabilità condizionale può essere ricondotta alla marginalizzazione.

Lemma 3.1 (Shafer).
Per ogni DAG con un numero finito di nodi $v_1, \ldots, v_n$ esiste un ordinamento dei nodi $\left(v_{\sigma(1)}, \ldots, v_{\sigma(n)}\right)$, non necessariamente unico, tale che i genitori di $v_{\sigma(i)}$ siano un sottoinsieme di $\left\{v_{\sigma(1)}, \ldots, v_{\sigma(i-1)}\right\}$. △

La dimostrazione di tale lemma può essere reperita in Koski e Noble (2009).

Definizione 3.32 (Rete Probabilistica Bayesiana).
Una rete probabilistica bayesiana è definita come una coppia $(\mathcal{G}, p)$ dove

- $\mathcal{G} = (V, E)$ è un grafo aciclico orientato composto dall'insieme dei nodi $V = \{1, \ldots, n\}$ per $d \in \mathbb{N}$ ed E è l'insieme degli archi orientati;

- p è una distribuzione di probabilità o una famiglia di distribuzioni di probabilità indicizzata dall'insieme dei parametri Θ sullo spazio delle variabili discrete $\{1, \ldots, n\}$.

La coppia $(\mathcal{G}, p)$ soddisfa i seguenti criteri:

- per ogni $\theta \in \Theta$, $p\,(.|\theta)$ è una distribuzione di probabilità con lo stesso spazio degli stati $\mathcal{X}$, dove $\mathcal{X}$ ha cardinalità finita.

In altre parole $\forall \theta \in \Theta$, $p\,(.|\theta) : \mathcal{X} \to [0,\, 1]$ e $\sum_{\underline{x} \in \mathcal{X}} p\,(\underline{x}|\theta) = 1$.

- Ad ogni nodo $X_v \in V$ senza variabili genitori viene assegnato un potenziale di probabilità denotato da p_{X_v} che dà la distribuzione della variabile X_v.

- Ad ogni nodo $X_v \in V$ con un insieme di genitori $\Pi_v = \left(X_{b_1^{(v)}}, \ldots, X_{b_m^{(v)}}\right)$ non vuoto viene assegnato un potenziale di probabilità condizionata (CPP[3]) denotato da $p_{X_v|\Pi_v}$ che contiene la funzione di distribuzione della probabilità condizionale della variabile X_v, date le variabili $\left\{X_{b_1^{(v)}}, \ldots, X_{b_m^{(v)}}\right\}$.

- La funzione di distribuzione congiunta può essere fattorizzata come segue:

$$p_{X_1,\ldots,X_n} = \prod_{v=1}^{n} p_{X_v|\Pi_v}$$

[3]Conditional Probability Potential.

- La fattorizzazione è minimale nel senso che per ogni ordinamento delle variabili tale che $\Pi_j \subseteq \{X_1, \ldots, X_{j-1}\}$, Π_j *è il più piccolo insieme di variabili tali che* $X_j \perp \Pi_j^c | \Pi_j$, *vale a dire che:*

$$\Pi_j = \bigcap \{A \subseteq \{X_1, \ldots, X_{j-1}\} \text{ tali che } X_j \perp A^c | A\}. \qquad \triangle$$

3.4.2 Indipendenza Condizionale e D-separazione

Esiste un forte legame tra indipendenza condizionale e d-separazione delle variabili di un DAG come mostra il seguente teorema:

Teorema 3.8 (D-separazione → Indipendenza Condizionale).
Sia $\mathcal{G} = (V, E)$ *un grafo ciclico orientato, e sia* p *una distribuzione di probabilità che fattorizza lungo* $\mathcal{G}$.

Se per ogni tripletta di insiemi disgiunti $A, B, C \subset V$ *vale che* A *e* B *sono d-separati, allora* $A \perp B | C$ *(*A *e* B *sono indipendenti dato* C*).* $\triangle$

Per la dimostrazione si veda Koski e Noble (2009).

La relazione inversa è più complicata da verificare e richiede l'introduzione di alcune nozioni supplementari.

Innanzitutto esiste una condizione necessaria e sufficiente perché una funzione di probabilità p su un insieme V di variabili casuali possa essere fattorizzata lungo un grafo $\mathcal{G}$.

Definizione 3.33 (Condizione di Markov locale orientata).
Sia $V = \{X_1, \ldots, X_n\}$ *un insieme di variabili discrete. Una funzione di probabilità* p *sul vettore casuale* $\underline{X} = (X_1, \ldots, X_n)$ *soddisfa la Condizione di Markov locale orientata rispetto al grafo aciclico orientato* $\mathcal{G} = (V, E)$ *(ed è detta localmente* $\mathcal{G}$*-markoviana) sse per ogni* $j \in \{1, \ldots, n\}$, X_j *è condizionalmente indipendente, dati i suoi genitori* Π_j, *da tutte le variabili dell"insieme* $V \setminus (V_j \cup Pi_j)$, *dove* V_j *è l'insieme di tutti i discendenti di* X_j, *ovvero*

$$V_j = \{Y \in V : \text{ esiste una traiettoria orientata da } X_j \text{ a } Y\}.$$

In questo caso vale che:

$$X_j \perp V \setminus (V_j \cup Pi_j) \, | \Pi_j. \qquad \triangle$$

Definizione 3.34 (Modello di Markov).
Sia $\mathcal{G} = (V, E)$ *un grafo aciclico orientato con* $V = \{X_1, \ldots, X_n\}$ *insieme di variabili casuali. Sia* $\mathcal{V}$ *l'insieme di tutti i sottoinsiemi di* V *e sia* p *una funzione di probabilità per il vettore casuale* $\underline{X} = (X_1, \ldots, X_n)$. *Sia*

$$\mathcal{I}(p) = \{(X, Y, S) \in V \times V \times \mathcal{V} | X, Y \notin S, \quad X \perp Y | S\}.$$

Il modello di Markov $\mathcal{M}_{\mathcal{G}}$ determinato dal grafo aciclico orientato $\mathcal{G}$ è l'insieme di asserzioni di indipendenza condizionale:

$$\mathcal{M}_{\mathcal{G}} = \{\mathcal{I} | \mathcal{I} = \mathcal{I}(p) \text{ per qualche } p \text{ che sia localmente } \mathcal{G} - \text{ Markoviana } \}.$$

In altre parole il modello di Markov è la collezione formata da tutti gli insiemi $\mathcal{I}$ di relazioni di indipendenza condizionale corrispondenti alle distribuzioni localmente $\mathcal{G}$- Markoviane.

Una distribuzione p si dice appartenere al modello markoviano di $\mathcal{G}$ ($p \in \mathcal{M}_{\mathcal{G}}$) sse $\mathcal{I}(p) \in \mathcal{M}_{\mathcal{G}}$ △

Teorema 3.9.
Sia $\mathcal{I}(p)$ l'insieme di tutte le asserzioni le indipendenze condizionali soddisfatte dalla funzione di probabilità p per un vettore casuale $\underline{X} = (X_1, \ldots, X_n)$.
$\mathcal{I}(p) \in \mathcal{M}_{\mathcal{G}}$ sse p fattorizza lungo $\mathcal{G}$.

Per la dimostrazione di veda (Koski e Noble, 2009).

Definizione 3.35 (I-map perfetta(p-map), Fedeltà).
Sia $\mathcal{G} = (V, E)$ un grafo aciclico orientato con $V = \{X_1, \ldots, X_n\}$ insieme di variabili casuali. Si dice che $\mathcal{G}$ è una I-map perfetta per una distribuzione di probabilità p su V se per ogni tripletta di sottoinsiemi distinti di variabili $A, B, C \subset V$ vale che

$$A \perp B | C \Leftrightarrow A \perp B ||_{\mathcal{G}} C.$$ △

Se $\mathcal{G}$ è una I-map perfetta per p, allora si dice che $\mathcal{G}$ è fedele a p.

Definizione 3.36 (Consistenza).
Sia $V = \{X_1, \ldots, X_n\}$ una collezione di variabili casuali sia $\mathcal{M}$ l'insieme di tutte le asserzioni di indipendenza condizionale. Sia $\mathcal{V}$ l'insieme di tutti i sottoinsiemi di V, allora per ogni $(X_i, X_j, S) \in V \times V \times \mathcal{V}$ tali che $X_i, X_j \notin S$, vale che:

$$(X_i, X_j, S) \in \mathcal{M} \Leftrightarrow X_i \perp X_j | S.$$

Un grafo aciclico orientato $\mathcal{G} = (V, E)$ è consistente con un insieme di asserzioni di indipendenze condizionali $\mathcal{M}$ sse

$$(X_i, X_j, S) \in \mathcal{M} \Leftrightarrow X_i \perp X_j ||_{\mathcal{G}} S.$$ △

Definizione 3.37 (I-sub-map, I-map, I-equivalenza (di Markov)).
Siano $\mathcal{G}_1$ e $\mathcal{G}_2$ due DAG sullo stesso insieme di variabili V.

Si dice che $\mathcal{G}_1$ è una I-sub-map di $\mathcal{G}_2$ se ogni coppia di variabili d-separate da un insieme in $\mathcal{G}_1$ è d-separato dallo stesso insieme anche in $\mathcal{G}_2$.

Si dice che $\mathcal{G}_1$ e $\mathcal{G}_2$ sono I-equivalenti se $\mathcal{G}_1$ è una I-sub-map di $\mathcal{G}_2$ e $\mathcal{G}_2$ è una I-sub-map di $\mathcal{G}_1$.

Teorema 3.10 (d-separazione → indipendenza condizionale).
Sia $\mathcal{G} = (V, E)$ un grafo ciclico orientato, e sia p una distribuzione di probabilità per il vettore $\underline{X} = (X_1, \ldots, X_n)$ che fattorizza lungo $\mathcal{G}$. Si dice che p e $\mathcal{G}$ sono reciprocamente fedeli sse per ogni tripletta di insiemi disgiunti di variabili $A, B, C \subset V$ vale che

$$A \perp B ||_{\mathcal{G}} C \Leftrightarrow p_{A,B|C} = p_{A|C}\, p_{B|C}.$$

In altre parole p e $\mathcal{G}$ sono reciprocamente fedeli sse per ogni tripletta di insiemi disgiunti di variabili $A, B, C \subset V$ vale che

$$A \perp B ||_{\mathcal{G}} C \Leftrightarrow A \perp B | C. \qquad \triangle$$

Per la dimostrazione di veda (Naïm *e altri*, 2007).

In generale, data una distribuzione congiunta di probabilità $p_{\underline{X}}$ su un vettore casuale $\underline{X} = (X_1, \ldots, X_n)$ non è sempre possibile trovare un grafo aciclico orientato che sia fedele alle relazioni di indipendenza di p.

É possibile introdurre il concetto di sufficienza bayesiana anche per reti probabilistiche bayesiane.

Sia $\mathcal{G} = (V, E)$ un grafo aciclico orientato con $V = \{X_1, \ldots, X_n\}$ dove le X_i sono state enumerate in modo che $\Pi_i \subseteq \{X_1, \ldots, X_{i-1}\}$. In ambito bayesiamo possiamo considerare il vettore dei parametri $\underline{\theta}$ come una realizzazione del vettore aleatorio $\underline{\Theta}$ dove, per $j = 1, \ldots, n$, è possibile determinare $\underline{\theta}_j$ a sua volta realizzazione del vettore aleatorio $\underline{\Theta}_j$.

In tal caso si definisce la seguente proprietà di un insieme di parametri:

Definizione 3.38 (Modularità dei parametri).
*Un insieme di parametri $\underline{\Theta}$ per una rete probabilistica bayesiana soddisfa la proprietà di **modularità dei parametri** se può essere scomposto in n distinti insiemi di parametri $\underline{\Theta}_1, \ldots, \underline{\Theta}_n$ tali che per $j = 1, \ldots, n$ il vettore dei parametri $\underline{\Theta}_j$ è legato direttamente solo al nodo X_j.* $\triangle$

L'ipotesi di modularità dei parametri consente di estendere un DAG inserendo dei nodi-parametro come genitori delle variabili del grafo dirigendo l'arco da ogni

nodo dell'insieme $\underline{\Theta}_j$ verso il rispettivo nodo X_j in modo da generare un altro grafo aciclico diretto dove $P_{X_1,\ldots,X_n|\underline{\Theta}}$ è decomponibile come

$$p_{X_1,\ldots,X_n|\underline{\Theta}} = \prod_{j=1}^{n} p_{X_j|\underline{\Theta},\Pi_j} \tag{3.24}$$

Sotto assunzione di modularità, $\underline{\Theta}_1, \ldots, \underline{\Theta}_n$ sono vettori aleatori indipendenti la cui distribuzione congiunta a priori è data dal prodotto delle distribuzioni individuali a priori: $\pi_{\underline{\Theta}} = \prod_{j=1}^{n} \pi_{\underline{\Theta}_j}$.

Utilizzando la seguente notazione, con $j = 1, \ldots, n$:

$$\widetilde{X}_j := \left(\left(\underline{X}_1, \underline{\Theta}_1\right), \ldots, \left(\underline{X}_{j-1}, \underline{\Theta}_{j-1}\right)\right) \tag{3.25}$$

e t_j funzione tale che:

$$t_j\left(\widetilde{X}_j\right) = \Pi_j \tag{3.26}$$

si ricava che:

$$\widetilde{X}_j \perp \left(\underline{X}_j, \underline{\Theta}_j\right) | \Pi_j \tag{3.27}$$

In altre parole l'insieme Π_j dei genitori di X_j è una statistica predittiva sufficiente per $\left(\underline{X}_j, \underline{\Theta}_j\right)$ nel senso che in $\widetilde{X}_j := \left(\left(\underline{X}_1, \underline{\Theta}_1\right), \ldots, \left(\underline{X}_{j-1}, \underline{\Theta}_{j-1}\right)\right)$ non ci sono ulteriori informazioni rilevanti per predire $\underline{X}_j$ o $\underline{\Theta}_j$.

In una rete probabilistica bayesiana i cui parametri soddisfano l'assunzione di modularità, (Π_j, X_j) sono una statistica sufficiente per il parametro $\underline{\Theta}_j$.

3.5 Stima dei parametri del modello

Per semplicità vengono presentate prima le tecniche di stima dei potenziali, che sebbene non siano operazioni semplici, quantomeno rappresentano un problema comunemente affrontato nell'ambito della stima bayesiana. Il passo successivo è quello di stimare la relazione causale tra i nodi, un problema che presenta difficoltà maggiori dato l'enorme numero di possibili configurazioni fra gli stessi.

3.5.1 Stima della Distribuzione congiunta: l'apprendimento delle probabilità potenziali

Una volta acquisita con l'apprendimento automatico, tramite l'implementazione delle conoscenze degli esperti o entrambi, la struttura delle relazioni causali, per

completare la stima di una rete probabilistica bayesiana è necessario provvedere alla stima della distribuzione condizionata dell'insieme delle variabili/nodi che compongono il DAG.

La distribuzione congiunta può essere partizionata in una produttoria di probabilità condizionate, dato che per ogni variabile X_j, il potenziale della probabilità $\left\{\left(p_{X_j} \mid \Pi_j\right)_{j=1}^{d}\right\}$ dipende solo dai genitori Π_j di X_j.

Per quanto riguarda la distribuzione a priori da attribuire ai CPP può essere espressa come un prodotto di distribuzioni di Dirichlet, infatti Geiger e Heckerman (1997) mostrano come, sotto assunzioni piuttosto ampie, sia inevitabile l'impiego di questa distribuzione, che tra l'altro permette un metodo di aggiornamento il più delle volte semplificato.

Oltre alla stima bayesiana dei CPP, viene presentato per completezza anche il metodo frequentista per eccellenza per la stima dei parametri: la massima verosimiglianza. Prima di procedere però, viene introdotta la notazione del caso e viene presentato il problema.

Notazione

Sia indicato $V = X_1, X_2, \ldots, X_d$ l'insieme delle variabili casuali considerate e sia indicato $\mathcal{X}_j = \left\{x_1^{(1)}, x_2^{(2)}, \ldots, x_d^{(k_j)}\right\}$ l'insieme di tutti i valori che può assumere ogni X_j. l'insieme di tutte le possibili configurazioni del risultato dell'esperimento è dato da $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$, ovvero

$$\mathcal{X} = \left\{\left(x_1^{(j_1)}, \ldots, x^{(j_d)}\right) \mid j_1 \in (1, \ldots, k_1), \ldots, j_d \in (1, \ldots, k_d)\right\},$$

dove, per non generare confusione fra matrici di variabili casuali e matrici di dati, ogni realizzazione (denominata anche *istanziazione*) viene indicata con il vettore sottosegnato $\underline{x} = \left(x_1^{(j_1)}, \ldots, x^{(j_d)}\right)$, preso come riga. In questo modo la matrice delle realizzazioni campionarie può essere indicata con:

$$\boldsymbol{x} = \begin{pmatrix} \underline{x}_{(1)} \\ \vdots \\ \underline{x}_{(n)} \end{pmatrix}$$

dove il vettore riga $\underline{x}_{(i)} = x_{i,1}^{(j_1)}, \ldots, x_{i,d}^{(j_d)}$ $(i = 1, \ldots, n)$, rappresenta la i-esima realizzazione e la colonna

$$\underline{\boldsymbol{x}}^{(j)} = \begin{pmatrix} x_{1,j}^{(j)} \\ \vdots \\ x_{n,j}^{(j)} \end{pmatrix}$$

rappresenta il vettore delle realizzazioni della j-esima variabile X_j.

Per completare il contesto dell'esperimento, indicato con Ω, viene introdotto $\widetilde{\Theta}$, ovvero lo spazio dei parametri della distribuzione in modo da ottenere $\Omega = \widetilde{\Theta} \times \mathcal{X}$.

Il DAG associato alla rete di relazioni causali viene indicato con $\mathcal{G} = (V, E)$ e da esso si ricavano le indicazioni per fattorizzare $p_{X_1,\ldots,X_d}$

Per ogni variabile X_j, si considerino tutti le possibili istanziazioni dell'insieme dei genitori Π_j e le si indichi con $\left(\pi_j^{(l)}\right)_{l=1}^{q_j}$, dove $\left(\pi_j^{(l)}\right)$ indica che la configurazione dei genitori della variabile X_j è nello stato $\left(\pi_j^{(l)}\right)$ e ci sono q_j possibili configurazioni di Π_j.

Impostando:

$$n_k\left(x_j^{(i)} \mid \pi_j^{(l)}\right) = \begin{cases} 1 & \text{se } \left(x_j^{(i)}, \pi_j^{(l)}\right) \text{ si trova in } \underline{x}_{(k)} \\ 0 & \text{altrove} \end{cases}$$

dove $\left(x_j^{(i)}, \pi_j^{(l)}\right)$ è una configurazione della famiglia $(X_j,\ \Pi_j)$.

Sia $\underline{\theta}$ l'insieme dei parametri, definiti come:

$$\theta_{jil} = p\left(\left\{X_j = x_j^{(i)}\right\} \mid \left\{\Pi_j = \pi_j^{(l)}\right\}\right).$$

per $j = 1, \ldots, d$, $i = 1, \ldots, k_j$, $l = 1, \ldots, q_j$, data la struttura del grafo $\mathcal{G} = (V, E)$.

La probabilità congiunta di un qualsiasi caso $\underline{x}^{(k)}$ può essere scritta come:

$$p_{\underline{X}|\underline{\Theta}}\left(\underline{x}_{(k)} \mid \underline{\theta}, E\right) = \prod_{j=1}^{d} \prod_{l=1}^{q_j} \prod_{i=1}^{k_j} \theta_{jil}^{n_k\left(x_j^{(i)}|\pi_j^{(l)}\right)}$$

con:

- d pari al numero dei nodi (delle variabili);
- k_j pari al numero dei possibili stati della variabile X_j;

- q_j il numero delle possibili configurazioni dei genitori di X_j;
- $\theta_{jil} := p\left(\left\{X_j = x_j^{(i)}\right\} \mid \left\{\Pi_j = \pi_j^{(l)}\right\}\right)$;

Ciò vuol dire che θ_{jil} è la probabilità condizionale della variabile X_j nello stato i, data la configurazione l dei genitori, ovvero $\pi_j^{(l)}$.

Se consideriamo la matrice (delle possibili realizzazioni):

$$\boldsymbol{X} = \begin{pmatrix} \underline{X}_{(1)} \\ \vdots \\ \underline{X}_{(n)} \end{pmatrix}$$

la matrice

$$\boldsymbol{x} = \begin{pmatrix} \underline{x}_{(1)} \\ \vdots \\ \underline{x}_{(n)} \end{pmatrix}$$

risulta essere composta da n osservazioni indipendenti, da cui:

$$p_{\underline{X}|\Theta}\left(\underline{x}_{(k)} \mid \underline{\theta}, \mathcal{G}\right) = \prod_{k=1}^{n} p_{\underline{X}|\Theta}\left(\underline{\boldsymbol{x}}_{(k)} \mid \underline{\theta}, \mathcal{G}\right) = \prod_{j=1}^{d}\prod_{l=1}^{q_j}\prod_{i=1}^{k_j}\prod_{k=1}^{n} \theta_{jil}^{n_k\left(x_j^{(i)}|\pi_j^{(l)}\right)} =$$
$$= \prod_{j=1}^{d}\prod_{l=1}^{q_j}\prod_{i=1}^{k_j} \theta_{jil}^{\sum_{k=1}^{n} n_k\left(x_j^{(i)}|\pi_j^{(l)}\right)}$$

Se usiamo la convenzione:

$$n\left(x_j^{(i)} \mid \pi_j^{(l)}\right) = \sum_{k=1}^{n} n_k\left(x_j^{(i)} \mid \pi_j^{(l)}\right)$$

Questo è il numero di volte in cui la configurazione $\left(x_j^{(i)} \mid \pi_j^{(l)}\right)$ appare in $\boldsymbol{x} = \begin{pmatrix} \underline{x}_{(1)} \\ \vdots \\ \underline{x}_{(n)} \end{pmatrix}$.

La funzione di verosimiglianza assume la seguente forma:

$$L\left(\underline{\theta}\right) = \prod_{k=1}^{n} p_{\underline{X}|\underline{\Theta}}\left(\underline{\boldsymbol{x}}^{(k)} \mid \underline{\theta}, \mathcal{G}\right) = \prod_{j=1}^{d}\prod_{l=1}^{q_j}\prod_{i=1}^{k_j} \theta_{jil}^{n\left(x_j^{(i)}|\pi_j^{(l)}\right)}$$

La verosimiglianza può essere fattorizzata in fattori locali genitore-figlio e in più in $d \times q_j$ separati stimatori di massima verosomiglinza tutti delle forma base e si ricava (si veda Koski e Noble, 2009) che:

$$\left(\hat{\theta}_{MLE,\, jil}\right) = \frac{n\left(x_j^{(i)} \mid \pi_j^{(l)}\right)}{n\left(\pi_j^{(l)}\right)}$$

dove coerentemente con la notazione precedente:

$$n\left(\pi_j^{(l)}\right) = \sum_{i=1}^{k_j} n\left(x_j^{(i)} \mid \pi_j^{(l)}\right)$$

è la frequenza della configurazione $\pi_j^{(l)}$ in $\boldsymbol{X}$, per cui la stima di massima verosimiglianza di θ_{jil} è data da:

$$\hat{\theta}_{MLE,\, jil} = \frac{\text{frequenza della configurazione della famiglia}}{\text{frequenza della configurazione dei genitori}}$$

Seguendo l'approccio bayesiano invece, si procede all'apprendimento della distribuzione a posteriori dei CPP. Per convenienza si indichi con:

$$\theta_{j\cdot l} = \left(\theta_{j1l}, \ldots, \theta_{jk_jl}\right)$$

la distribuzione di probabilità sullo stato di X_j, data $\pi_j^{(l)}$la configurazione dei genitori.

Si consideri come distribuzione a priori su $\theta_{j\cdot l}$ una Dirichlet:

$$Dir\left(\alpha_{j1l},\ldots,\alpha_{jk_jl}\right).$$

la distribuzione a posteriori è ancora una Dirichlet:

$$\theta_{j\cdot l} \mid \left(\underline{x}_{(1)},\ldots,\underline{x}_{(n)}\right) \sim Dir\left(n\left(x_j^{(1)} \mid \pi_j^{(1)}\right)+\alpha_{j1l},\ldots,n\left(x_j^{(k_j)} \mid \pi_j^{(1)}\right)+\alpha_{jk_jl}\right).$$

La configurazione di famiglia al nodo j, per esempio: $n\left(x_j^{(1)} \mid \pi_j^{(l)}\right),\ldots,n\left(x_j^{(k_j)} \mid \pi_j^{(l)}\right)$ viene registrata come una memoria dell'esperienza passata.

La distribuzione a posteriori di $\theta_{j\cdot l}$ dipende solo dalle frequenze assolute della configurazione di famiglia al nodo j e non dalla configurazione ad ogni nodo. Ne segue che:

$$\left(n\left(x_j^{(1)} \mid \pi_j^{(l)}\right),\ldots,n\left(x_j^{(k_j)} \mid \pi_j^{(l)}\right)\right)_{l=1}^{q_j}$$

è una statistica predittiva sufficiente per $\theta_{j\cdot l}=\left(\theta_{j1l},\ldots,\theta_{jk_jl}\right)$. Seguendo queste considerazioni è possibile ricavare la distribuzione predittiva di un nuovo caso $\underline{x}_{n+1}$ utilizzando la densità a posteriori.

In presenza di un insieme di istanziazioni $\boldsymbol{x}$, θ_{jil} definito come :

$$\tilde{\theta}_{jil}=p\left(\left\{X_j=x_j^{(i)}\right\} \mid \left\{\Pi_j=\pi_j^{(l)}\right\},\{\boldsymbol{X}=\boldsymbol{x}\}\right)$$

sarà stimato come:

$$\tilde{\theta}_{jil}=p\left(\left\{X_{n+1,j}=x_j^{(i)}\right\} \mid \left\{\Pi_j=\pi_j^{(l)}\right\},\{\boldsymbol{X}=\boldsymbol{x}\}\right).$$

In Koski e Noble (2009) si dimostra che:

$$\tilde{\theta}_{jil}=\frac{n\left(x_j^{(i)} \mid \pi_j^{(l)}\right)+\alpha_{jil}}{n\left(\pi_j^{(l)}\right)+\sum_{i=1}^{k_j}\alpha_{jil}}$$

3.5.2 Stima del Modello grafico

La stima dei potenziali del modello rappresenta una sfida di gran lunga più semplice rispetto a quella che pone la ricerca di una struttura grafica compatibile con le relazioni causali esistenti tra i dati a disposizione.

In generale è possibile integrare anche nell'apprendimento del modello grafico le conoscenze a priori, ma nel seguito sarà trattato solo il caso più propriamente induttivo dell'apprendimento non vincolato, il cui obbiettivo è quello di dare il maggior spazio possibile alle evidenze.

Non tutte le strutture di indipendenza condizionale possono essere rappresentate tramite un grafo aciclico orientato ed inoltre tra le condizioni fondamentali per l'utilizzo del modello delle reti probabilistiche bayesiane vi è la richiesta di sufficienza causale, che non rappresenta altro se non la necessità di considerare tutti le variabili rilevanti per la spiegazione dei fenomeni indagati.

Una volta stabilita la fattibilità dell'apprendimento si presenta il problema pratico della sua realizzazione.

In generale i metodi di identificazione di un grafo aciclico orientato da una collezione di dati si dividono principalmente in due grandi categorie:

- *Approcci che si basano sulla ricerca di indipendenze condizionali.* Introdotto inizialmente da Spirtes e Glymour (1991); Spirtes *e altri* (2000) e da Pearl e Verma (Verma e Pearl, 1990; Pearl, 2000), tale approccio consiste nel cercare tutte le indipendenze condizionali che esistono fra le variabili analizzate allo scopo di ricostruire, in base ad esse, la struttura del DAG.

- *Approcci che impiegano una funzione di punteggio.* L'introduzione di una funzione di punteggio che si prefigge di quantificare l'adeguatezza di una rete bayesiana al problema da risolvere, riducendo la scelta del grafo alla massimizzazione di tale funzione.

La difficoltà associata all'apprendimento della struttura risiede nel fatto che si tratta di analizzare e confrontare una quantità enorme di possibili combinazioni di archi e di indipendenze condizionali. Robinson (1977) dimostra che per un numero d di nodi, il relativo numero di possibili grafi aciclici orientati corrisponde a:

$$N(d) = \sum_{i=1} d(-1)^{i+1} \binom{d}{i} 2^{i(d-i)} N\left(d-i\right) \tag{3.28}$$

Tale valore cresce in maniera super-esponenziale al crescere del numero dei nodi. Per $d = 5$ il valore di $N(d)$ è già pari a 29.281.

Le soluzioni finora impiegate riguardano principalmente euristiche che hanno lo scopo di ridurre lo spazio delle possibili soluzioni.

Un altro problema da non trascurare è l'orientamento degli archi, perché, esistono diversi grafi che rappresentano la medesima struttura di relazioni causali

(marvok-equivalenti), per cui è necessario, una volta ricavata la classe di equivalenza dalle evidenze empiriche, determinare le direzioni più probabili e trasmetterle lungo il grafo rispettando la struttura di *parentela*.

Ipotesi di base

Per costruzione (in base alla definizione di *I-map*) il grafo che costituisce una rete probabilistica bayesiana deve rispettare la condizione di markovianità locale, la quale impone che ciascun nodo sia indipendente da tutti i suoi non discendenti, dati i suoi genitori.

Tra le condizioni da rispettare, nell'associazione di una struttura grafica ad una collezione di dati è decisamente rilevante la seguente:

Definizione 3.39 (Sufficienza cusale).
Sia V l'insieme di tutte le variabili inserite in un grafo aciclico orientato $\mathcal{G} = (V, E)$ allo scopo di rappresentare le caratteristiche rilevate su un determinato insieme di dati.

Si dice che V è sufficiente per tali dati sse si verifica uno dei seguenti casi:
per ogni Y, causa comune a più variabili di V
- Y appartiene ancora a V
oppure
- Y esercita un'influenza costante su tutte le variabili incluse in V. △

che in altri termini significa che l'insieme delle variabili V è sufficiente per rappresentare tutte le relazioni di indipendenza che esistono nei dati rilevati.

Un'altra condizione essenziale è l'esistenza di *fedeltà* tra la distribuzione di probabilità soggiacente ai dati ed un eventuale DAG. Tutto ciò si traduce nell'ipotizzare l'esistenza si una grafo aciclico che sia una *P-map* del modello di indipendenza associato a tale distribuzione.

Classi di equivalenza di Markov

Con il teorema 3.7 è stata introdotta la nozione di classe di equivalenza di Markov, che consente di ricavare direttamente una caratterizzazione dei sottoinsiemi dello spazio di tutti i grafi possibili dato una certa collezione di variabili causali.

L'equivalenza di Markov per due grafi aciclici orientati corrisponde all'affermazione che tali grafi rappresentano le stesse relazioni di indipendenza condizionale. Tutto ciò consente di ridurre lo spazio di ricerca ai grafi Markov equivalenti e successivamente procedere all'orientamento degli archi seguendo opportuni criteri grafici.

Poiché tutti i grafi aciclici orientati equivalenti possiedono la stessa struttura di archi non orientati e le stesse serie convergenti, una classe di equivalenza può

essere rappresentata da un grafo aciclico parzialmente orientato (PDAG) avente una struttura non orientata comune e caratterizzato da archi reversibili (che appartengono ad una serie convergente o la cui inversione ne genera una) caratteristici della classe di equivalenza.

Il grafo così ottenuto viene detto completo o essenziale (CPDAG).

Grazie al concetto di equivalenza di Markov è possibile identificare, per mezzo degli algoritmi di ricerca che di seguito saranno presentati, una struttura di base generale semi-orientata che rispetti le indipendenze condizionali soggiacenti il grafo e successivamente applicare un algoritmo che stabilisca l'orientamento più opportuno degli archi reversibili.

Chickering (2002) propone un algoritmo per passare da un DAG rappresentativo di una rete bayesiana ad un PDAG, ma è altresì possibile ottenere un DAG da un PDAG come mostra l'algoritmo 3 proposto da Dor e Tarsi (1992)

Algoritmo 3 Algoritmo PDAG → DAG

Dati: grafo aciclico parzialmente orientato (PDAG)
Passo 1 $\mathcal{B} \leftarrow$ PDAG (grafo completamente orientato, estensione consistente del PDAG);
Passo 2. $\mathcal{A} \leftarrow$ lista degli archi inclusi in $\mathcal{B}$;
Passo 3. Ripetere:
Ricerca di un nodo X_i tale che valgano entrambi:
- non esiste alcun arco $X_i \rightarrow X_j$ in $\mathcal{A}$;
- per ogni X_i t.c. esiste $X_i - X_j$ in $\mathcal{A}$;
X_j è adiacente a tutti gli altri nodi adiacenti a X_i.

Se X_i non esiste ALLORA:
Il PDAG $\mathcal{B}$non ammette alcuna estensione completamente orientata;
ALTRIMENTI:
per ogni X_i t.c. esiste $X_i - X_j \in \mathcal{A}$:
$X_i \rightarrow X_j$ rientra in $\mathcal{A}$;
$\mathcal{A} \leftarrow \mathcal{A} \backslash (X_i, X_j)$;

•. FINO a che $\mathcal{A} \neq \varnothing$

3.5.3 Ricerca di Indipendenze Condizionali

In questo paragrafo vengono trattati i metodi di ricerca euristica basati sul principio della ricerca di indipendenze condizionali (detti anche metodi basati su vincoli[4]), sviluppati soprattutto nell'ambito dell'Intelligenza Artificiale da Pearl, Verma,

[4]Constraint Based Methods.

Spirtes, Glymour e Sheines.

Il processo di ricerca consiste nell'effettuare una serie di test di indipendenza condizionale per determinare le struttura di indipendenza e soprattutto le strutture convergenti ed in seguito propagare l'orientamento degli archi reversibili lungo tutto il grafo.

Test di Indipendenza Condizionale

I test statistici comunemente utilizzati per la ricerca di indipendenze condizionali si basano sul test χ^2 di indipendenza e sul test del rapporto di verosimiglianza G^2.

Data una coppia di variabili X ed Y, aventi rispettivamente h e k modalità, il test χ^2 mette a confronto due modelli statistici, quello osservato O_{xy} caratterizzato da una determinata distribuzione congiunta e quello teorico T_{xy} in cui la stessa distribuzione congiunta è fattorizzabile nel prodotto delle marginali. La differenza dei modelli (rappresentato l'uno dalle frequenze congiunte $n_{i,j}$ e l'altro dalle frequenze teoriche di indipendenza $\frac{n_{i,\cdot}\, n_{\cdot,j}}{n}$) viene calcolata tramite la distanza:

$$\chi^2 = \sum_{x=1}^{h} \sum_{y}^{k} \frac{(O_{xy} - T_{xy})^2}{T_{xy}} = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{\left(n_{i,j} - \frac{n_{i,\cdot\,,j}}{n} n_{i,j}\right)^2}{\frac{n_{i,\cdot}\, n_{\cdot,j}}{n} n_{i,j}}. \tag{3.29}$$

Tale statistica segue una distribuzione del χ^2, e se il valore teorico sotto ipotesi nulla è maggiore del valore osservato non è possibile rifiutare l'ipotesi si indipendenza fra le variabili.

Spirtes e Glymour (1991) propongono di sostituire la statistica del χ^2 con il rapporto di verosimiglianza:

$$G^2 = \sum_{x=1}^{h} \sum_{y}^{k} O_{xy} \ln\left(\frac{O_{xy}}{T_{xy}}\right) = 2 \sum_{i=1}^{h} \sum_{j=1}^{k} n_{i,j} \ln\left(\frac{n_{i,j}\, n}{n_{i,\cdot}\, n_{\cdot,j}}\right) \tag{3.30}$$

che segue ugualmente la distribuzione χ^2.

Per adattare la ricerca di dipendenze al caso di triplette di variabili X, Y, Z, dove Z si manifesta con un numero r di modalità, è possibile estendere il test χ^2 utilizzando la seguente statistica

$$\chi^2 = \sum_{x=1}^{h} \sum_{y}^{k} \sum_{z}^{r} \frac{(O_{xyz} - T_{xyz})^2}{T_{xyz}}. \tag{3.31}$$

Algoritmo PC e IC

L'individuazione delle indipendenze condizionali permette di costruire la struttura grafica delle reti probabilistiche bayesiane, ma il numero di test di indipendenza

da effettuare è esponenziale in rapporto al numero di variabili implicate, per cui sono stati proposti degli algoritmi che limitano lo spazio di ricerca delle soluzioni.

Spirtes *e altri* (2000) propongono la procedura PC (algoritmo 4) che limita i test di indipendenza alle relazioni di ordine 0 ($X_i \perp X_j$), poi alle relazioni condizionali di ordini superiori (($X_i \perp X_j | X_h$) e via dicendo).

Algoritmo 4 Algoritmo PC

Dati: $V = \{X_1, \ldots, X_n\}$

Passo 1. Costruzione di un grafo non orientato completamente connesso e tale che per ogni coppia di nodi X ed Y l'insieme dei nodi $S_{X,Y}$ che li rende condizionatamente indipendenti sia vuoto e abbia cardinalità 0 ($S_{X,Y} = \varnothing$ e $|S_{X,Y}| = k = 0$);

Passo 2. Per ogni coppia di nodi X ed Y, rimuovere l'arco congiungente sse dato l'insieme A_X dei nodi adiacenti ad X (escluso Y) vale che per ogni sottoinsieme $S \subset A$ di dimensione k, X ed Y risultano indipendenti condizionatamente ad S.

Assegnare $S_{X,Y} = S$.

Passo 3. Se è stato rimosso almeno un arco, ripetere il punto 2 incrementando k.

Passo 4. Per ogni coppia di nodi X ed Y non adiacenti, individuare i nodi $Z \in A_X \cap A_Y$ ed orientare gli archi in una connessione convergente se $Z \notin S_{X,Y}$.

Passo 5.Orientazione iterativa degli archi seguenti.

- Se X ed Y non adiacenti ma esiste Z tale che $X \to Z$ e $Y - Z$, orientare l'arco $\langle Y, Z \rangle$ verso Y in modo da ottenere una connessione seriale $Z \to Y$;
- Se esiste $Z \in A_X$ tale che sia possibile individuare una traiettoria orientata che porta da X a Z, orientare l'arco che li congiunge in maniera da evitare un ciclo.

Passo 6. Se esiste un arco orientato da X ad Y, per ogni $Z \in A_Y$, orientare l'arco $\langle Y, Z \rangle$ in direzione di Z se $Z \notin A_Y$.

L'algoritmo IC (Causazione Induttiva[5]) proposto da Pearl (2000) è basato sullo stesso principio ma procede aggiungendo gli archi in base al risultato dei test di indipendenza anziché eliminarli.

Questi metodi evidenziano una differenza esistente fra i tipi di archi potenzialmente inseribili nel DAG. Gli archi per cui viene subito determinata la direzione rappresentano relazioni causali molto più vincolanti degli altri. Quindi è possibile rilevare che questo metodo di ricerca stabilisce anche una valutazione della forza dei legami causali.

[5]Inductive Causation.

3.5.4 Metodi basati sulla Funzione di Punteggio

Una serie di metodi euristici di ricerca fondamentali nell'apprendimento della struttura grafica di una rete probabilistica bayesiana si basa su un'opportune funzione di punteggio (*score*) che assegna al grafo un valore indicativo sia della sua bontà di adattamento ai dati sia della complessità della sua struttura. La ricerca del DAG viene ricondotta alla massimizzazione di tale funzione.

Una funzione di punteggio per essere utilizzabile deve possedere la proprietà di *scomponibilità locale*, ovvero deve essere esprimibile come la somma dei punteggi locali al livello di ciascun nodo.

Poiché una ricerca esaustiva non è praticabile, gli algoritmi impiegati da questi metodi adottano diversi compromessi che vanno dalla riduzione dello spazio delle soluzioni (ricerca limitata a particolari configurazioni o imposizione di un ordinamento sui nodi) alla ricerca *greedy*, come viene definita in inglese, che cerca di coprire il maggior numero possibile di grafi candidati.

Prima di presentare alcuni utili algoritmi di ricerca di seguito vengono presentate le più utilizzate funzioni di punteggio per i grafi.

Un criterio abbastanza comune a tutti queste funzioni è il principio di parsimonia, riconducibile ancora una volta al *rasoio di Ockham*, in base al quale, a parità di bontà di adattamento, le soluzioni più complesse vengono penalizzate rispetto a quelle più semplici. Ne consegue che nella maggior parte dei casi le funzioni sono composte da una parte che tiene conto della verosimiglianza $\mathrm{L}(\mathcal{D}|\mathcal{G})$, dove $\mathcal{D}$ rappresenta la base dei dati ai quali il grafo $\mathcal{G}$ è associato, ed una parte che tiene conto del numero d di parametri necessari per rappresentare il grafo.

Alcuni esempi ampiamente utilizzati in generale per la valutazione della bontà di adattamento di un modello ai dati sono l'AIC[6]:

$$Score_{AIC}\left(\mathcal{D},\mathcal{G}\right) = \log \mathrm{L}\left(\mathcal{D}|\mathcal{G}\right) - d \tag{3.32}$$

ed il BIC[7]

$$Score_{BIC}\left(\mathcal{D},\mathcal{G}\right) = \log \mathrm{L}\left(\mathcal{D}|\mathcal{G}\right) - \frac{d}{2}\log N. \tag{3.33}$$

dove N rappresenta il numero delle osservazioni a disposizione.

Un criterio più attinente all'ambito dell'intelligenza artificiale è invece il principio di *lunghezza di descrizione minimale* (MDL)[8] introdotto da Rissanen (1978). Tale criterio afferma che il modello che meglio rappresenta i dati è quello che minimizza la somma di due termini:

- la lunghezza della codifica del modello

[6]Akaike Information Criterion.
[7]Bayesian Information Criterion.
[8]Minimum Description Length.

- la lunghezza della codifica dei dati quando sono rappresentati da tale modello.

Tale principio è stato declinato in diversi modi nella sua applicazione alle reti bayesiane.

In un ottica più soggettivista, impostando un modello bayesiano sulle ipotesi relative alla struttura del DAG ricercato, è possibile utilizzare come punteggio la probabilità a posteriori:

$$Score_{Bayes}\left(\mathcal{G},\mathcal{D}\right)=P\left(\mathcal{G}|\mathcal{D}\right)=\frac{P\left(\mathcal{D}|\mathcal{G}\right)P\left(\mathcal{G}\right)}{P\left(\mathcal{D}\right)}\propto P\left(\mathcal{D}|\mathcal{G}\right)P\left(\mathcal{G}\right). \tag{3.34}$$

Nel caso si assuma una distribuzione multinomiale sulle variabili e una distribuzione di Dirichlet sui parametri del modello, la probabilità a posteriori può essere scritta in forma esplicita (Cooper e Herskovits, 1992) e, sotto l'ipotesi di indipendenza locale dei parametri, vale che:

$$P\left(\mathcal{G}|\mathcal{D}\right)=\prod_{i=1}^{n}\left(\prod j=1^{q_i}\frac{\Gamma\left(a_{ij}\right)}{\Gamma\left(n_{ij}+a_{ij}\right)}\prod k=1^{r_i}\frac{\Gamma\left(n_{ijk}+a_{ijk}\right)}{\Gamma\left(a_{ijk}\right)}\right) \tag{3.35}$$

n è il numero delle variabili X_i

q_i è il numero delle possibili configurazioni di Π_{X_i}

r_i è il numero dei possibili stati di X_i

a_{ijk} sono i parametri della distribuzione di Dirichlet relativi alla variabile i al suo valore k e alla configurazione j dei suoi genitori.

n_{ij} sono le frequenze assolute e n_{ijk} sono le frequenze assolute osservate a parità di condizioni.

Algoritmi di ricerca: riduzione dello spazio delle soluzioni

Una volta determinata la funzione di punteggio più opportuna è necessario procedere all'ottimizzazione della stessa.

Tra i metodi che impiegano la riduzione dello spazio di ricerca delle soluzioni, un'euristica molto interessante è quella degli **alberi di copertura massima** (MWST)[9]: una struttura ad albero che passa per tutti i nodi e che massimizza un punteggio definito per tutti gli archi possibili.

L'applicazione del criterio varia a seconda della funzione di punteggio che si intende utilizzare. Un algoritmo particolarmente diffuso è quello di Kruskal (algoritmo 5) che parte da un insieme di n alberi mono-nodo e li fonde successivamente in relazione al peso degli archi determinato dalla funzione di punteggio.

[9]Maximum Weight Spanning Tree.

Algoritmo 5 Algoritmo Kruskal di costruzione dell'albero ottimale

Dati:

$V = \{X_1, \ldots, X_n\}$

$\mathcal{A}$: elenco degli archi non orientati $\langle X_i, X_j \rangle$ ordinati per punteggio non crescente;

Passo 1. Indicato con $\mathcal{T}_{X_i}$ un qualsiasi albero passante per X_i, per ogni X_i costruire un albero mono-nodo $\mathcal{T}_{X_i} = \{X_i\}$;

Passo 2. Inizializzare l'albero ottimale non orientato $\mathcal{G}$ come insieme vuoto.

Passo 3. Presa nell'ordine ogni coppia di nodi $(X_i, X_j) \in \mathcal{A}$:

Se l'albero $\mathcal{T}_{X_i}$ passante per X_i è diverso da $\mathcal{T}_{X_j}$ passante per X_j allora:

- aggiungere l'arco $\langle X_i, X_j \rangle$ alla struttura dell'albero ottimale:

$\mathcal{G} \leftarrow \mathcal{G} \cup \langle X_i, X_j \rangle$.

- Fondere i due alberi $\mathcal{T}_{X_i}$ e $\mathcal{T}_{X_j}$ in $\mathcal{T}' = \mathcal{T}_{X_i} \cup \mathcal{T}_{X_j}$ e assegnare $\mathcal{T}'$ come nuovo albero sia a X_i a per X_j:

$\mathcal{T}_{X_i} \leftarrow \mathcal{T}'$

$\mathcal{T}_{X_j} \leftarrow \mathcal{T}'$

L'albero di copertura massima che si ottiene è un albero non orientato che lega tutte le variabili. Per orientare gli archi di questa struttura è possibile impiegare l'algoritmo 3 oppure, più semplicemente è possibile scegliere un nodo radice e orientare ogni arco a partire da tale nodo, memorizzando il padre di ciascuna variabile, servendosi poi di questa informazione per orientare tutti gli altri archi.

I vantaggi di questo algoritmo risiedono nel fatto che esso permette di ottenere rapidamente un albero orientato molto vicino alla struttura originale, e soprattutto la configurazione risulta molto semplice e poco intricata. Inoltre poiché ogni nodo viene automaticamente inserito nella struttura, è possibile trovare legami altrimenti trascurati.

D'altro canto proprio questa proprietà diventa scomoda perché forza alcune variabili ad appartenere al grafo, anche quando esse non sarebbe realmente influenti sul problema analizzato.

Algoritmi di ricerca: algoritmi meta-euristici

Per le euristiche che non impiegano uno spazio delle soluzioni ridotto, il problema più comune è rappresentato dalla possibilità che il raggiungimento di un massimo locale fermi inopportunamente l'algoritmo di ricerca.

Per ovviare a tale problema sono stati sviluppati i cosiddetti algoritmi meta-euristici, che affiancano ai metodi di ricerca tradizionali delle tecniche che consentono di uscire dai punti di ottimo locali. Per ottenere tale risultato viene consentito

all'algoritmo di effettuare delle mosse peggiorative della funzione di punteggio, imponendo però di non effettuare più la mossa precedente (ovviamente migliorativa) allo scopo di impedire un ciclo infinito.

Algoritmo 6 Algoritmo Tabu Search

Passo 1. Generare una soluzione iniziale S e porre $k := 1$, $S^* := S$, $k^* := 1$, $TL := \{s\}$;
Passo 2. Se (condizione d'arresto) return;
Passo 3. Determinare la miglior soluzione $\overline{S}$ nell'intorno della soluzione corrente da cui vengono però escluse le soluzioni appartenenti alla lista tabu TL.
Passo 4. Se $Score\left(\overline{S}\right) < Score\left(S^*\right)$ porre $S^* := \overline{S}$, $k^* := k$;
Passo 5. Porre $k := k + 1$ e $S := \overline{S}$;
Passo 5. Inserire $\overline{S}$ in TL al posto della soluzione "più vecchia" e tornare al Passo 2;

Fanno parte di tale classe di algoritmi di ricerca introdotti da (Glover, 1990a,b,c), i **Tabu**[10] **Search** (algoritmo 6), i quali predispongono una lista contenente le ultime t mosse effettuate che diventano vietate (*tabu* appunto). Indicata con S^i la soluzione alla i-esima interazione, alla generica interazione k la *lista tabu* sarà così composta:

$$TL := \left\{S^{k-1}, S^{k-2}, \ldots, S^{k-t}\right\} \tag{3.36}$$

La lista così strutturata impedisce che si creino cicli di ampiezza t; la lunghezza t prende il nome di *tabu tenure*.

L'arresto di questo algoritmo è condizionato al raggiungimento di un numero, fissato a priori, di iterazioni o al fatto che sia trascorso un prefissato tempo di calcolo, oppure al raggiungimento della soluzione ottima S^*. Tali criteri possono essere ulteriormente raffinati.

Questi sono gli algoritmi meta-euristici che nella pratica forniscono i risultati migliori, benché non vi siano prove formali di convergenza.

Le prestazioni degli algoritmi Tabu dipendono da molti fattori quali la gestione della lista tabu, i metodi che permettono di verificare se la soluzione trovata è tabu, la lunghezza t della lista stessa (che è direttamente proporzionale sia ai cicli che vengono impediti sia al tempo di calcolo).

I metodi Tabu possono integrare euristiche che limitano lo spazio delle possibili soluzioni. Tra queste rientrano i metodi che imponendo un ordinamento sui nodi consentono di escludere le soluzioni in contraddizione con tale ordinamento, infatti

[10]Per indicare tale algoritmo si utilizza anche la parola Taboo.

se si stabilisce che X_i viene prima di X_j, allora l'arco orientato da X_j verso X_i non sarà ammissibile.

In questa maniera lo spazio delle soluzioni viene ridotto, ma resta ugualmente abbastanza ampio da richiedere l'impiego di un algoritmo di ricerca. In quest'ottica nasce il Tabu Order, un applicazione del Tabu Serch allo spazio delle soluzioni generato da archi ordinati.

Capitolo 4

Applicazione: Analisi di soddisfazione e fedeltà della clientela

L'applicazione proposta per effettuare il confronto fra i due approcci presentati nei capitoli precedenti fa riferimento all'ambito economico, ed in particolare si esplica in uno studio di soddisfazione e fedeltà della clientela.

Come introduzione alle analisi, viene fornita una breve presentazione del contesto applicativo, ovvero le maggiori teorie economiche che interpretano il concetto di soddisfazione del cliente e i metodi statistici proposti per la rilevazione empirica

Il passo immediatamente successivo è una descrizione dei dati, con particolare riguardo al significato delle variabili indagate ed al metodo utilizzato per la rilevazione. La fonte di tali dati è l'indagine campionaria effettuata sotto la direzione di Fornell *e altri* (1996) dal National Quality Research Center allo scopo di costruire l'*American Customer Satisfaction Index*[1] (di seguito indicato con l'acronimo ACSI).

Il campione in questione è ristretto a soli 4 settori dell'industria americana e precisamente:

- aviazione civile;
- forniture energetiche;
- telecomunicazioni via cavo e satellitari;
- settore alberghiero.

Per quanto riguarda le analisi vere e proprie, il modello ad equazioni strutturali viene stimato con il principale metodo *component based*, ovvero il PLS-PM nella versione implementata nel software XLSTAT 2010.1[2]. Per il modello di re-

[1] http://www.theacsi.org/

[2] Addinsoft, http://www.xlstat.com

ti bayesiane si impiega principalmente l'algoritmo di apprendimento Tabu Order implementato nel software BayesiaLab 4.6.8[3].

La sezione relativa ai risultati si compone di due analisi principali distinte una in ambito SEM e l'altra nell'ambito delle reti bayesiane, e di una sottosezione relativa all'integrazione degli approcci che prevede prima l'impiego dei grafi appresi con le reti bayesiane come proposte di nuovi schemi strutturali all'interno del modello SEM e poi l'applicazione dell'analisi di reti bayesiane sui punteggi ottenuti con la stima PLS.

L'intero studio viene effettuato avendo soprattutto cura di evidenziare non tanto la validità dei risultati dei singoli approcci, ampiamente accertata in letteratura, quanto le potenzialità della loro integrazione.

4.1 L'ambito applicativo e i dati

4.1.1 La soddisfazione e la fedeltà della clientela

Prima di presentare i risultati è opportuno fare una panoramica del contesto in cui si colloca l'analisi. Essendo la soddisfazione del cliente il fulcro dello studio in oggetto, ne vengono brevemente esposti la storia e le interpretazioni più comuni, nonché le più utilizzate tecniche di misurazione. Inoltre viene brevemente accennato il passaggio, in atto ormai da diverso tempo, dall'orientamento alla soddisfazione a quello verso la fedeltà del consumatore, oggi vero obiettivo delle strategie di marketing.

A questo punto si impone una piccola precisazione: di seguito, per brevità, con il termine prodotto di un azienda si intende il bene o il servizio fornito e il termine *consumatore* viene inteso, con una grossolana approssimazione, sinonimo di *utente*.

La qualità dei prodotti offerti è da sempre considerata come un fattore determinante negli scambi commerciali. Nelle economie primitive, il rapporto tra acquirente e venditore (che spesso coincide con il produttore) è tanto diretto che la mancata soddisfazione, derivante dalla percezione di qualità scadente, viene recepita in maniera *istantanea* sotto forma di lamentela e mancati introiti. Per tale ragione, in un primo momento la misurazione della qualità si è limitata per lo più agli aspetti tecnici, essendo diretta la conoscenza dell'opinione del cliente riguardo al prodotto offerto.

La rivoluzione industriale e l'avvento della produzione di massa hanno cambiato radicalmente i termini del problema aumentando, spesso in maniera abissale, la distanza fra produzione e consumo, con la conseguente necessità di trovare metodi statistici adeguati alla rilevazione dell'opinione della clientela sulla qualità offerta

[3]S.A. Bayesia, http://www.bayesia.com/

dalle aziende, la quale assume nuove interpretazioni; in tal senso, i punti di vista sono contrastanti e possono essere ricompresi in due grandi categorie.

L'approccio *ingegneristico* ritiene che l'aspetto tecnico-produttivo sia ancora centrale della definizione della qualità e, riallacciandosi ad una concezione primordiale, presuppone che i fattori determinanti siano legati solo ed esclusivamente alle proprietà intrinseche di un prodotto, quali le caratteristiche dei materiali, le prestazioni, i consumi, l'efficienza del servizio, la fruibilità ecc. Stando a tale concezione, avvalorata anche dalla necessità di omologazione della produzione di massa, nella fase di sviluppo di un prodotto i fattori propulsivi non sono le esigenze del consumatore ma gli standard tecnici di eccellenza produttiva, i quali determinano un sicuro successo sul mercato.

Dalla parte opposta, i sostenitori di una teoria della *percezione*, ritengono, in primo luogo, che potrebbe non esistere un prodotto in assoluto migliore, sotto tutti i punti di vista, ed in secondo luogo che la competenza tecnica del cliente potrebbe non essere sufficiente per apprezzare appieno lo sforzo produttivo.

Un esempio in tal senso viene dal settore informatico: esiste un computer migliore di ogni altro, a parità di fascia di prezzo? Tutti gli acquirenti che si collocano in quella fascia, sono in grado di apprezzarne la presunta eccellenza tecnica? Infine, la qualità di un prodotto rilevata dall'azienda potrebbe non coincidere con quella percepita dal consumatore. La misura della qualità viene quindi spostata da un piano oggettivo ad uno soggettivo, dove il concetto di qualità "tecnica" viene sostituito dal concetto di qualità "percepita", ponendo le basi per l'analisi della soddisfazione della clientela.

La rilevazione della soddisfazione della clientela diventa un'operazione sistematica negli anni Cinquanta, in alcune grandi aziende americane, e ben presto conquista una posizione privilegiata nella definizione delle strategie di mercato.

Il concetto di soddisfazione si evolve nel tempo perché non è interessante sapere soltanto se un prodotto ha soddisfatto oppure no il cliente, ma risulta più utile, ai fini delle strategie aziendali, conoscere la distribuzione degli utenti per diversi livelli di soddisfazione, le soglie di attesa e di tolleranza, la gradazione dell'importanza dei bisogni espressi al fine di porsi degli obiettivi di miglioramento dei risultati aziendali.

Vi sono diverse ragioni che determinano la necessità di una visione chiara delle cause del superamento delle aspettative e ancor più della presenza di aree di insoddisfazione, alcune delle quali sono:

- *Fornire una misura delle prestazioni.*

Le aziende valutano i propri risultati principalmente in base agli indici di bilancio, tra cui i più noti sono il fatturato e gli utili, ma ormai la soddisfazione della clientela viene considerata uno degli indicatori più attendibili del rapporto dell'azienda con il mercato, aggiungendo informazioni relative alle prestazioni.

Adattare la struttura di un'azienda alle esigenze del cliente produce stabilità nel lungo termine.

- *Fornire una previsione di profittabilità*

Oggi si assume comunemente l'esistenza di una relazione tra soddisfazione della clientela e fedeltà all'azienda, e tali indicatori possono essere considerati non solo utili per la valutazione delle prestazioni correnti di un'azienda, ma anche della profittabilità di più lungo periodo, soprattutto tenendo conto che spesso i costi di acquisizione di nuovi clienti sono molto superiori a quelli di mantenimento.

- *Facilitare il Bench marking*

In generale, un confronto sulla percezione dei consumatori sulla qualità offerta aiuta a determinare il vantaggio/svantaggio competitivo rispetto ai concorrenti e ad individuare la *best practice*, le strategie da prendere come punto di riferimento se si punta a migliorare la propria posizione sul mercato.

- *Supportare le strategie di marketing*

I punti precedenti entrano spesso a far parte di un'analisi strategica più ampia, quale ad esempio quella di SWOT. Essa è un supporto alle decisioni aziendali che si propone di evidenziare i punti di forza (**S**trenghts) e di debolezza (**W**eakness) di un'impresa allo scopo di far emergere opportunità (**O**pportunities) e minacce (**T**hreats) ed indicare azioni concrete che possono condurre ad un miglioramento delle prestazioni dell'azienda e soprattutto del suo posizionamento sul mercato. La misura della soddisfazione della clientela, in rapporto ai fattori che la determinano e ai fenomeni su cui influisce, rientra in un ambito più generale assumendo una posizione centrale.

L'aumento della competitività è possibile solo operando in accordo con i fabbisogni della clientela per determinarne la fedeltà e l'attaccamento al prodotto/azienda, ma rilevare la semplice presenza o assenza di soddisfazione non è molto utile dal punto di vista operativo, la percentuale di clienti soddisfatti non fornisce sufficienti informazioni, quindi per individuare aree di miglioramento è necessario avere una profonda conoscenza di tutti gli aspetti che determinano la soddisfazione del consumatore.

Dovendo trattare un concetto molto complesso che è possibile scomporre nelle sue manifestazioni più elementari è logico a questo punto far riferimento a quanto esposto principalmente nel capitolo 2 e trattare il problema della rilevazione della soddisfazione (e di conseguenza della fedeltà) come un problema di misurazione di una variabile latente tramite una o più variabili manifeste. In letteratura, infatti, i modelli per la misura della qualità in generale e della soddisfazione in particolare possono essere distinti in diverse categorie, le principali ricalcano il modello di misura e quello causale.

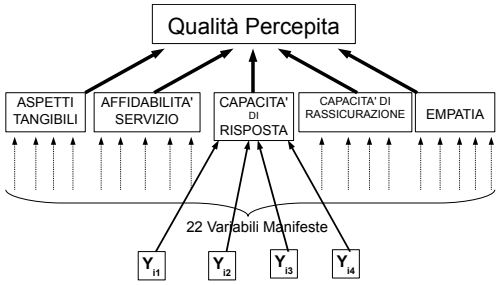

Figura 4.1: Modello di misurazione per la qualità percepita.

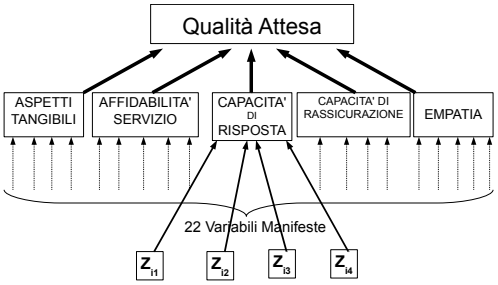

Figura 4.2: Modello di misurazione per la qualità attesa.

I modelli formativi

I modelli formativi assumono che la soddisfazione sia un concetto multidimensionale che è possibile ricostruire tramite una funzione algebrica (additiva o moltiplicativa) delle sue componenti elementari. Il riferimento è al modello di misura formativo presentato nel paragrafo 2.1.

Il modello più rappresentativo in tale categoria è il **SERVQUAL**, che viene proposto da Parasuraman *e altri* (1994) allo scopo di fornire una misura della soddisfazione del cliente prendendo in considerazione:

- le aspettative di qualità che esso ha su un prodotto;

- la percezione di qualità riscontrata sullo stesso.

Alla base di tale modello si pone il *paradigma della discrepanza* che prevede che la soddisfazione del cliente sia pari alla differenza fra le proprie aspettative e la qualità riscontrata. Ovviamente sia le aspettative sia le percezioni sulla qualità possono essere scomposte in elementi fondamentali, precisamente in base a 5 dimensioni della soddisfazione:

- *Aspetti tangibili*: aspettativa/soddisfazione relativa a caratteristiche fisiche del prodotto, quali estetica, funzionalità, consumi. Risulta evidente che l'aspetto tecnico è diventato solo una parte della questione.

- *Affidabilità*: capacità dell'azienda produttrice di rispettare tutte le condizioni di fornitura del bene/servizio al momento dell'offerta.

- *Capacità di risposta*: attitudine dell'azienda a venire prontamente incontro alle esigenze del cliente con il supporto tecnico e l'assistenza, ma anche in fase di ideazione del prodotto.

- *Capacità di rassicurazione*: attitudine a far sentire il cliente seguito con competenza e cortesia.

- *Empatia*: capacità dell'impresa di rispondere ai bisogni del cliente fornendo un servizio personalizzato.

Le 5 dimensioni sono a loro volta indagate con un numero di variabili manifeste tale da coprirne gli aspetti più importanti. Nell'ultima versione sono presenti in media 4 indicatori per dimensione, per un totale di 22.

La rilevazione prevede l'utilizzo di una scala Likert che va da "completamente in disaccordo" a "completamente d'accordo". Ovviamente, le rilevazioni vengono effettuate quanto più omogeneamente possibile, sia con riguardo alle aspettative che alla percezione.

Per completare il quadro viene richiesto all'intervistato di fornire un peso che rifletta l'importanza relativa di ciascuna dimensione nella determinazione della qualità percepita/attesa.

Il valore dell'indice di soddisfazione della clientela secondo il modello SERVQUAL si ottiene come media aritmetica ponderata:

$$CSI_{SERVQUAL} = \frac{\sum_{h=i}^{N} \left\{ \sum_{j=1}^{5} w_{jh} \left[\sum_{i=1}^{q_j} \frac{y_{jih} - z_{jih}}{q_j} \right] \right\}}{N}$$

dove:

y_{jih}: rappresenta la percezione dell'h-esimo rispondente in relazione all'indicatore i-esimo della dimensione j-esima.

z_{jih}: rappresenta l'aspettativa dell'h-esimo rispondente in relazione all'indicatore i-esimo della dimensione j-esima.

w_{jh}: rappresenta il peso di importanza relativa dell'h-esimo rispondente dell'area j-esima.

Questo modello trova dei forti limiti nel non tenere in considerazioni le reti di relazioni causali in cui si trova ad agire la soddisfazione e soprattutto nella completa esclusione dal modello della fedeltà del cliente, che è la variabile economica veramente centrale nelle analisi.

I modelli strutturali

Quando la soddisfazione della clientela viene inserita in uno schema di dipendenze in cui essa rappresenta alternativamente il fattore determinante per alcuni fenomeni e l'effetto di altri, il modello di misura diventa parte integrante di un più ampio modello causale e la costruzione di un indicatore di soddisfazione della clientela può avvenire tramite con la stima di un modello ad equazioni strutturali.

L'esempio più conosciuto di tale modello è un indice a valenza nazionale, proposto e sviluppato principalmente da Fornell (1992); Fornell *e altri* (1996); Fornell e Larcker (1981), prima per la Svezia (**C**ustomer **S**atisfaction **B**arometer), successivamente per Stati Uniti (**A**merican **C**ustomer **S**atisfaction **I**ndex) e infine adottato anche dalla Comunità Europea (**E**uropean **C**onsumer **S**atisfaction **I**ndex). Il progetto nasce con lo scopo di promuovere la ricerca della qualità e l'orientamento

al mercato dell'industria svedese a livello di Paese, ma il suo schema può essere applicato anche alle singole aziende senza alcuna difficoltà.

I concetti latenti presi in considerazione sono:
- la **soddisfazione** della clientela;
- le **aspettative** sul prodotto;
- la **qualità** del prodotto percepita dal cliente;
- il **valore** del prodotto percepito dal cliente;
- l'**immagine** del prodotto;
- la **fedeltà** del cliente al prodotto;
- i **reclami** generati dal prodotto e il loro trattamento.

Per quanto riguarda la soddisfazione, essa viene prima definita come concetto latente determinato da variabili rilevabili quali le soddisfazione complessiva del cliente, la soddisfazione delle aspettative, la coincidenza fra ideale di prodotto e reale qualità percepita, poi viene inserita in un diagramma di relazioni causali che deriva da un'intensa e pluriennale attività di ricerca empirica, i cui risultati possono essere sintetizzati nel modello generale in figura 4.3, dove vengono rappresentati la direzione ed il verso delle relazioni ed il modello di misura della sola variabile SODDISFAZIONE a scopo esemplificativo.

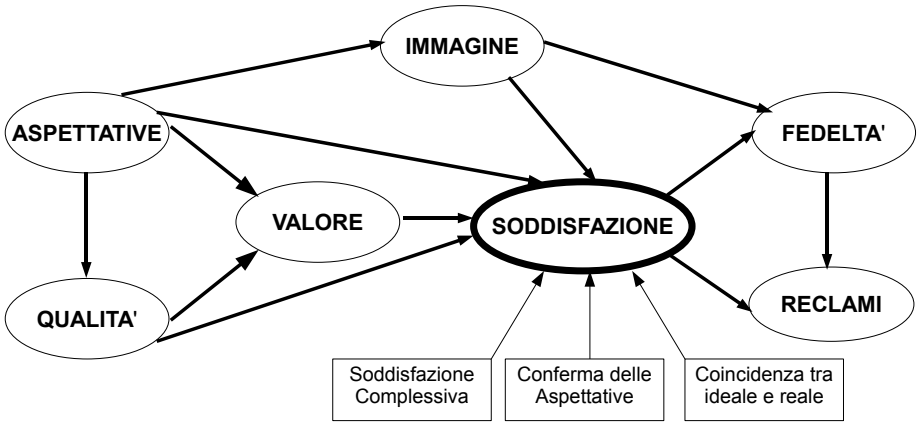

Figura 4.3: Modello di equazioni strutturali per la soddisfazione della clientela.

Per ogni nazione, per ogni mercato, per ogni azienda è necessario stimare la forza e la significatività di ciascun legame poiché, a priori, nessuno di essi è indispensabile. Ci sono prodotti per cui l'immagine non conta molto e viene esclusa dal modello esplicativo e ci sono settori in cui sono le aspettative ad essere ininfluenti.

Infine uno dei pregi maggiori dei modelli strutturali è l'aver portato la fedeltà del consumatore ad assumere un ruolo centrale nella ricerca. Grazie alla relazione

soddisfazione→fedeltà, tutti gli altri elementi dello schema causale possono essere interpretati come leve operative non solo per il raggiungimento dell'apprezzamento da parte della clientela ma anche per il mantenimento del rapporto di fiducia con la stessa.

In questo modo il modello diventa molto flessibile: permettendo di ricoprire un'ampia casistica presenta un quadro più ampio del rapporto cliente-azienda e si propone come un migliore indicatore di prestazioni e di profittabilità, nonché uno strumento strategico maggiormente operativo. Non solo rende possibili confronti interaziendali, ma anche con la media nazionale e fra Paesi diversi, assumendo così anche un ruolo macroeconomico.

Negli ultimi anni si è sviluppato un approccio ai modelli strutturali che consiste nell'applicazione di tecniche dell'intelligenza artificiale alla ricerca di strutture causali latenti tra i dati: le reti probabilistiche bayesiane.

L'obiettivo di tali modelli non è tanto quello di validare le relazioni esistenti tra i fenomeni quanto piuttosto di rilevarne di nuove, prendendo in considerazione legami tralasciati dai classici modelli lineari.

L'apporto creativo delle reti si presta ad essere integrato nella struttura "confermativa" dei SEM, fornendo degli spunti innovativi nell'organizzazione della struttura causale, soprattutto nelle relazioni tra variabili latenti e nella migliore definizione dei modelli di misura.

4.1.2 I dati

Come anticipato il database è parte della più generale rilevazione effettuata trimestralmente dal National Quality Research Center allo scopo di stimare l'ACSI.

La popolazione campionata coincide con quella presente sul territorio statunitense (il periodo particolare viene omesso per motivi di riservatezza) la quale viene intervistata per ottenerne l'opinione su determinati beni e servizi di diversi settori industriali strategici dell'economia americana. I settori analizzati in questa sede sono quelli dell'aviazione civile, delle forniture energetiche, delle telecomunicazioni via cavo/satellitari ed il settore alberghiero.

Nel databese analizzato, per ogni settore vengono valutate 4 aziende e per ogni azienda sono intervistati 250 individui, per un totale di 4000 rilevazioni.

Gli intervistati esprimono il proprio giudizio su una scala di gradimento/soddisfazione crescente in dieci punti sugli aspetti descritti in tabella 4.1

Nel caso particolare non sono presenti i fenomeni IMMAGINE e LAMENTELE. Inoltre, per varie ragioni vengono escluse dall'analisi le variabili di tolleranza. In primo luogo è stato riscontrato un numero molto elevato di mancate risposte per

Etichetta	Variabile Manifesta	Variabile Latente
soddis	Soddisfazione complessiva	**SODDISFAZIONE**
conferma	Conferma delle aspettative	
ideale	Vicinanza all'ideale di prodotto	
commlessivoA	Aspettative di qualità complessiva	**ASPETTATIVE**
personalizA	Aspettative di personalizzazione del prodotto	
noproblemA	Aspettative di affidabilità	
complessivoQ	Qualità complessiva	**QUALITA'**
personalizQ	Soddisfazione dei bisogni personali	
noproblemQ	Assenza di problematiche	
pq	Prezzo data la qualità	**VALORE**
qp	Qualità dato il prezzo	
riacquisto	Intenzioni di riacquisto	**FEDELTA'**
toll_%piu	Tolleranza ad un aumento di prezzo	
toll_%meno	Tolleranza ad una diminuzione di prezzo della concorrenza	

Tabella 4.1: Variabili manifeste e relative variabili latenti rilevate nell'indagine ACSI.

entrambe: il 38% per la tolleranza ad un aumento percentuale di prezzo del prodotto analizzato e il 72% per la tolleranza alla diminuzione percentuale di prezzo dei prodotti concorrenti.

Questi dati, uniti alla considerazione che la persona media ha scarsissima familiarità con le percentuali, hanno fatto sospettare una scarsa comprensibilità della formulazione della domanda e la conseguente incapacità di rispondere per l'intervistato. Sotto queste ipotesi nasce il dubbio che anche le risposte complete non siano del tutto attendibili, da cui la decisione di eliminare del tutto le due variabili dall'analisi, lasciando solo l'intenzione di riacquisto come indicatrice di FEDELTA'.

Analisi descrittive

Le principali caratteristiche socio-demografiche rilevate sugli intervistati vengono riassunte di seguito.

La percentuale di donne (59,7%) è superiore di 20 punti a quella degli uomini presenti nel campione. L'etnia prevalente è quella bianca, con l'85.6% della popo-

lazione, seguita da quella afroamericana con 6.23% e da quella ispanica, rilevata separatamente, con 4.82%.

Per quanto riguarda il livello d'istruzione, questo vede una prevalenza di titoli di studio medio alti: meno del 25% del campione non ha un titolo universitario (figura 4.4).

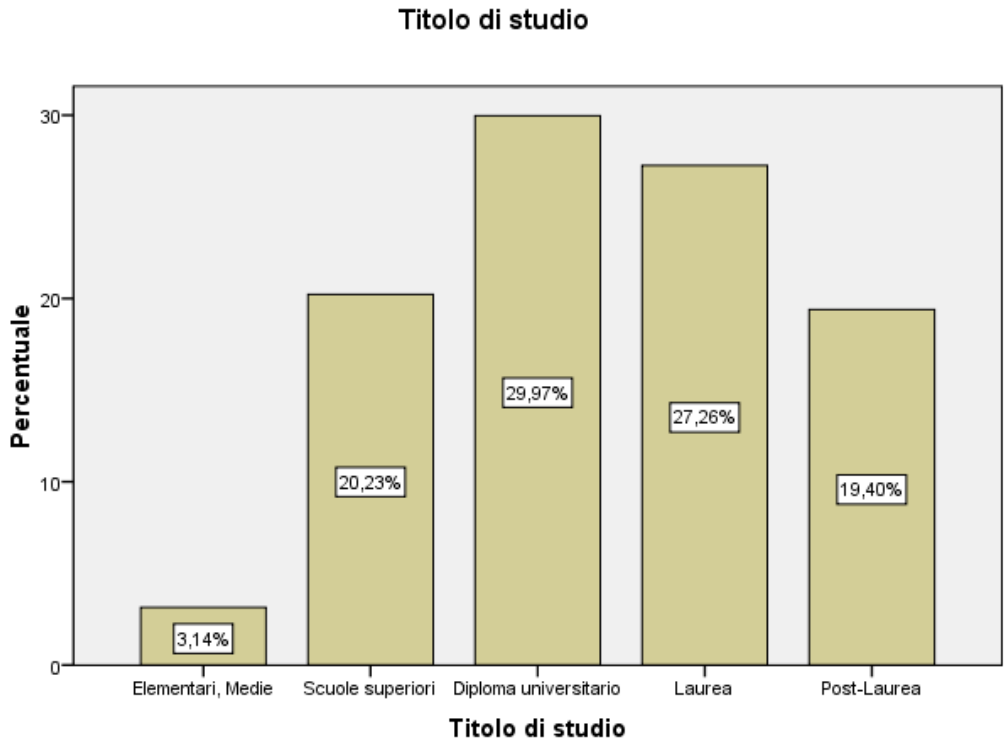

Figura 4.4: Grafico a barre per la distribuzione percentuale dei livelli d'istruzione.

La rilevazione della fascia di reddito mostra una forte rappresentatività della fascia più alta (i percettori di un reddito superiore ai 100.000$ rappresentano il 22.4% del campione), seguita dai redditi compresi, con il 18.07%, dai redditi compresi tra i 40 ed i 60.000$ e con il 16.02% dai redditi tra i 60 e gli 80.000$. Le restanti fasce di reddito attestate intorno all'11% (figura 4.5).

L'età dei soggetti si aggira intorno ad una media di 48 anni, con una deviazione standard di 15 anni circa.

La presenza di dati mancanti nelle variabili rimaste dopo l'eliminazione delle due variabili di tolleranza è alquanto irrisoria. Per le variabili di giudizio si va da un massimo del 3% ad un minimo dello 0.1%(tabella 4.2). Lo stesso accade per le

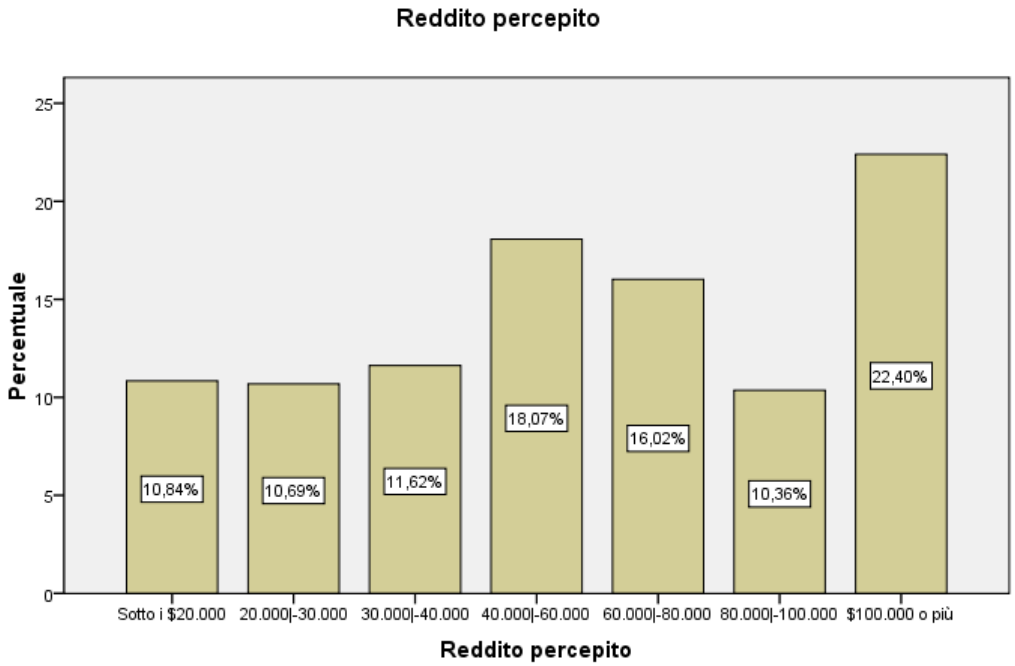

Figura 4.5: Grafico a barre per la distribuzione percentuale delle classi di reddito.

variabili socio-demografiche (tabella 4.3), fatta eccezione per il reddito, da sempre un dato *sensibile*, che raggiunge il 17% di dati mancanti.

Dopo aver effettuato un'analisi dei dati mancanti con il test MCAR[4] di Little (1988), è possibile ipotizzare che i dati presentino una struttura MAR[5] e procedere all'imputazione degli stessi. Il metodo è uno dei più semplici nonché più diffusi, ovvero quello della sostituzione dei mancanti con la mediana della variabile.

La scelta è caduta sulla mediana, non solo perché è un indice più robusto ma anche perché preserva le caratteristiche delle variabili, le quali benché comunemente trattate come continue, hanno una natura ordinale. Inoltre tale metodo utilizza come valore sostitutivo un elemento già presente nel database, evitando di introdurre una nuova modalità.

Sempre per la stessa regione anche per le variabili socio-demografiche vengono utilizzate la mediana e, ove opportuno la moda, per imputare i dati mancanti.

La scarsa presenza di dati mancanti non sembra richiedere, a parere dell'autrice, l'utilizzo di tecniche d'imputazione più potenti, ma computazionalmente più

[4]Missing Completely at Random.
[5]Missing At Random.

Variabili manifeste	**N. osserv. valide**	**Mancanti** N.	%	**Media**	**Dev.St.**	**Min**	**Max**
soddis	3988	12	0,3%	7,56	2,22	1	10
conferma	3972	28	0,7%	6,69	2,24	1	10
ideale	3886	114	2,9%	6,45	2,39	1	10
complessivoA	3917	83	2,1%	7,52	1,92	1	10
personalizzA	3947	53	1,3%	7,78	2,00	1	10
noproblemA	3920	80	2,0%	7,25	2,44	1	10
complessivoQ	3995	5	0,1%	7,70	2,05	1	10
personalizzQ	3996	4	0,1%	7,63	2,31	1	10
noproblemQ	3976	24	0,6%	7,69	2,46	1	10
pq	3898	102	2,6%	6,60	2,51	1	10
qp	3925	75	1,9%	7,14	2,34	1	10
riacquisto	3879	121	3,0%	7,07	2,79	1	10

Tabella 4.2: Statistiche descrittive per le variabili di giudizio.

impegnative come può essere l'imputazione multipla (Little e Rubin, 2002).

Tale operazione preliminare si è resa opportuna, benché esistano forme di trattamento dei dati mancanti anche all'interno degli stessi algoritmi di stima dei modelli qui analizzati (sia per i SEM che per le reti) che avrebbero senza dubbio portato risultati migliori della solo imputazione con la mediana.

La scelta è stata determinata in questo caso dalla necessità di utilizzare per le analisi comparativa preliminare un insieme di dati comuni, lasciando ad un eventuale lavoro successivo l'onere di confrontare i due diversi metodi di trattamento dei dati mancanti.

Prima di procedere con le analisi causali vengono presentati in tabella 4.2 alcuni indici descrittivi dei giudizi degli intervistati sulle variabili analizzate.

Appare evidente che il giudizio espresso in genere non è certo negativo, oscillando tra il sufficiente (6) ed il buono (circa 8).

Restando in tema di giudizi, poiché l'interpretabilità dei risultati dell'apprendimento tramite reti bayesiane migliora con il diminuire delle modalità delle variabili analizzate, è stata effettuata una ricodifica dei dati allo scopo di passare da 10 a sole 4 modalità, come segue:

1: Giudizio nettamente negativo ($1 \vdash 4$)

2: Giudizio tendenzialmente sufficiente ($5 \vdash 6$)

Variabile	**N. osserv. valide**	**Mancanti** N.	%	**Moda**	**Mediana**	**Media**
Età	3902	98	2,5%	50	47.5	48.1
Genere	4000	0	0%	F	-	-
Reddito	3321	679	17%	$> 100.M\$$	$40 - 60.M\$$	-
Livello istruzione	3980	20	0,5%	Diploma univer.	Diploma univer.	-
Etnia	3935	65	1,6%	bianco	-	-
Etnia ispanica	3973	27	0.7%	NO	-	-

Tabella 4.3: Statistiche descrittive per le variabili socio-demografiche.

4: Giudizio buono ($7 \vdash 8$)
4: Giudizio ottimo($9 \vdash 10$)

Tale trasformazione, non compromette il sistema di valutazione se si tiene conto che ha lo scopo di inverte l'operazione comunemente effettuata dagli intervistati, i quali, generalmente, prima formulano un giudizio di valore e poi lo convertono in un numero sulla scala richiesta.

Avendo presentato i dati, di seguito si procede all'esposizione dei risultati.

4.2 Il modello strutturale: stima PLS-PM

4.2.1 Le opzioni dell'algoritmo

I dati rilevati per l'ACSI vengono comunemente modellati da equazioni strutturali, i cui parametri sono stimati con il metodo PLS-PM. Adottando il metodo deduttivo, in base alle conoscenze a priori a disposizione nell'ambito di ricerca. In questa sede si formula l'ipotesi che il modello segua lo schema classico dell'analisi della soddisfazione e della fedeltà della clientela in ambito ACSI (ridotto, come giustificato nel paragrafo 4.1.2) che viene riportato in figura 4.6. Tale scelta è supportata dai molteplici studi di comportamento, oltre che statistici, che hanno affiancato lo sviluppo della metodologia di indagine utilizzata per l'ACSI. Le domande del questionario sono state costruite e testate appositamente per poter catturare le manifestazioni delle variabili latenti d'interesse, per cui il modello di misura è scelto quasi per costruzione.

Nonostante ciò, il modello riflessivo adottato non è stato accettato a priori, ma valutato alla luce di alcune analisi preliminari. La scelta del modello riflessivo viene avvalorata dal fatto che, come mostra la tabella 4.4, l'analisi delle componenti principali per blocchi evidenzia la presenza di una sola componente per ciascun

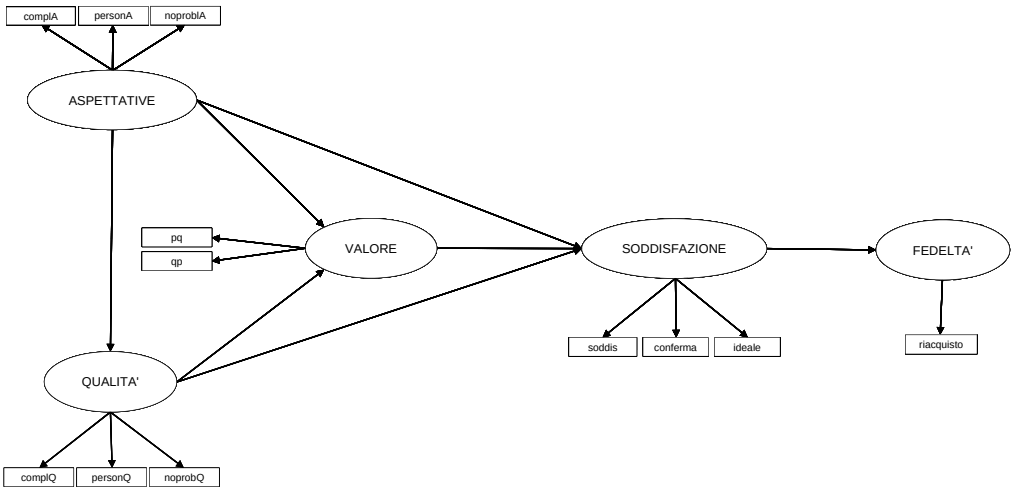

Figura 4.6: Schema ACSI per l'analisi della soddisfazione e della fedeltà della clientela.

blocco e sia l'α di Cronbach che il ρ di Dillon-Goldstein sono praticamente tutti maggiori di 0.70 tranne l'α delle ASPETTATIVE che è di poco minore. In definitiva sembra confermata la consistenza interna del modello, ed in particolare l'unidimensionalità e l'omogeneità dei blocchi.

Per quanto riguarda lo schema causale che lega i costrutti teorici, esso proviene da anni di ricerche di marketing. Il metodo PLS-PM non fornisce un potente apparato di test, come fanno i metodi *covariance based*, perché non è noto un ben identificato criterio di ottimizzazione globale a cui far riferimento per la costruzione di una funzione di adattamento. D'altro canto tale metodo di stima consente di utilizzare validi test di adattamento per valutare la conformità ai dati senza dover chiamare in causa pesanti ipotesi probabilistiche.

Una volta validata la stima sarà possibile interpretare i risultati, ovvero:

- i pesi dei fattori nella determinazione del relativo concetto latente;
- i coefficienti del modello strutturale, che legano tra loro le variabili latenti;
- i punteggi, calcolati per ogni individuo, delle variabili latenti.

Soprattutto quest'ultimo valore è interessante perchè permette di fornire una valutazione media sulle aziende e sui settori d'industria analizzati da parte dei consumatori.

Variabile latente	Dimensioni	α di Cronbach	ρ di D.G.	Autovalori(ACP)
ASPETTATIVE	3	0.667	0.820	1.525 0.674 0.310
QUALITA'	3	0.826	0.897	2.167 0.491 0.255
VALORE	2	0.870	0.940	1.828 0.235
SODDISFAZIONE	3	0.867	0.919	2.237 0.363 0.232

Tabella 4.4: Consistenza interna dei blocchi.

Il software utilizzato per l'analisi è XLSTAT, programma statistico ideato da Addinsoft come componente aggiuntiva di Excel che sfrutta la potenza del foglio di calcolo come motore per effettuare avanzate analisi statistiche. Nello specifico viene impiegato il modulo XLSTAT-PLSPM, particolare applicazione che implementa i risultati delle più recenti scoperte nel campo dei PLS-PM.

Prima di procedere con l'analisi è importante definire le impostazioni iniziali.

Per quanto riguarda i dati, il trattamento preliminare di imputazione dei mancanti e di raggruppamento è stato effettuato proprio per rendere superflue le operazioni specifiche di ciascun algoritmo e fornire un identico dataset di partenza. In particolare, non è necessario utilizzare le opzioni di trasformazione fornite dal software per normalizzare le variabili in quanto le stesse sono state rilevate, e poi raggruppate, mantenendo la stessa unità di misura, le loro medie sono confrontabili e si vuole fare in modo che le rispettive varianze forniscano indicazione dell'importanza relativa. In definitiva, nell'analisi saranno utilizzate le variabili originali.

Come conseguenza di un utilizzo delle variabili originali nell'analisi, si ritiene utile ricavare i punteggi dei costrutti latenti utilizzando i pesi normalizzati in modo da ottenere i risultati nella stessa unità di misura delle manifeste e da rendere confrontabili le valutazioni medie.

Relativamente all'algoritmo di stima esterno del modello, avendo scelto lo schema riflessivo, viene utilizzato il Modo di stima A, con peso iniziale delle variabili manifeste pari al primo autovalore. Per quanto riguarda il modello interno, invece, lo schema utilizzato è quello del centroide poiché la struttura interna è ben rodata e non ci sono coefficienti di correlazione molto vicini allo zero.

Infine, dal momento che non sono richieste assunzioni probabilistiche per l'applicazione del metodo PLS-PM, è necessario fare ricorso a procedure di validazione

incrociata [6] per stimare la significatività dei parametri e degli indici usati. In particolare, nell'analisi sono stati utilizzati, ove necessario, il metodo bootstrap con 200 replicazioni e una procedura blindfolding con passo pari a 30.

4.2.2 Validazione del modello di misura

Una volta effettuata l'analisi, ma prima dell'interpretazione dei risultati, si presenta la necessità di valutare la bontà del modello ottenuto, nel caso particolare della stima PLS-PM con l'obiettivo di ottenere una buona capacità predittiva. La validazione del modello viene effettuata separatamente per ogni sua parte, quindi si ottiene una validazione del modello di misura, del modello strutturale e del modello complessivo. Dopo aver verificato che ogni modello di misura sia ben specificato

	ASPETTATIVE	QUALITA'	VALORE	SODDISFAZIONE	FEDELTA'
complessivoA	**0.769**	0.399	0.350	0.416	0.285
personalizzA	**0.826**	0.470	0.364	0.444	0.281
noproblemA	**0.745**	0.385	0.247	0.302	0.180
complessivoQ	0.512	**0.871**	0.611	0.751	0.518
personalizzaQ	0.477	**0.897**	0.612	0.750	0.519
noproblemQ	0.404	**0.822**	0.477	0.573	0.385
pq	0.352	0.575	**0.940**	0.688	0.511
qp	0.414	0.660	**0.943**	0.747	0.544
soddis	0.455	0.787	0.741	**0.912**	0.608
conferma	0.419	0.696	0.676	**0.893**	0.550
ideale	0.437	0.649	0.612	**0.862**	0.553
riacquisto	0.315	0.550	0.561	0.642	**1.000**

Tabella 4.5: Cross-loadings

è necessario controllare anche che ciascuna variabile manifesta sia monofattoriale, ovvero che misuri un solo fattore. Nella tabella 4.5 sono riportati per ciascuna variabile sia i loadings[7] che i loadings incrociati[8] e si nota come, per ogni variabile, la relazione con la variabile latente del proprio blocco di riferimento sia più forte rispetto a quella con i costrutti restanti.

Per la validazione del modello esterno è opportuno considerare i valori assunti dalle comunalità e dalla comunalità media. Il VALORE spiega larga parte della variabilità delle sue manifeste (in media l'88.6%), mentre le ASPETTATIVE si limitano a spiegarne in media il 61%. Nel complesso si può però affermare che il

[6] Cross-validation

[7] Misura del contributo che ogni singolo indicatore apporta separatamente alla rilevanza della corrispondente variabile latente.

[8] Misura del contributo che ogni indicatore apporta separatamente alla rilevanza delle variabili latenti ed esso non direttamente collegate.

Variabile latente	**Variabili manifesta**	**Comunalità**	**Comunalità Media (AVE)**
ASPETTATIVE	complessivoA	0.592	0.610
	personalizzA	0.682	
	noproblemA	0.556	
QUALITA'	complessivoQ	0.758	0.746
	personalizzQ	0.805	
	noproblemQ	0.676	
VALORE	pq	0.883	0.886
	qp	0.889	
SODDISFAZIONE	soddis	0.832	0.791
	conferma	0.797	
	ideale	0.743	
Media			0.746

Tabella 4.6: Comunalità

modello di misura sia ben strutturato, infatti come da tabella 4.6, l'AVE risulta sempre maggiore di 0.5.

Per il modello strutturale si procede con l'analisi degli indici di ridondanza per ciascun blocco, che si affianca agli R^2 riportati in tabella 4.7, allo scopo di valutare la bontà del modello strutturale, tenendo però in considerazione anche il modello di misura, poiché la ridondanza valuta la capacità esplicativa di ogni variabile latente esogena sulle variabili manifeste legate alle latenti da essa dipendenti. Le

Variabile latente	R^2	R^2 corretto	R^2 (Bootstrap)	Limite inferiore (95%)	Limite superiore (95%)
QUALITA'	0.288	0.288	0.287	0.257	0.313
VALORE	0.435	0.435	0.434	0.409	0.459
SODDISFAZIONE	0.742	0.742	0.742	0.726	0.757
FEDELTA'	0.412	0.412	0.412	0.379	0.441
Media	0.470				

Tabella 4.7: R^2

variabili latenti mostrano nel complesso dei buoni valori dell'R^2, migliori per la SODDISFAZIONE e peggiori per la QUALITA'. Anche in tabella 4.8 le ridondanze mostrano come il *riacquisto* (unica manifesta della FEDELTA') e le manifeste della SODDISFAZIONE siano tutto sommato ben spiegate, mentre non si può dire altrettanto delle variabili di VALORE e di QUALITA' che presentano dei valori piuttosto bassi di ridondanza. Come già anticipato nel capitolo 2 è necessario

Variabile latente	Variabili manifeste	Ridondanze	Ridondanze Medie
QUALITA'	complessivoQ	0.218	0.215
	personalizzQ	0.232	
	noproblemQ	0.195	
VALORE	pq	0.384	0.386
	qp	0.387	
SODDISFAZIONE	soddis	0.617	0.587
	conferma	0.592	
	ideale	0.552	
FEDELTA'	riacquisto	0.412	0.412
Media			0.400

Tabella 4.8: Ridondanze

fornire una validazione incrociata degli indici utilizzati. Nello specifico si nota come le stime bootstrap dell'R^2 siano abbastanza stabili mentre, come risulta dalle tabelle 4.9 e 4.10, se le ridondanze cross-validate subiscono un lieve peggioramento (rendendo le ridondanze medie della quantità non interpretabili), le comunalità subiscono una più drastica riduzione [9].

Variabile latente	Variabili manifeste	Comunalità	Comunalità medie	Comunalità blindfolding	Com. Media blindfolding
ASPETTATIVE	complessivoA	0.592	0.610	0.277	0.222
	personalizzA	0.682		0.311	
	noproblemA	0.556		0.124	
QUALITA'	complessivoQ	0.758	0.746	0.511	0.463
	personalizzQ	0.805		0.516	
	noproblemQ	0.676		0.380	
VALORE	pq	0.883	0.886	0.527	0.524
	qp	0.889		0.521	
SODDISFAZIONE	soddis	0.832	0.791	0.582	0.543
	conferma	0.797		0.564	
	ideale	0.743		0.485	

Tabella 4.9: Validazione incrociata delle comunalità tramite blindfolding.

Sulla validità del modello interno è possibile fare ancora delle considerazioni. Come mostra la tabella 4.11, le variabili latenti hanno validità discriminante ovvero misurano cinque concetti differenti. In ultimo è possibile valutare la bontà di adattamento complessiva del modello grazie all'indice di GoF, per il cui calcolo vengono presi in considerazione sia il modello esterno che quello interno. Il

[9] entrambi calcolati utilizzando la procedura di blindfolding.

Variabile latente	Variabili manifeste	Ridondanze	Media Ridondanze	Ridondanze blindfolding	Media Ridondanze blindfolding
QUALITA'	complessivoQ	0.218	0.215	0.172	0.090
	personalizzQ	0.232		0.113	
	noproblemQ	0.195		0.010	
VALORE	pq	0.384	0.386	0.282	0.335
	qp	0.387		0.393	
SODDISFAZIONE	soddis	0.617	0.587	0.601	0.524
	conferma	0.592		0.481	
	ideale	0.552		0.492	
FEDELTA'	riacquisto	0.412	0.412	0.366	0.366

Tabella 4.10: Validazione incrociata delle ridondanze tramite blindfolding

	ASPET	QUAL	VAL	SODD	FED	Comunalità Media (AVE)
ASPETTATIVE	**1**	0.288	0.166	0.242	0.099	0.610
QUALITA'	0.288	**1**	0.431	0.642	0.302	0.746
VALORE	0.166	0.431	**1**	0.581	0.315	0.886
SODDISFAZIONE	0.242	0.642	0.581	**1**	0.412	0.791
FEDELTA'	0.099	0.302	0.315	0.412	**1**	
Comunalità Media(AVE)	0.610	0.746	0.886	0.791		**0**

Tabella 4.11: Validità discriminante. Tabella delle correlazioni al quadrato.

PLS-PM non è ancora orientato verso l'ottimizzazione dell'indice globale GoF che fornisce ugualmente un buon indicatore empirico.

Per il modello nell'insieme, i valori assunti dal GoF sono abbastanza alti, come da tabella 4.12: un GoF assoluto pari a 0.592 è da considerarsi un buon risultato. A maggior ragione un valore pari a 0.979 per il GoF relativo è indice di bontà del modello, dovuta in maniera pressoché omogenea sia al modello interno (0.980) che a quello esterno (0.999). Sempre dalla tabella 4.12 è possibile leggere la significatività di questo parametro stimata con il metodo bootstrap.

	GoF	GoF (Bootstrap)	Limite inferiore (95%)	Limite superiore (95%)
Assoluto	0.592	0.591	0.581	0.601
Relativo	0.979	0.978	0.974	0.981
Modello esterno	0.999	0.999	0.999	0.999
Modello interno	0.980	0.979	0.975	0.982

Tabella 4.12: Bontà globale di adattamento

4.2.3 Risultati della stima

Dopo aver effettuato una procedura di validazione del modello è infine possibile commentare la stima dei parametri ottenuti.

Per quanto riguarda il modello di misura, in base ai risultati riportati in tabella 4.13, si può rilevare che tutte le variabili contribuiscono più o meno allo stesso modo alla determinazione della corrispondente latente.

Variabile latente	**Variabile manifesta**	**Peso esterno**	**Peso est. (norm.)**	**Peso est. (Bootstrap)**	**Lim. Inf. 95%**	**Lim. Sup. 95%**
ASPETTATIVE	complessivoA	0.447	0.317	0.448	0.425	0.475
	personalizzA	0.515	0.365	0.515	0.493	0.542
	personalizzA	0.449	0.318	0.450	0.423	0.475
QUALITA'	complessivoQ	0.396	0.336	0.397	0.385	0.410
	personalizzQ	0.426	0.361	0.426	0.415	0.440
	noproblemQ	0.357	0.303	0.358	0.346	0.369
VALORE	pq	0.504	0.482	0.505	0.494	0.516
	qp	0.543	0.518	0.544	0.533	0.556
SODDISFAZIONE	soddis	0.414	0.357	0.414	0.405	0.423
	conferma	0.375	0.324	0.375	0.366	0.386
	ideale	0.369	0.319	0.369	0.358	0.380
FEDELTA'	riacquisto	0.907	1.000	0.907	0.896	0.921

Tabella 4.13: Pesi esterni

Per quanto riguarda le relazioni strutturali, in figura 4.7 sono riportati i coefficienti di path (che risultano tutti significativi): lo spessore delle frecce rappresenta graficamente il contributo diretto apportato a ciascuna variabile.

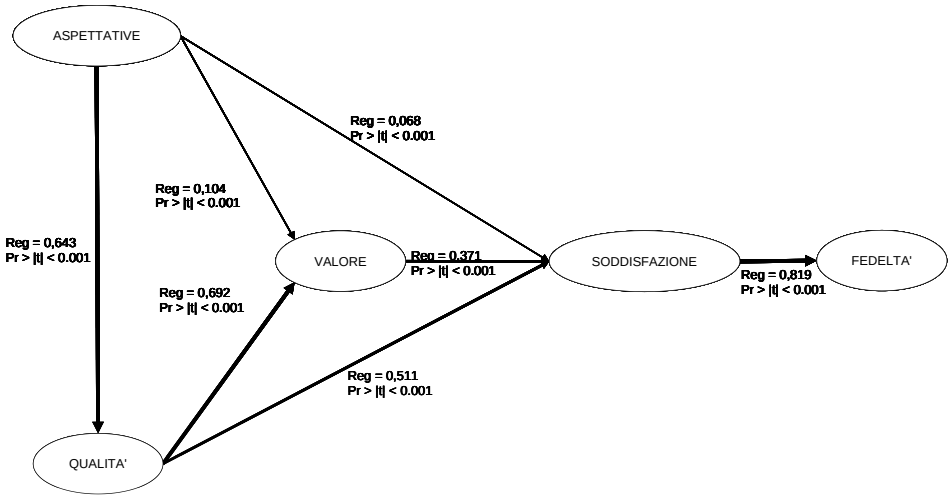

Figura 4.7: Stima PLS-PM dei parametri del modello strutturale.

Si può notare come le ASPETTATIVE, al contrario della QUALITA', abbiano un'influenza diretta piccola, benché significativa, sulla SODDISFAZIONE la cui equazione è data da:

SODDISFAZIONE =
= -0.043+0.068*ASPETTATIVE+0,511*QUALITA'+0,371*VALORE.

Se però si considerano anche gli effetti totali riportati in tabella 4.14, appare chiaro che le ASPETTATIVE agiscano anche indirettamente, raggiungendo un effetto totale consistente (0.600).

La SODDISFAZIONE è un fattore determinante per la FEDELTA' con un effetto pari a 0.819 e grazie a ciò anche le altre variabili latenti esercitano un effetto indiretto non trascurabile sulla FEDELTA' stessa, particolarmente nel caso della QUALITA' percepita.

Uno dei vantaggi maggiori della stima PLS-PM è che essa fornisce anche i punteggi delle variabili. In tal modo è possibile valutare il livello di giudizio degli intervistati sui costrutti latenti d'interesse. Come si vede in tabella 4.15, a fronte di una scarsa variabilità, il punteggio medio è generalmente "buono" (ovvero 3), con il VALORE a 2.744, definibile in termini di "quasi buono", come pure la FEDELTA' (2.889) e

	ASPET.	QUAL.	VAL.	SODD.	FED
ASPETTATIVE					
QUALITA'	0.643				
VALORE	0.549	0.692			
SODDISFAZIONE	0.600	0.768	0.371		
FEDELTA'	0.492	0.629	0.304	0.819	

Tabella 4.14: Effetti totali.

la SODDISFAZIONE (2.761), mentre con le ASPETTATIVE (3.014) a livelli "più che buoni" come la QUALITA' (3.094).

Variabile	Minimo	Massimo	Media	Deviazione std.
ASPETTATIVE	1	4	3.014	0.709
QUALITA'	1	4	3.094	0.848
VALORE	1	4	2.744	0.955
SODDISFAZIONE	1	4	2.761	0.864
FEDELTA'	1	4	2.889	1.102

Tabella 4.15: Descrittive dei punteggi delle variabili latenti.

Essendo a disposizione gli identificativi per gruppi di settore, è possibile valutare la media dei punteggi anche in base al settore industriale di appartenenza. Il settore aeronautico è caratterizzato dalle minori ASPETTATIVE e dalla maggiore FEDELTA', il settore con le peggiori performance è quello della TV via cavo e satellitare, quello che invece ha raccolto le valutazioni generalmente migliori è il settore alberghiero.

Con i dati a disposizione è possibile analizzare le valutazioni espresse da diversi sottogruppi, utilizzando per esempio le indicazioni sull'azienda erogatrice del prodotto/servizio, sesso, età livello culturale dell'intervistato ecc. In generale è possibile definire dei sottogruppi omogenei al loro interno rispetto a caratteristiche non rilevate nel questionario, utilizzando l'integrazione REBUS dell'algoritmo PLS-PM, alla ricerca di eventuale variabilità latente.

La variabilità latente

Dopo aver effettuato una stima PLS-PM globale, è sempre opportuno effettuare un'analisi REBUS, alla ricerca di variabilità latente. Sebbene gli indicatori analizzati non abbiano evidenziato particolari criticità nel modello, vale la pena verificare se non sia possibile migliorarne la capacità predittiva e/o descrittiva cercando nei

VARIABILE	Settore	No. di oss.	Media	Dev. St.
ASPETTATIVE	Aviazione civile	1000	*2.800*	0.696
	TV cavo/satellite	1000	2.947	0.736
	Hotel	1000	**3.201**	0.611
	Forniture energetiche	1000	3.106	0.718
QUALITA'	Aviazione civile	1000	2.992	0.829
	TV cavo/satellite	1000	*2.839*	0.903
	Hotel	1000	**3.328**	0.737
	Forniture energetiche	1000	3.215	0.830
VALORE	Aviazione civile	1000	2.945	0.887
	TV cavo/satellite	1000	2.374	0.995
	Hotel	1000	**3.053**	0.831
	Forniture energetiche	1000	*2.605*	0.942
SODDISFAZIONE	Aviazione civile	1000	2.760	0.789
	TV cavo/satellite	1000	*2.468*	0.945
	Hotel	1000	**2.976**	0.762
	Forniture energetiche	1000	2.841	0.866
FEDELTA'	Aviazione civile	1000	**3.093**	1.053
	TV cavo/satellite	1000	*2.624*	1.183
	Hotel	1000	2.841	1.031
	Forniture energetiche	1000	2.997	1.077

Tabella 4.16: Punteggi medi per settore.

dati dei sottogruppi più omogenei. Tale operazione potrebbe essere effettuata con un'analisi di cluster, ma il metodo di stima PLS-PM integra un algoritmo più efficiente che effettua una procedura di split ottimizzando i risultati della stima: il metodo REBUS presentato nel paragrafo 2.5.4.

Come facevano pensare i buoni risultati dell'analisi globale, non sembra che ci siano dei sottogruppi rilevanti all'interno dei dati. Con la procedura automatica di troncamento in 3 gruppi, i modelli separati ottenuti non mostrano miglioramenti sostanziali.

Il Gof globale è più alto dei tre Gof di gruppo, e l'R^2 mostra un leggero miglioramento solo per il gruppo 3, peggiorando però negli altri casi. In base al principio di economicità della descrizione si ritiene opportuno mantenere il modello globale.

4.3 Il modello strutturale: l'apprendimento tramite reti probabilistiche bayesiane

Nell'introduzione del presente capitolo è stato già osservato che le reti bayesiane stanno assumendo un ruolo sempre maggiore nell'ambito della ricerca economica

Variabile latente	**R^2-Glob**	**R^2-M1**	**R^2-M2**	**R^2-M3**
ASPETTATIVE				
QUALITA'	**0.288**	0.167	0.239	0.380
VALORE	**0.435**	0.157	0.169	0.173
SODDISFAZIONE	**0.742**	0.611	0.571	0.560
FEDELTA'	**0.412**	0.218	0.115	0.205
Media	**0.470**	0.289	0.274	0.330

	GoF-Glob	GoF-1	GoF-2	GoF-3
Assoluto	**0.592**	0.446	0.426	0.474
Relativo	**0.979**	0.976	0.964	0.973
Mod. esterno	**0.999**	0.998	0.997	0.999
Mod. interno	**0.980**	0.978	0.967	0.975

Tabella 4.17: Indici di bontà di adattamento per i modelli ottenuti con l'algoritmo REBUS.

e nelle strategie di mercato, soprattutto nei settori dove sono richiesti strumenti agili e potenti per il supporto alle decisioni.

In questo paragrafo vengono mostrate le potenzialità dell'apprendimento bayesiano in relazione alle teorie di marketing della soddisfazione e della fedeltà della clientela, ovvero in un ambito da sempre "dominio" dei modelli ad equazioni strutturali.

4.3.1 L'apprendimento preliminare ed il trattamento delle variabili latenti

Nel capitolo 3 è stato messo in evidenza che le reti bayesiane hanno la capacità di estrarre informazione dai dati grezzi senza richiedere pesanti ipotesi preliminari.

In questa applicazione, l'unica operazione preliminare effettuata è stata quella di raggruppare le modalità delle variabili di giudizio, allo scopo di rendere i risultati più leggibili.

Il primo approccio all'apprendimento della struttura è indirizzato al raggiungimento di quest'ultima condizione.

Nel presente lavoro è stato stabilito, per opportunità, un nodo obiettivo intorno al quale far ruotare l'analisi. Infatti, al contrario dei modelli strutturali, per le reti bayesiane la sintesi dei risultati è molto più difficile, nel senso che le reti sono caratterizzate da un maggior numero di parametri, quelli caratterizzanti la distribuzione congiunta, rispetto alle stime PLS-PM, che hanno la possibilità

di sintetizzare i risultati relativi alla forza delle relazioni tramite dei sistemi di equazioni.

Difficilmente la distribuzione congiunta presenta una forma nota e soprattutto la comprensione dell'eventuale formula che la riassume e le sue implicazioni non sono di immediata interpretazione quanto i parametri di una o più rette di regressione.

L'obiettivo dell'analisi è stato fissato sulla variabile manifesta di riacquisto, non solo perché essa si prospetta come un nodo *foglia* piuttosto che *radice*, quindi dipendente, direttamente o meno, un po' da tutte le altre variabili, ma anche perché il fattore che si trova a rappresentare, la fedeltà del cliente, è spesso l'obiettivo implicito di tutte le indagini di mercato.

Un altro punto fondamentale da considerare è la presenza di variabili latenti. Tra le poche condizioni necessarie per l'applicazione del modello di reti probabilistiche bayesiane rientra la condizione di sufficienza causale, ovvero la capacità del grafo di ricomprendere tutte le variabili aventi un'influenza significativa nella struttura causale. Le variabili latenti rientrano proprio in questa definizione quindi dovrebbero essere ricomprese nel grafo finale.

Nella teoria delle reti bayesiane non è stato ancora sviluppato, e soprattutto implementato un metodo autonomo per ricavare le variabili latenti, quindi la tecnica più diffusa per l'estrazione dei fattori è il clustering non supervisionato per variabili.

Ad esempio il software qui utilizzato per l'apprendimento della struttura delle reti si basa sulla divergenza di Kullback-Leiber calcolata sull'intensità degli archi congiungenti le variabili per individuare i nodi meno distanti e creare dei raggruppamenti di variabili tra loro simili per poi indurne la variabile latente rappresentativa del blocco (Bayesia, 2010).

Allo scopo di disporre degli archi da utilizzare nell'apprendimento delle variabili latenti, viene inizialmente impiegato l'algoritmo degli alberi di massima copertura(MWST)[10], introdotto nel paragrafo 3.5.4, con funzione di punteggio data dalla MDL[11] perché, fornendo una struttura ad albero, questo metodo garantisce un apprendimento rapido e una maggiore leggibilità del grafo ottenuto.

In questo primo passo dell'analisi la variabile obiettivo viene esclusa in ragione del fatto che si intende tenerla separata dal raggruppamento. Comunque, effettuando l'operazione di clustering tenendo in considerazione anche la variabile di riacquisto, con soli 4 gruppi, tale variabile rientra nel blocco della SODDISFAZIONE; con 5 gruppi, invece, il riacquisto è la prima ad uscire dal blocco, facendo gruppo a sè.

[10]Maximum Weight Spanning Tree.
[11]Lunghezza di descrizione minimale.

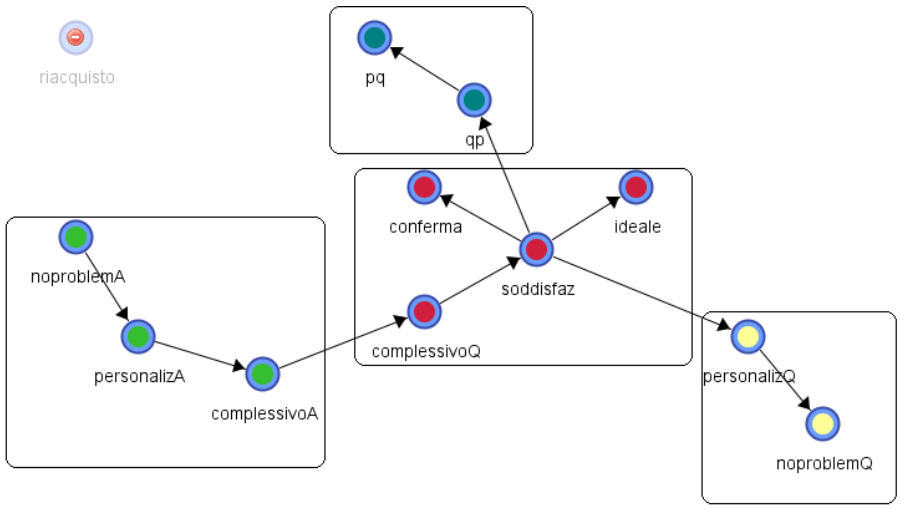

Figura 4.8: Struttura grafica ottenuta con l'algoritmo MWST e su cui è stato effettuato il clustering.

Una volta determinate le variabili maggiormente simili con le tecniche di raggruppamento (vedi figura 4.8) in base alle caratteristiche medie dei gruppi viene generato un nuovo fattore latente come esemplifica la figura 4.9.

In base alle variabili manifeste che compongono il blocco e alle relazioni probabilistiche che esse hanno con il relativo fattore è possibile caratterizzare il fattore stesso.

I cluster ottenuti in questa sede non sono molto diversi dai modelli di misura impiegati nel PLS-PM, come si poteva prevedere essendo il formulario dell'indagine costruito appositamente per creare tale struttura.

A questo punto inizia ad intravedersi l'utilità di un apprendimento induttivo dai dati in quanto le relazioni probabilistiche evidenziate dalla rete bayesiana sembrano mostrare che la variabile manifesta complessivoQ percepita sia più prossima al fattore di soddisfazione che a quello di qualità come sarebbe giusto presupporre. In questa considerazione rientra il fatto che la rete, rilevando relazioni non necessariamente lineari potrebbe aver colto aspetti sfuggiti al ricercatore nell'impostare la struttura PLS-PM.

Per la precisione si ritiene necessario specificare che il cambiamento di blocco della variabile manifesta complessivoQ percepita si verifica anche quando la variabile obiettivo riacquisto viene inserita nell'apprendimento.

A scopo esemplificativo in figura 4.9 viene riportato il grafo di misura e la relativa distribuzione congiunta del fattore SODDISFAZIONE. Infatti, per tutte le altre variabili latenti la caratterizzazione è la medesima del PLS-PM, tranne ovviamente per la variabile QUALITA' che ha perso una variabile manifesta. Chiaramente,

la visualizzazione completa della distribuzione richiede una certa interattività con il software, difficile da riprodurre su carta. Per tale motivo è stata impostata una variabile obiettivo, o meglio, una modalità obiettivo della variabile interessata (indicata con un nodo concentrico in figura 4.9).

La variabile latente risultante dal grafo in figura 4.9 può essere ancora caratterizzata come SODDISFAZIONE del cliente in quanto la qualità complessiva percepita è molto vicina sintatticamente a tale concetto, come dimostra anche il fatto che nelle stime PLS-PM la variabile latente QUALITA' che essa caratterizza è la più influente sulla SODDISFAZIONE.

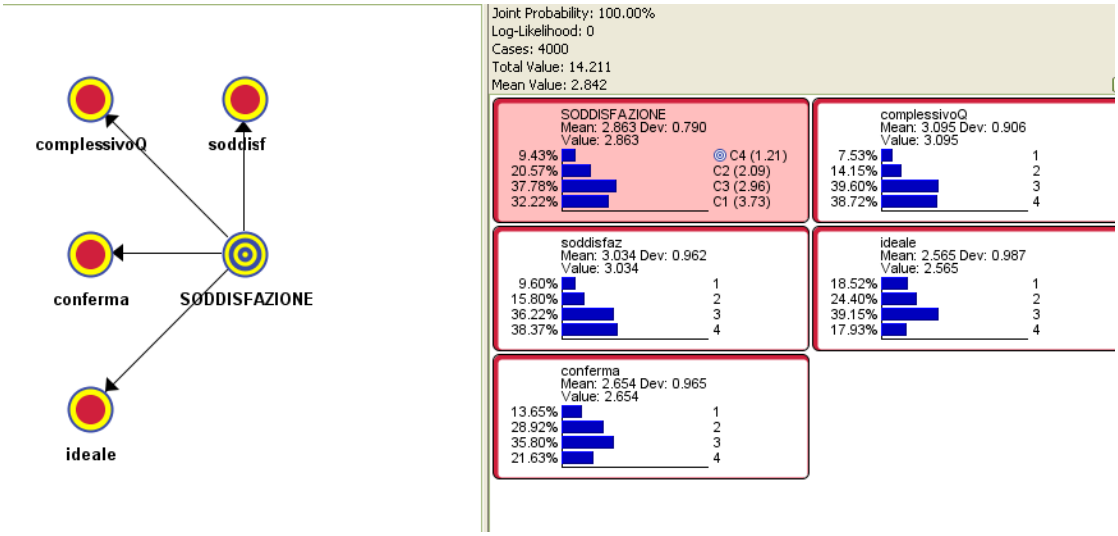

Figura 4.9: Struttura grafica ottenuta con l'algoritmo Tabu Order applicato sul nuovo gruppo di variabili che include anche i fattori latenti.

4.3.2 La rete probabilistica bayesiana finale

Una volta determinati anche i fattori, la rete raggiunge finalmente la condizione di sufficienza causale (vedi paragrafo 3.5.2) ed è possibile procedere all'apprendimento della struttura grafica finale, reintegrando anche la variabile obiettivo nell'analisi.

Allo scopo di utilizzare un algoritmo di ricerca più efficiente viene impiegato l'algoritmo Tabu Order, presentato nel paragrafo 3.5.4, con lunghezza della lista tabu pari a dieci. Il risultato ottenuto è mostrato nella figura 4.10.

Il grafo ottenuto con questo primo passaggio non è ancora molto esplicativo quindi si sceglie di trascurare i legami meno forti modificando il coefficiente di complessità strutturale, che altro non è se non la componente del punteggio MLD che pesa la complessità del modello.

Nelle impostazioni dell'algoritmo esso varia tra 0 e 150, più alto si impone il coefficiente e maggiore è la forza che deve avere un legame perchè appaia nel modello. Nel caso di coefficiente posto uguale a 0 si favorisce il grafo completamente connesso, al contrario per un coefficiente pari a 150 si tende ad ottenere una struttura molto semplificata. Nel caso in questione, esso è stato posto pari a 5 per ottenere una migliore leggibilità delle relazioni.

Già in questa fase è possibile integrare l'apporto del modello PLS-PM nell'apprendimento delle reti bayesiane riprendendo le considerazioni fatte nel paragrafo 4.2 sui vari modelli di misura che evidenziavano un chiaro schema formativo. In questo caso è possibile proibire archi che vadano dalle variabili manifeste alle variabili latenti.

Inoltre per completare il quadro è utile stabilire anche dei vincoli sulle relazioni tra variabili manifeste in modo da avere alla fine uno schema opportunamente caratterizzato esclusivamente da relazioni latente→manifesta e latente →latente.

Il risultato finale, mostrato in figura 4.11 evidenzia quindi lo spostamento della variabile manifesta qualità percepita dal blocco della QUALITA' a quello della SODDISFAZIONE, arricchendo quest'ultimo, ma privando la QUALITA' di una sua caratterizzazione fondamentale.

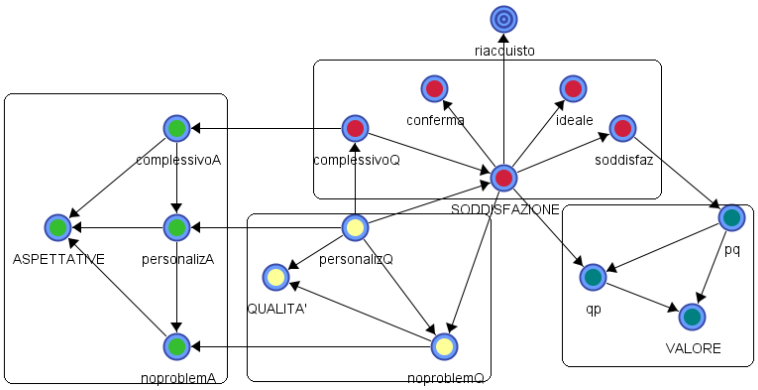

Figura 4.10: Struttura grafica ottenuta con l'algoritmo Tabu Order applicato sul nuovo gruppo di variabili che include anche i fattori latenti.

L'ultima considerazione da effettuare riguarda la direzione delle frecce tra le variabili latenti. È stato anticipato nel capitolo 3 che gli algoritmi di apprendimento nella maggior parte dei casi determinano lo scheletro e le strutture convergenti del grafo più adatto secondo l'opportuno criterio di valutazione.

Ma a ben vedere è possibile invertire il verso delle frecce a proprio piacimento, nel rispetto delle strutture convergenti, oppure anche creandone di nuove se esistono conoscenze a priori che ne impongono la presenza nel grafo.

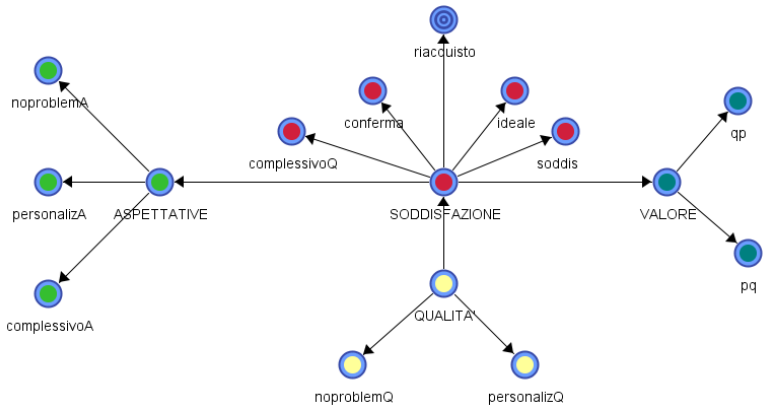

Figura 4.11: Grafo ottenuto dall'integrazione delle conoscenze a priori sui modelli di misura.

Nel caso specifico, il grafo ottenuto originariamente e mostrato in figura 4.11 con l'algoritmo Tabu order prevede due frecce dirette da SODDISFAZIONE verso ASPETTATIVE e VALORE. Quest'orientamento entra in contraddizione con le conoscenze a priori che abbiamo sulla struttura delle variabili.

In effetti le variabili relative alle aspettative sono state costruite in modo da rilevare concetti che vengono temporalmente prima di quelli rilevati dalle variabili di soddisfazione.

Inoltre anche il concetto di VALORE non sembra essere logicamente determinabile dalla SODDISFAZIONE, specie in considerazione delle domande a cui è associata la variabile. Infatti per entrambe le variabili manifeste pq e qp viene chiesta espressamente una valutazione di rapporto qualità/prezzo o viceversa, che a rigor di logica è una componente della soddisfazione piuttosto che il contrario.

Anche in questo caso benché non accettabile logicamente, la proposta nata dalla struttura inferita dalle reti bayesiane ha contribuito a dare una nuovo spunto di riflessione anche per la struttura causale del modello.

A questo punto, risulta necessario imporre una nuova struttura di convergenze nel grafico, che nella proposta iniziale non era prevista, il cambio della struttura porta necessariamente ad un cambio della classe di equivalenze della soluzione, ma l'operazione diventa obbligata poiché deriva da conoscenze a priori non trascurabili.

Il grafo che si ottiene in ultima analisi è quello di figura 4.12, in cui lo spessore delle frecce è proporzionale alla mutua informazione dell'arco (indicata sullo stesso, insieme al suo valore percentuale).

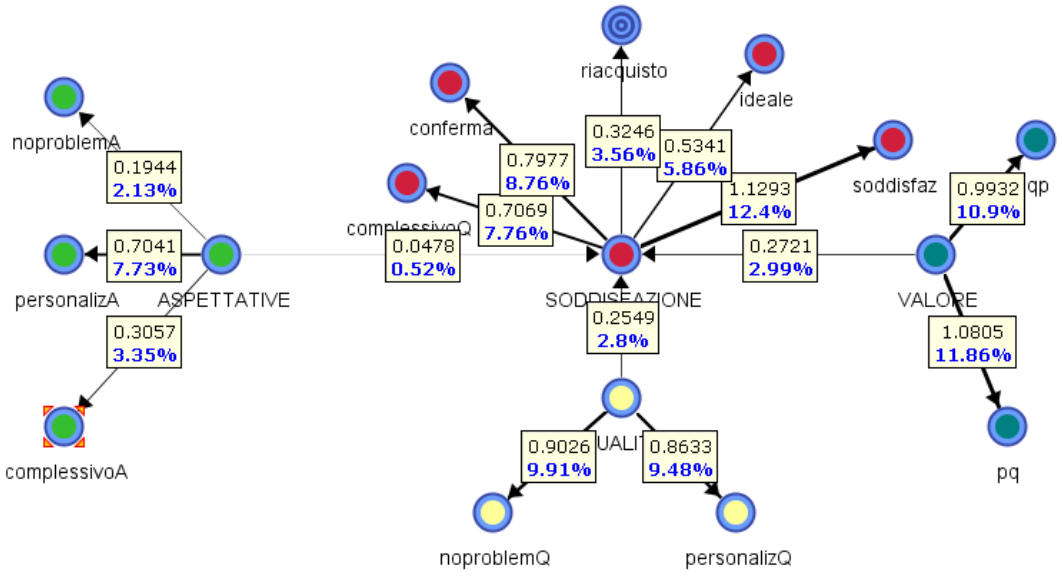

Figura 4.12: Struttura grafica ottenuta con l'algoritmo Tabu Order integrato con alcune conoscenze a priori sulle relazioni e sulle direzioni delle stesse.

4.3.3 Gli strumenti per la valutazione del modello

Gli strumenti per la valutazione del modello ottenuto non sono altrettanto sviluppati come per i metodi SEM. È Possibile far ricorso a tecniche di validazione incrociata, ad esempio introducendo il metodo jackknife per valutare la stabilità degli archi per il risultato di apprendimento. Il metodo consiste nel dividere il database non ordinato in un numero k di sottocampioni di cui $k-1$ vengono usati per l'apprendimento.

Effettuando la validazione incrociata per la struttura con cambio di orientazione, è stata inserita una struttura convergente ulteriore che non esiste nella proposta originaria dell'algoritmo, l'apprendimento sui sotto campioni non ritrova mai questa convergenza, che però è obbligata per questioni teoriche.

Ovviamente il punteggio ottenuto dalla classe di equivalenza appresa con l'algoritmo Tabu è migliore del modello modificato, ma se le conoscenze a priori impongono tale scelta dell'orientamento, bisogna tenerne conto.

Una volta validata la struttura, è possibile utilizzare la distribuzione condizionata che il software apprende automaticamente tramite inferenza bayesiana (vedi paragrafo 3.5.1), come leva per l'analisi del riacquisto.

Poiché esistono difficoltà a mostrare il meccanismo alquanto dinamico dell'utilizzo delle probabilità congiunte, viene proposto un esempio in cui si mostra la variazione (freccia grigia) della distribuzione congiunta, una volta fissata una modalità della variabile SODDISFAZIONE (figura 4.13).

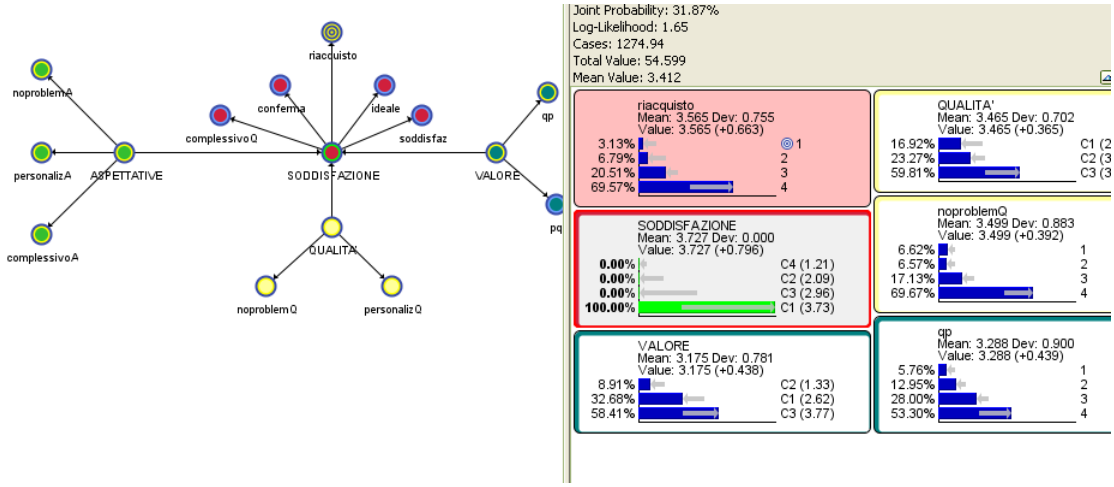

Figura 4.13: Struttura grafica e modalità di visualizzazione delle probabilità condizionate.

Dall'indagine della distribuzione è possibile anche valutare l'effetto del Markov blanket: in figura 4.13 viene mostrata la distribuzione condizionata del grafo imponendo una modalità della variabile SODDISFAZIONE come evidenza, mentre in figura 4.14 si evidenzia l'ininfluenza sulla variabile di riacquisto del fissare una modalità delle variabili d-separate, ad esempio la variabile VALORE, una volta che è stata fissata la modalità dell'unico genitore: la SODDISFAZIONE (si noti che non è presente alcuna freccia grigia sulla distribuzione del riacquisto).

4.4 L'integrazione degli approcci

Poiché l'autrice ritiene opportuno sfruttare le potenzialità di entrambi gli approcci presentati, allo scopo di migliorare i risultati dell'analisi statistica di relazioni causali, in questo paragrafo vengono presentate alcune proposte di integrazione dei metodi SEM e delle reti bayesiane.

4.4.1 Stima PLS-PM con strutture di derivazione bayesiana

La prima proposta consiste nell'utilizzare i risultati di apprendimento delle reti bayesiane per determinare il modello strutturale da stimare con il PLS-PM.

Come suggerito dall'apprendimento bayesiano, si possono valutare i miglioramenti ottenibili spostando la variabile *complessivoQ* dal modello di misura della QUALITA' a quello della SODDISFAZIONE. In linea di principio tale operazione

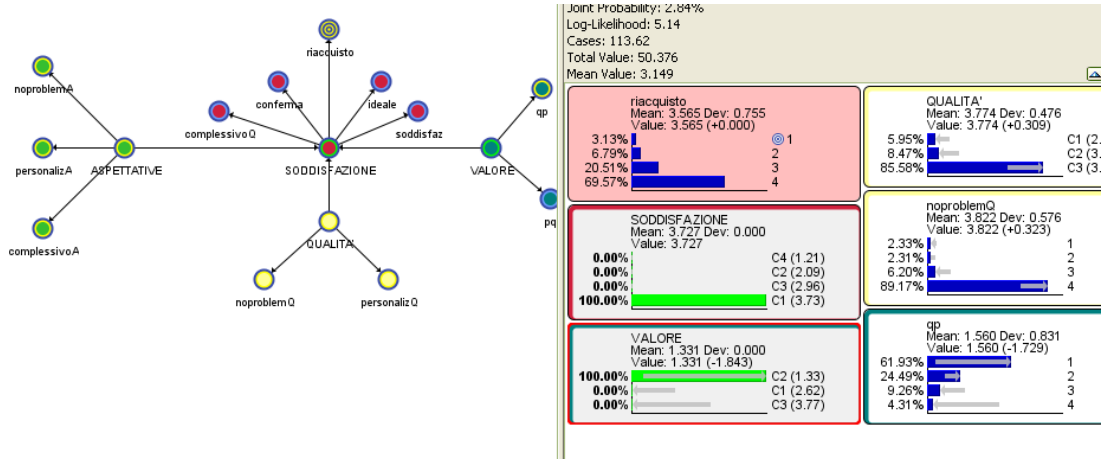

Figura 4.14: Struttura grafica e modalità di visualizzazione delle probabilità condizionate.

non pone grossi problemi in quanto la variabile manifesta misura la qualità complessivamente percepita, quindi può logicamente essere interpretata come misura della SODDISFAZIONE. D'altro canto la QUALITA' resta orfana di un indicatore importante.

La matrice di correlazione non dà indicazioni univoche al riguardo poiché le correlazioni con gli indicatori di SODDISFAZIONE non sono molto diverse da quelle degli indicatori di QUALITA' (tabella 4.18)

	QUALITA'		SODDISFAZIONE		
Variabili	personalizzQ	noproblemQ	soddis	conferma	ideale
complessivoQ	0.715	0.559	0.739	0.646	0.612

Tabella 4.18: Correlazioni di complessivoQ

Se consideriamo gli indicatori di omogeneità e unidimensionalità (tabella 4.19), i nuovi modelli di misura possono essere accettati senza problemi.

Variabile latente	Dimensioni	α di Cronbach	ρ di D.G. (ACP)	Autovalori
QUALITA'	2	0.735	0.884	1.656
				0.436
SODDISFAZIONE	4	0.892	0.926	2.767
				0.375
				0.309
				0.202

Tabella 4.19: Validità del modello esterno *Pls-Bn1*

A questo punto bisogna valutare se sia opportuno il salto concettuale di spostare *complessivoQ* dal blocco in cui è stata inserita per costruzione.

Confrontando il modello originario con il nuovo modello che viene indicato come *Pls-Bn1*, i valori che si ottengono non indicano un netto miglioramento. La comunalità di *complessivoQ* passa da 0.758 del modello originario a 0.727 nel modello *Pls-Bn1*. L'indice di GoF subisce un lieve peggioramento (passando da a 0.592 a 0.579) e per il resto, come da tabella 4.20, gli scostamenti sono minimi e comunque non sufficienti per giustificare uno sconvolgimento nella struttura che ha una sua consistenza logica e una stabilità pluriennale.

	modello con complessivoQ in SODD.			modello originario		
Variabile latente	R^2	Com. Medie (AVE)	Ridond. medie	R^2	Com. Medie (AVE)	Ridond. Medie
ASPETTATIVE		0.609			0.610	
QUALITA'	0.248	0.790	0.196	0.288	0.746	0.215
VALORE	0.393	0.886	0.348	0.435	0.886	0.386
SODDISFAZIONE	0.745	0.757	0.564	0.742	0.791	0.587
FEDELTA'	0.412	1.000	0.412	0.412	1.000	0.412
Media	0.449	0.746	0.380	0.470	0.746	0.400

Tabella 4.20: Validità del modello strutturale *Pls-Bn1*

Un'altra possibilità di integrazione è quella di utilizzare il grafo ottenuto al paragrafo 4.3 e mostrato in figura 4.12 per costruire il modello causale. Il risultato di una stima PLS-PM basata su tale modello strutturale è mostrato in figura 4.15. Si nota come la variabile manifesta *complessivoQ* sia stata attribuita alla misura della SODDISFAZIONE e come siano stati eliminati tutti i legami tra le variabili latenti esplicative della stessa. A questo punto è possibile valutare se vi sono miglioramenti significativi nella performance del modello di *derivazione bayesiana*[12].

I risultati ottenuti, in termini di miglioramento di potere predittivo e potere esplicativo del modello non sono eclatanti. Confrontando la tabella 4.21 con i risultati delle tabelle 4.9 e 4.10 del paragrafo 4.2.2 non emergono sostanziali cambiamenti. Le comunalità medie della SODDISFAZIONE sono lievemente aumentate mentre quelle del VALORE sono diminuite.

Gli R^2 delle uniche due variabili esogene del modello *bayesiano* SODDISFAZIONE (0.746) e FEDELTA' (0.412) sono rimasti pressoché immutati rispetto al modello originario in cui sono pari rispettivamente a 0.742 e a 0.412, e l'indice di GoF (assoluto=0.658 e relativo=0.987) ha subito solo un leggero aumento: nel modello originario era pari 0.592 in valore assoluto e a 0.979 in valore relativo.

[12]Da qui in avanti denominato per semplicità modello *bayesiano*.

Variabile latente	Variabile manifesta	Comunalità	Comunalità medie	Ridondanze	Ridondanze medie
ASPETTATIVE	complessivoA	0.613	0.612		
	personalizzA	0.694			
	noproblemA	0.528			
QUALITA'	personalizzQ	0.820	0.757		0.208
	noproblemQ	0.760			
VALORE	pq	0.885	0.790		0.481
	qp	0.886			
SODDISFAZIONE	soddis	0.833	0.886	0.621	0.513
	conferma	0.763		0.569	
	ideale	0.706		0.527	
	complessivoQ	0.727		0.543	
FEDELTA'	riacquisto	1.000	1.000	1.102	0.412
			0.747		0.404

Tabella 4.21: Validità del modello *bayesiano*.

	ASPET.	QUAL.	VAL.	SODD.	FED.
ASPETTATIVE					
QUALITA'	0.000				
VALORE	0.000	0.000			
SODDISFAZIONE	0.145	0.417	0.375		
FEDELTA'	0.123	0.354	0.319	0.850	

Tabella 4.22: Effetti totali modello di derivazione bayesiana

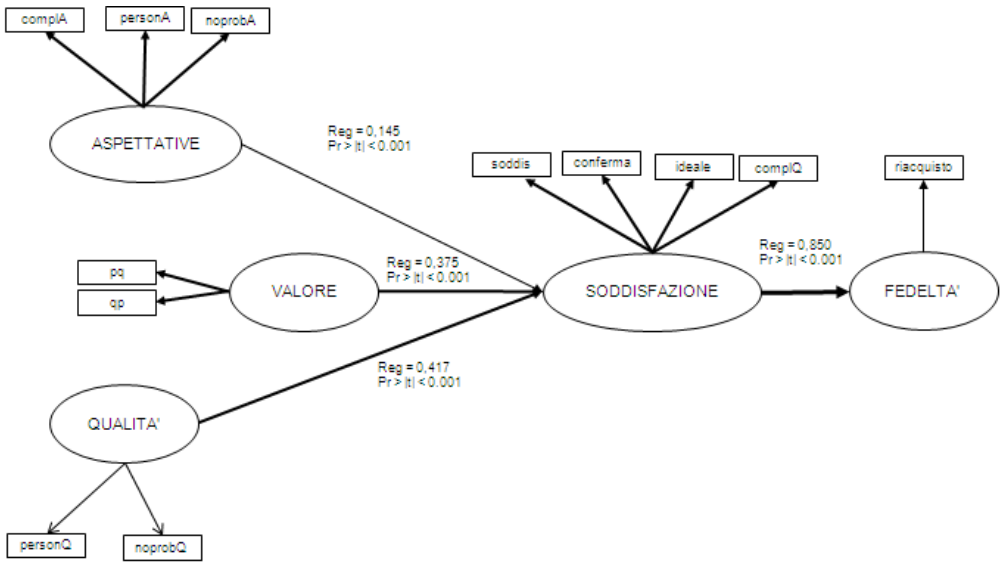

Figura 4.15: Stima PLS-PM dei parametri del modello SEM impostato su struttura ricavata dall'apprendimento tramite reti probabilistiche bayesiane.

In definitiva, a fronte di un miglioramento così modesto, è preferibile mantenere la struttura classica del modello, anche in ragione di un'altra considerazione. Il modello *bayesiano* esclude dall'analisi l'influenza indiretta delle variazioni delle ASPETTATIVE e della QUALITA', il cui effetto totale, con questo modello causale, viene notevolmente ridimensionato. Per esempio, l'azione delle ASPETTATIVE sulla SODDISFAZIONE passa da 0.600 a 0.145 e per quanto riguarda la QUALITA', si passa da 0.768 a 0.417. Nella scelta influisce anche l'opportunità di utilizzare gli effetti indiretti fra variabili, che dall'analisi PLS-PM risultano significativi nonché consistenti.

4.4.2 Apprendimento di reti probabilistiche bayesiane sui punteggi derivanti dalla stima PLS-PM

Un'altra interessante opportunità che può derivare dall'integrazione dei due metodi è la ricerca di una struttura causale alternativa a quella ACSI in base all'applicazione degli algoritmi di apprendimento sui punteggi ottenuti con una stima PLS-PM.

In questo caso non c'è la necessità di effettuare alcun raggruppamento e quindi si applica direttamente una algoritmo di apprendimento Tabu Order sui punteggi ottenuti dalla stima PLS-PM (opportunamente raggruppati per ottenere solo quattro modalità, le stesse delle variabili manifeste di giudizio) col fine di ottenere il grafo delle relazioni mostrato in figura 4.16.

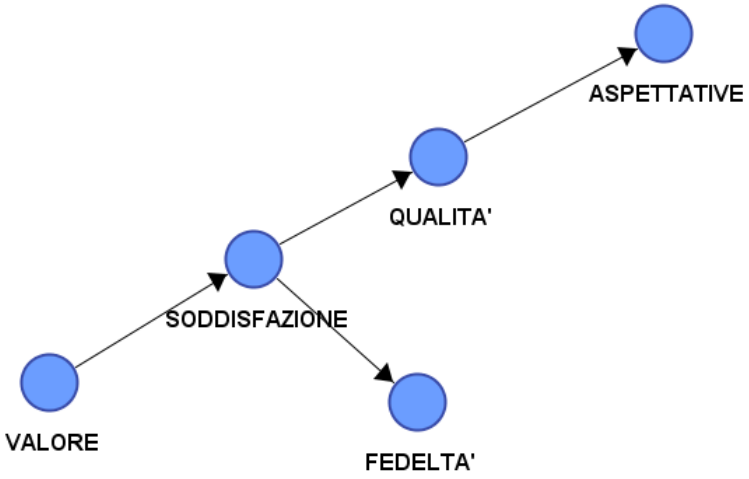

Figura 4.16: Struttura di rete bayesiana appresa sui punteggi delle variabili latenti PLS-PM.

Il risultato mostra ancora la centralità della SODISFAZIONE, ma presenta sempre il problema si trascurare i legami indiretti tra ASPETTATIVE, QUALITA' e VALORE.

In ultimo bisogna considerare che ancora una volta il verso delle frecce non corrisponde ai vincoli logici imposti sulle variabili manifeste e di conseguenza sui concetti latenti da esse determinati.

Cambiare il verso dell'arco da ASPETTATIVE a QUALITA' non crea problemi mentre solo uno dei nodi tra QUALITA' e VALORE può avere un arco diretto verso SODDISFAZIONE e l'altro dovrebbe essere contrario per non creare una struttura convergente.

Poiché tale opzione è in contraddizione con la teoria, come già osservato nel paragrafo 4.3, anche in questo caso è necessario cambiare la classe di equivalenza di Markov introducendo una struttura convergente imposta dalle conoscenze teoriche a priori.

Sull'apprendimento dei punteggi PLS-PM, le reti bayesiane non riescono a proporre nuove idee di relazione, ma questo risultato era prevedibile considerando che la struttura del questionario determina necessariamente la struttura dei dati, i quali rispecchiano lo schema causale ACSI in base al quale sono stati rilevati.

Capitolo 5

Conclusioni

Come si è potuto notare i modelli ad equazioni strutturali si adattano piuttosto bene ai dati analizzati, ma tale risultato era abbastanza prevedibile dato che la struttura del questionario deriva proprio dall'intenzione di implementare gli algoritmi PLS-PM nelle analisi di ricerca di mercato e di soddisfazione della clientela.

In effetti, in questa applicazione stupisce piuttosto la necessità di dover escludere delle variabili dall'analisi, come è accaduto per le due variabili di tolleranza.

Per quanto riguarda l'applicazione delle reti probabilistiche bayesiane al caso, bisogna ammettere che, seppur *adattate* allo scopo di analizzare dei dati continui con opportuni accorgimenti quali il raggruppamento delle modalità delle variabili analizzate e la conseguente perdita di informazioni, esse hanno dato qualche spunto di novità.

Dall'analisi è emersa l'opportunità di aggregare la variabile di percezione della qualità (che pure nell'analisi PLS-PM mostrava una forte connotazione monofattoriale) al modello di misura del fattore di SODDISFAZIONE piuttosto che a quello della QUALITA' per cui è stata costruita.

Poiché il modello globale derivante dall'apprendimento con le reti bayesiane non presenta un miglioramento rispetto a quello base dell'ACSI nei termini delle valutazioni effettuate con i modelli strutturali, nasce il sospetto che l'analisi bayesiana abbia rilevato delle relazioni non lineari che i SEM non sono ancora adatti a percepire.

Tale proposta di modifica dei modelli di misura, può essere intesa anche nel senso di un suggerimento per una migliore caratterizzazione della parte di questionario dedicata alla rilevazione della qualità globale, in modo da distinguere più nettamente i due concetti latenti di qualità e di soddisfazione.

Per quanto riguarda la struttura causale suggerita dall'analisi delle reti bayesiane, essa non ha trovato un riscontro positivo a causa dell'ambiguità relativa alla direzione della causalità fra variabili latenti: la prima opzione proposta dall'algoritmo è alquanto improbabile in quanto stabilisce la direzione di causalità dalla

SODDISFAZIONE al VALORE ed alle ASPETTATIVE, in netta contraddizione con il fatto che le aspettative della clientela vengono rilevate come una variabile temporalmente precedente alla soddisfazione della stessa.

È più intuitivo implementare la direzione della causalità validata con il modello PLS-PM nell'algoritmo di apprendimento bayesiano, sotto forma di vincoli sulle relazioni. Oltre al vincolo piuttosto elementare di impedire relazioni tra variabili manifeste, una volta inserite le variabili latenti nello schema causale, si presenta la necessità di stabilire la forma dello schema di misura.

Seguendo le indicazioni dell'analisi PLS-PM si è optato per il modello riflessivo (che come si diceva è anche quello su cui è stata costruita la rilevazione). Infatti il modello bayesiano non è riuscito da solo a rilevare tali strutture.

Infine, anche l'analisi bayesiana sui punteggi delle variabili latenti ottenuti con i PLS-PM non ha dato grandi spunti di novità, ribadendo la posizione centrale della SODDISFAZIONE nello schema causale, ma lasciando ancora indeterminato il problema della direzione delle frecce.

Su tutto, inoltre, risalta la limitatezza degli strumenti di valutazione della bontà del modello di reti bayesiane rispetto agli strumenti offerti dai modelli strutturali. Ovviamente l'orientamento di questi ultimi, più nei metodi di stima LISREL che PLS-PM, è proprio la ricerca di strutture causali stabili e, in un ottica principalmente deduttiva, la validazione della teoria attraverso il confronto con i dati.

Diametralmente opposto si posiziona l'obiettivo dell'analisi bayesiana, orientata principalmente alla ricerca di informazione all'interno dei dati, in un'ottica puramente induttiva che si propone di produrre automaticamente teorie originali e scoprire legami tralasciati delle teorie più accreditate.

Ovviamente è più facile che portino buoni risultati metodologie impiegate nell'analisi di strutture di relazioni causali già consolidate, anche se su dati sempre differenti, piuttosto che algoritmi operanti principalmente "alla cieca", senza direzioni privilegiate di ricerca, ma d'altro canto anche senza vincoli nell'indagine.

Per tali ragioni viene proposta un'integrazione delle due metodologie, proprio allo scopo di sfruttare le potenzialità di entrambe, nella fase della ricerca statistica in cui esse si rivelano più utili.

Il livello attuale dello sviluppo degli algoritmi di intelligenza artificiale non consente ancora di raggiungere il tanto agognato obiettivo di un apprendimento completamente induttivo, ma almeno riesce a riportare in ambito statistico la parte della ricerca causale che si occupa della formulazione delle teorie, da Popper relegata ad un ambito di studio prettamente psicologico.

Non si vede ancora come possa essere convertita automaticamente la mera associazione statistica in un legame causale ma, quantomeno, è stato possibile

introdurre le leggi probabilistiche all'interno del processo creativo, indirizzandolo verso risultati più coerenti con i dati offerti dalla realtà.

I difetti più grossi riscontrati negli algoritmi di apprendimento implementati nelle reti probabilistiche bayesiane, sono da ricondurre alla difficoltà nel trattamento dei dati continui (che nella maggior parte dei casi vengono raggruppati per permetterne l'apprendimento) e all'impossibilità di determinare univocamente il verso delle relazioni causali. Gli algoritmi sono in grado di stabilire le strutture di convergenza, ma all'infuori di esse, la direzione degli archi non è definitiva e può essere sostituita indifferentemente da tutte quelle ad essa equivalenti in senso markoviano.

D'altro canto nei modelli ad equazioni strutturali il trattamento della non linearità riveste ancora un ruolo secondario, benché le prospettive di sviluppo in tal senso siano interessanti (Esposito Vinzi e Russolillo, 2010; Russolillo, 2009).

È pur vero però che il trattamento delle variabili latenti nei modelli PLS-PM presenta dei grossi vantaggi rispetto alle mere tecniche di clustering di variabili implementate nelle reti bayesiane. Infatti i punteggi determinati con il metodo PLS-PM prendono in considerazione l'intera struttura causale, mentre il clustering si limita a valutare la forza dei nodi di variabili manifeste ed a caratterizzare la variabile latente in base alle caratteristiche del blocco di appartenenza anche se potrebbe essere interessante implementare misure probabilistiche non lineari nella determinazione dei modelli di misura.

Anche il metodo di inferenza è diverso: per valutare le relazioni di dipendenza della variabile target le reti bayesiane fanno ricorso alle evidenze, alle casistiche ed al condizionamento, i PLS invece si basano sulla significatività delle relazioni.

Inoltre gli algoritmi di apprendimento delle reti probabilistiche bayesiane basati su punteggio inoltre, ottimizzano una funzione che valuta la struttura nel suo complesso, mentre il metodo PLS-PM permette di valutare separatamente il modello di misura ed il modello strutturale oltre che quello globale.

In ultimo nei modelli PLS-PM viene implementata la ricerca di variabilità latente all'interno dell'algoritmo di stima (con il metodo REBUS per esempio) in modo tale che il modello risultante non prenda in considerazione esclusivamente dei vincoli di distanza fra gli individui o fra le variabili, ma anche la bontà di adattamento dei sottomodelli su popolazioni più omogenee.

Il caso presentato in questa tesi è piuttosto orientato verso i metodi PLS-PM e dunque sarebbe interessante effettuare delle ricerche di lungo periodo in cui applicare gli algoritmi di apprendimento di reti probabilistiche bayesiane su dati più *neutri* e testare gli schemi direzionali più credibili utilizzando i modelli ad equazioni strutturali con il metodo di stima PLS-PM.

Infine, un altro spunto interessante potrebbe essere quello di confrontare le prestazioni dei nuovi algoritmi di stima PLS-PM che si stanno sviluppando in

ambito non lineare con i risultati ottenuti in questo campo dalle reti probabilistiche bayesiane.

Bibliografia

Anderson E.; Fornell C. (2000). Foundations of the American customer satisfaction index. *Total Quality Management & Business Excellence*, **11**(7), 869–882.

Bacon F.; Jardine L.; Silverthorne M. (2000). *The new organon*. Cambridge Univ Pr.

Bayesia S. (2010). *BayesiaLab.4.6.8 User guide*. Bayesia S.A.

BayesiaLab.4.6.8 (2010).

Bollen K. A. (1989). *Structural equations with latent variables*. A Wiley-Interscience Publication.

Chickering D. (2002). Learning equivalence classes of Bayesian-network structures. *The Journal of Machine Learning Research*, **2**, 445–498.

Chin W. (1998). The Partial Least Squares Approach to Structural Equation Modeling. *Modern methods for business research*, p. 295.

Cifarelli D. (1998). *Introduzione al calcolo delle probabilità*. McGraw-Hill Libri Italia.

Cifarelli D.; Muliere P. (1989). *Statistica bayesiana*. Iuculano. Editore. Pavia.

Cooper G.; Herskovits E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, **9**(4), 309–347.

De Finetti B. (1931). Sul significato soggettivo della probabilita. *Fundamenta mathematicae*, **17**, 298–329.

De Finetti B. (1937). La prï¿½vision: ses lois logiques, ses sources subjectives. In *Annales de l'Institut Henri Poincarï¿½*, volume 7, pp. 1–68.

Devlin K. (2009). *La lettera di Pascal. Storia dell'equazione che ha fondato la teoria della probabilità*. Rizzoli.

Dor D.; Tarsi M. (1992). A simple algorithm to construct a consistent extension of a partially oriented graph. *Cognitive Systems Laboratory, UCLA, Computer Science Department*.

Dowe P. (2010). Physical causation.

Efron B. (1982). The jackknife, the bootstrap and other resampling plans. In *CBMS-NSF regional conference series in applied mathematics*, volume 38. Siam.

Efron B.; Tibshirani R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Efron B.; Tibshirani R.; Tibshirani R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Esposito Vinzi V. (2008). The contribution of pls regression to pls path modelling: formative measurement model and causality network in the structural model. In *Joint Statistical Meetings (JSM) 2008, American Statistical Association, Denver, Colorado, United States of America, August 7th 2008.*

Esposito Vinzi V. (2009). Pls path modeling and pls regression: a joint partial least squarescomponent-based approach to structural equation modeling. In *IFCS@GFKL - Classification as a Tool for Research (IFCS 2009)*. University of Technology, Dresden, Germany, March 14th 2009 (Plenary Invited Speaker).

Esposito Vinzi V.; Russolillo G. (2010). Partial least squares path modeling and regression. *Wiley interdisciplinary reviews: computational statistics. New York: Wiley*.

Esposito Vinzi V.; Trinchera L.; Squillacciotti S.; Tenenhaus M. (2008). REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling. *Applied Stochastic Models in Business and Industry*, **24**(5), 439–458.

Esposito Vinzi V.; Chin W.; Henseler J.; Wang H. (2010a). Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields.

Esposito Vinzi V.; Trinchera L.; Amato S. (2010b). PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement. *Handbook of Partial Least Squares*, pp. 47–82.

Fornell C. (1992). A national customer satisfaction barometer: The swedish experience. *The Journal of Marketing*, **56**(1), pp. 6–21.

Fornell C.; Larcker D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, **18**(1), 39–50.

Fornell C.; Johnson M. D.; Anderson E. W.; Cha J.; Bryant B. E. (1996). The american customer satisfaction index: Nature, purpose, and findings. *The Journal of Marketing*, **60**(4), pp. 7–18.

Geiger D.; Heckerman D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *The Annals of Statistics*, **25**(3), 1344–1369.

Glover F. (1990a). Tabu search: A tutorial. *Interfaces*, **20**(4), 74–94.

Glover F. (1990b). Tabu search, Part I. *ORSA journal on Computing*, **2**(1), 4–32.

Glover F. (1990c). Tabu search, Part II. *ORSA Journal on computing*, **2**(1), 4–32.

Gupta S.; Kim H. W. (2008). Linking structural equation modeling to bayesian networks: Decision support for customer retention in virtual communities. *European Journal of Operational Research*, **190**(3), 818 – 833.

Jensen F.; Nielsen T. (2007). *Bayesian networks and decision graphs*. Springer Verlag.

Jöre skog K. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, **36**(4), 409–426.

Jöreskog K. (1970). A general method for analysis of covariance structures. *Biometrika*, **57**(2), 239.

Jöreskog K.; Sörbom D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software.

Jöreskog K.; Sörbom D. (2001). *LISREL 8: New statistical features*. Scientific Software International.

Jöreskog K.; Wold H. (1982). The ML and PLS techniques for modeling with latent variables: historical and comparative aspects. *Systems under indirect observation: Causality, structure, prediction*, **1**, 263–270.

Jöreskog K.; Sörbom D.; Magidson J.; Cooley W. (1979). *Advances in factor analysis and structural equation models*. Abt books Cambridge, MA.

Kaplan D. (2008). *Structural equation modeling: Foundations and extensions*. Sage Publications, Inc.

Kolmogorov A. (1950). Foundations of the theory of probability.

Koski T.; Noble J. (2009). *Bayesian networks: an introduction*. Wiley.

Kotler P.; Armstrong G. (2010). *Principles of marketing*. Prentice Hall.

Kumuthini J.; Jouffe L.; Bessant C. (2007). A novel graph optimisation algorithm for the extraction of gene regulatory networks from temporal microarray data. In *Proceedings of the 1st international conference on Bioinformatics research and development*, BIRD'07, pp. 28–37, Berlin, Heidelberg. Springer-Verlag.

Landenna G.; Marasini D.; Ferrari P. (1997). *Probabilità e variabili casuali*. Il Mulino, Bologna.

Laplace P.; Laplace P.; Dale A. (1995). *Philosophical essay on probabilities*. Springer.

Lewis D. (2001). *Counterfactuals*. Wiley-Blackwell.

Lidstone G. (1920). Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, **8**(182-192), 80.

Little R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, **83**(404), 1198–1202.

Little R.; Rubin D. (2002). Statistical analysis with missing data. *Wiley series in probability and statistics*.

Lohmöller J. (1987). *LVPLS program manual, version 1.8, Technical report, Zentralarchiv für Empirische Sozialforschung*. Universität Zu Köln, Köln.

Lohmöller J. (1989). *Latent variable path modeling with partial least squares*. Physica-Verlag Heidelberg.

Lyttkens E.; Areskoug B.; Wold H. (1975). The convergence of NIPALS estimation procedures for six path models with one or two latent variables. *Rapport technique, University of Göteborg*, **23**.

Menzies P.; Price H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science*, **44**(2), 187–203.

Naïm P.; Wuillemin P.; Leray P.; Pourret O.; Becker A. (2007). Réseaux bayésiens, Edition Eyrolles.

Parasuraman A.; Zeithaml V. A.; Berry L. L. (1985). A conceptual model of service quality and its implications for future research. *The Journal of Marketing*, **49**(4), pp. 41–50.

Parasuraman A.; Zeithaml V. A.; Berry L. L. (1994). Reassessment of expectations as a comparison standard in measuring service quality: Implications for further research. *The Journal of Marketing*, **58**(1), pp. 111–124.

Pearl J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pearl J. (2000). *Causality: models, reasoning, and inference*. Cambridge Univ Pr.

Piccinato L. (2009). *Metodi per le decisioni statistiche*. Springer.

Pólya G.; Conway J. (2004). *How to solve it: A new aspect of mathematical method*. Princeton University Press Princeton, NJ.

Popper K. (1934). Logic der Forschung. *Vienna, Julius Springer Verlag.(1959) The Logic of Scientific Discovery, New York, Basic Books.(1962) Conjectures and Refutations: The Growth of Scientific Knowledge, New*.

Ramsey F. (1931). General propositions and causality. *London: Kegan Paul, Trench & Trubner*, pp. 237–55.

Rissanen J. (1978). Modeling by shortest data description. *Automatica*, **14**(5), 465–471.

Robinson R. (1977). Counting unlabeled acyclic digraphs. *Combinatorial mathematics V*, pp. 28–43.

Russell B. (1912). On the notion of cause. In *Proceedings of the Aristotelian Society*, pp. 1–26. JSTOR.

Russolillo G. (2009). *Partial Least Squares Methods for Non-Metric Data*. Tesi di Dottorato di Ricerca, Università degli Studi di Napoli-Federico II.

Salmon W. (1984). *Scientific explanation and the causal structure of the world*, volume 121. Princeton University Press Princeton, NJ.

Savage L. (1972). *The foundations of statistics*. Dover Pubns.

Spearman C. (1904). General Intelligence Objectively Determined and Measured. *The American Journal of Psychology*, **15**(2), 201–292.

Spirtes P.; Glymour C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, **9**(1), 62.

Spirtes P.; Glymour C.; Scheines R. (1993). Causation, Prediction, and Search (Lecture Notes in Statistics 81).

Spirtes P.; Glymour C.; Scheines R. (2000). *Causation, prediction, and search*. The MIT Press.

Steiger J.; Lind J. (1980). Statistically based tests for the number of common factors. In *annual meeting of the Psychometric Society, Iowa City, IA*, volume 758.

Tenenhaus M. (1998). *La régression PLS: théorie et pratique*. Editions Technip.

Tenenhaus M.; Amato S.; Esposito Vinzi V. (2004). A global goodness-of-fit index for pls structural equation modelling. In *Proceedings of the XLII SIS Scientific Meeting*, volume Contributed Papers, CLEUP, Padova, pp. 739–742.

Tenenhaus M.; Esposito Vinzi V.; Chatelin Y.; Lauro C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, **48**(1), 159–205.

Thurstone L. (1947). *Multiple Factor Analysis*. The University of Chicago Press Chicago, IL.

Trinchera L. (2007). *Unobserved Heterogeneity in Structural Equation Models: a new approach to latent class detection in PLS Path Modeling*. Tesi di Dottorato di Ricerca, Università degli Studi di Napoli-Federico II.

Verma T.; Pearl J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–270. Elsevier Science Inc.

Williamson J. (2005). *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press, USA.

Wold H. (1966a). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, **1**, 391–420.

Wold H. (1966b). Nonlinear estimation by iterative least squares procedures. *Festschrift for J. Neyman*, p. 411.

Wold H. (1973). Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments, in: PR Krishnaiah (Ed.), Multivariate Analysis.

Wold H. (1975a). Path models with latent variables: the NIPALS approach. *Quantitative sociology: International perspectives on mathematical and statistical modeling*, pp. 307–357.

Wold H. (1975b). Soft modeling by latent variables: the nonlinear iterative partial least squares approach. *Perspectives in probability and statistics, papers in honour of MS Bartlett*, pp. 520–540.

Wold H. (1979). *Model construction and evaluation when theoretical knowledge is scarce: An example of the use of partial least squares*. Université de Genève, Faculté des Sciences Économiques et Sociales.

Wold H. (1982). Soft modeling: the basic design and some extensions. *Systems under indirect observation*, **2**, 589–591.

Wold S.; Martens H.; Wold H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Matrix Pencils*, pp. 286–293.

Wold S.; Ruhe A.; Wold H.; Dunn III W. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, **5**, 735.

Wold S.; Kettaneh-Wold N.; Skagerberg B. (1989). Nonlinear PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, **7**(1-2), 53–65.

Wold S.; Sjöström M.; Eriksson L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, **58**(2), 109–130.

Wright S. (1921). Correlation and causation. *Journal of agricultural research*, **20**(7), 557–585.

Wright S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, **5**(3), 161–215.

Wu W. W. (2010). Linking bayesian networks and pls path modeling for causal analysis. *Expert Systems with Applications*, **37**(1), 134 – 139.

XLSTAT2010.1 (2010).

Zanella A. (2001). Valutazione e modelli interpretativi di Customer Satisfaction: una presentazione di insieme. *Università Cattolica del S. Cuore, Istituto di Statistica, Serie EPN*, **105**, 7–20122.