UNIVERSITÀ DEGLI STUDI DI PADOVA

Facoltà di SCIENZE STATISTICHE

Corso di Laurea Specialistica in

SCIENZE STATISTICHE, DEMOGRAFICHE E SOCIALI

Tesi di laurea

Combinazione di modelli grafici per l'analisi di network biologici

Relatrice: Prof.ssa Monica Chiogna

Correlatrice: Dott.ssa Maria Sofia Massa

Laureando: Lorenzo Maragoni

A chi, in qualsiasi modo, ci ha creduto.

Indice

1. Introduzione

1.1. I network biologici

Lo studio statistico dei network biologici – appartenente alla branca della *genomica* – si basa sull'analisi dei profili di espressione dei geni: sulle somiglianze e differenze riscontrate nel comportamento di geni diversi, dello stesso gene in tempi successivi, dello stesso gene in individui con caratteristiche differenti. L'osservazione di un legame statistico in una coppia o un gruppo di geni porta a ipotizzare un effettivo collegamento biologico tra gli stessi, e può rappresentare il punto di partenza per analisi più approfondite.

I profili di espressione sono descritti dai *dati da microarray*, i quali, tramite apposite tecniche, raccolgono l'espressione di segmenti di DNA a partire da materiale biologico. L'obiettivo delle analisi è di arrivare a ricostruire network di grandi dimensioni che mostrino i legami (e le assenze di legame) all'interno di un ampio insieme di geni. La comprensione della struttura di un network biologico può essere di fondamentale

importanza nello studio di malattie ereditarie o a carattere degenerativo, come i tumori: a tale proposito, è evidente che i risultati ottenuti con metodi statistici dovranno poi essere confermati da un punto di vista biologico. Anche rimanendo nel contesto degli studi statistici, tuttavia, è importante ricordare che le conclusioni tratte da studi di tipo puramente osservazionale dovranno poi essere comprovate da studi perturbativi, in cui il ricercatore sia in grado di manipolare l'espressione di singoli geni di interesse, e di studiare le conseguenze di questi cambiamenti.

1.2. La combinazione di modelli grafici

I modelli grafici sono uno strumento di analisi multivariata, utile per esplorare la struttura delle relazioni esistenti all'interno di un insieme numeroso di variabili, e adatto pertanto all'analisi dei network biologici. Uno dei punti di forza principali dei modelli grafici sta nel consentire rappresentazioni di semplice interpretazione: un modello grafico è costituito di due classi di elementi: un insieme di *nodi* – che rappresentano le variabili – e un insieme di *archi* – che vengono tracciati tra due nodi se tra le corrispondenti variabili esiste una relazione.

I modelli grafici possono comprendere variabili di natura diversa (discrete o continue) e includere di conseguenza una varietà di distribuzioni. Un caso semplice, ma interessante dal punto di vista applicativo e di principale interesse per questa tesi, è quello in cui si assume che tutte le variabili coinvolte nel network seguano una distribuzione normale. In tal caso, il modello grafico sarà detto *gaussiano*, e si potrà indicare con la sigla GGM.

La necessità (opportunità) di combinare modelli grafici diversi nasce dai contesti in cui capita di disporre di informazioni provenienti da studi diversi su uno stesso problema, o su

problemi parzialmente sovrapponibili; nell'ambito della genomica, in particolare, è frequente incontrare ricerche che hanno in comune soltanto una parte delle variabili oggetto di studio. Questa tesi non ha l'obiettivo di trattare considerazioni riguardanti la comparabilità tra ricerche diverse (in termini di numerosità campionaria, tecniche di campionamento, scale di misurazione delle variabili); piuttosto, le domande di interesse riguarderanno l'opportunità o meno di effettuare indagini *ex novo*, includendo un grande insieme di variabili, quando è invece possibile in qualche modo sfruttare risultati già ottenuti da altre ricerche, anche se questi, presi singolarmente, riguardano solo una parte delle variabili oggetto di studio. Si indagheranno, in pratica, le condizioni e le conseguenze del combinare modelli grafici, nel contesto dell'analisi di network biologici.

1.3. Organizzazione della tesi

Sarà innanzitutto effettuata una panoramica sui metodi proposti in letteratura per affrontare l'analisi dei network biologici (*capitolo 2*). In un secondo tempo, sarà approfondita la natura dei modelli grafici, i loro metodi di applicazione e di combinazione, presentando alcuni risultati di interesse raggiunti in precedenti ricerche (*capitolo 3*). Successivamente saranno mostrati i risultati ottenuti tramite due studi di simulazione originali relativi alla combinazione di modelli grafici (*capitolo 4*) e una loro applicazione a dati reali (*capitolo 5*). La tesi sarà chiusa da una sintesi delle conclusioni raggiunte (*capitolo 6*).

2. Metodi di analisi di network biologici

2.1. Il concetto di network

All'interno di una struttura di DNA, i geni sono organizzati in network altamente strutturati. Tra alcuni geni esiste un legame più stretto, tra altri questo è estremamente labile: sintetizzando l'insieme dei legami in una rappresentazione grafica, è possibile ottenere un'immagine chiara di quali siano gli insiemi di geni interconnessi.

Per comprendere la struttura di un network, sono stati sviluppati diversi approcci dal punto di vista biologico e statistico. In una situazione ottimale si dovrebbe essere in grado di intervenire dall'esterno sullo stato della cellula, così da poter analizzare le reazioni del network conseguenti a questa manipolazione: studi di questo tipo sono detti *perturbativi*. Dati gli alti costi, in termini di tempo e tecnologia necessaria, necessari per effettuare studi di questo genere, spesso è più conveniente (o è semplicemente l'unica possibilità)

compiere studi *osservazionali*, indagando i legami tra i geni tramite somiglianze e differenze nei loro naturali profili d'espressione, senza intervenire direttamente sullo stato delle cellule. Nonostante la manipolazione attiva del materiale genetico sia l'unico modo per indagare i meccanismi causali sottostanti al network in analisi, in questa tesi gli studi considerati sono di tipo osservazionale: non si dovrà dunque cadere nell'errore di leggere in chiave causale relazioni che si riferiscono soltanto a correlazione o dipendenza statistica. Al tempo stesso, d'altra parte, sarà sempre sottinteso l'invito a prendere i risultati proposti come punti di partenza per la formulazione di nuove ipotesi, da verificare con analisi più approfondite.

Diversi metodi sono stati proposti in letteratura per trattare dati provenienti da studi osservazionali nel campo biologico. La differenza tra gli approcci dipende dalle diverse declinazioni del concetto di "relazione" tra variabili adottate nei singoli studi: tra i metodi di più larga diffusione si possono ricordare i *network di co-espressione*, i *modelli completamente condizionati* (*full conditional models*), i *modelli con dipendenza di primo ordine* o *di ordine qualsiasi* (*low-order conditional models*), i *network bayesiani*. Il riferimento principale per una panoramica su questi metodi si trova in Markowetz e Spang (2007), a cui si rimanda per approfondimenti nei casi in cui un'altra fonte non sia espressamente citata.

Il caso di studio preso come riferimento è un network costituito da un insieme di *p* variabili, $V = \{X_1,...,X_p\}$; per il quale la stima del modello si basa su un campione di numerosità *N*. Se ogni variabile rappresenta il nodo di un grafo e ogni arco tra due nodi rappresenta l'esistenza di una relazione tra le rispettive variabili, stimare un modello equivale a identificare l'insieme degli archi *E* (nel nostro contesto ipotizzati non orientati)

che rappresentano la struttura di dipendenza tra i nodi, relativamente al campione in esame. Il risultato delle analisi sarà un grafo del tipo $G = (V, E)$.

2.2. Network di co-espressione

La costruzione di modelli basata sulla *co-espressione* – ovvero, sulla somiglianza nel profilo di espressione dei geni – trova giustificazione nell'idea che questa similarità sia manifestazione di un vero legame biologico fra i geni in questione. Per quanto non possa in alcun caso esserne garanzia, la co-espressione ha spesso mostrato di essere legata a effettive somiglianze biologiche all'interno del network. Basandosi su questa ipotesi (che deve nei singoli casi essere approfondita e verificata), un network di co-espressione si costruisce assegnando ad ogni coppia di geni un punteggio relativo alla frequenza con cui, nel campione osservato, manifestano un profilo di espressione simile: i nodi corrispondenti a geni con un punteggio superiore a una certa soglia vengono collegati da un arco.

Prima ancora di definire una soglia di "somiglianza" opportuna per gli obiettivi dello studio, è però necessario tradurre il concetto stesso di *somiglianza* tra profili di espressione: tra le molteplici possibilità a disposizione, una delle scelte più comuni è di utilizzare il valore di *correlazione* (ricordando che, nel caso di modelli grafici gaussiani, si ha che l'incorrelazione equivale all'indipendenza). I network basati sull'uso di questa misura hanno un'interpretazione semplice, e possono essere ricostruiti accuratamente anche nei casi in cui $N << p$. E' necessario però tenere sotto controllo il livello di significatività delle relazioni trovate, dato che – in network non troppo estesi – è possibile che tutti i geni finiscano in qualche modo per essere correlati, così che profili di espressione simili si manifestino a volte per puro caso, piuttosto che per ragioni connesse

a legami biologici. Per controllare la significatività delle relazioni può essere utile ricorrere a tecniche di validazione incrociata (*split-half*) o a permutazioni dei dati originari, per le quali stimare di volta in volta il modello migliore e confrontarlo con quello ottenuto con i dati di partenza – si vedano Bickel (2005), oppure Yamanishi et al. (2004) per approfondimenti. E' necessario in ogni caso tenere presente che la correlazione è una misura lineare di dipendenza: per ricostruire la struttura di un network includendo relazioni non lineari occorre ricorrere a misure diverse, ad esempio di tipo non parametrico, come quelle basate sul metodo del nucleo proposte da Yamanishi et al. (2004).

Le analisi di co-espressione possono avere obiettivi più ampi che la ricostruzione del network in sé: possono ad esempio essere estese a serie temporali – introducendo come fattori di ritardo le correlazioni con valori espressi dagli stessi geni in istanti precedenti, come mostra Bickel (2005) – o a studi di co-espressione differenziale (Kostka e Spang, 2004), che hanno l'obiettivo di trovare i geni che mostrano comportamenti differenti sotto differenti condizioni (ad esempio, nel caso di buono o cattivo stato di salute della cellula).

Il limite principale della correlazione come strumento per la ricostruzione di network biologici sta nel fatto che essa misura soltanto le relazioni marginali tra le variabili: il valore di correlazione tra due variabili non tiene conto degli eventuali legami con il resto delle variabili presenti nel network. Per questo motivo risulta impossibile distinguere tra una co-espressione frutto di un legame diretto tra due geni, e una co-espressione "indiretta" dovuta all'azione confondente di geni intermedi. Gli studi perturbativi sono la chiave per superare problemi della classe di quello citato, e permettono di risolvere situazioni come quella rappresentata nella Figura 1.1 (tratta da Markowetz e Spang, 2007), in cui i geni in gioco sono tutti legati, ma i legami presenti potrebbero essere originati da meccanismi biologici diversi.

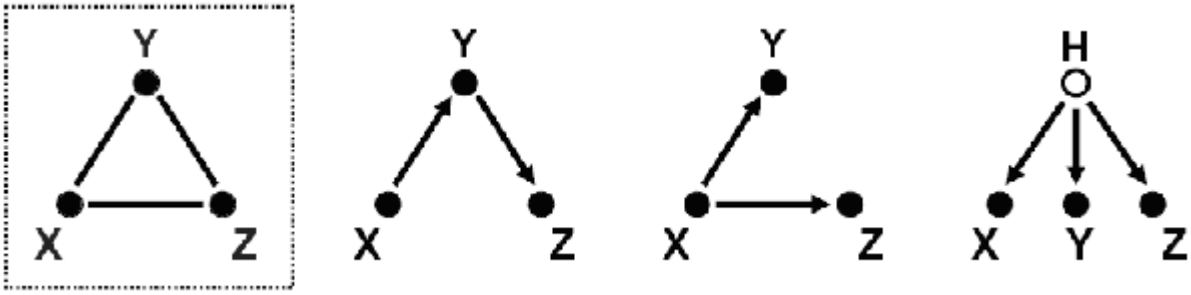

Figura 1.1: differenti meccanismi possibili dietro una co-espressione. Da sinistra a destra: manifestazione della co-espressione; geni regolati in cascata; un gene regola gli altri due; un gene regolatore nascosto governa tutti e tre i geni.

Per indagare quale tra i possibili meccanismi sia quello effettivo, è necessario modificare dall'esterno lo stato di uno dei geni e vedere quali conseguenze ha questa modifica. Se il meccanismo "vero" fosse che *X* regola *Y* che a sua volta regola *Z*, si troverebbe che modificando *Y* si modifica anche *Z*, ma non *X*; se il meccanismo fosse invece che *Y* regola sia *X* che *Z*, modificando *Y* si troverebbero variazioni in entrambe le altre due variabili. Ancora, se al variare di *Y* non seguisse alcuna modifica né in *X* né in *Z*, ci si troverebbe nella situazione in cui tutti e tre i geni sono legati, ma sono a loro volta governati da un regolatore "nascosto" *H* (esterno all'insieme considerato).

2.3. Indipendenza condizionata

Come si è accennato in precedenza, studi di tipo perturbativo sono di difficile realizzazione. Tuttavia, rimanendo nel contesto degli studi osservazionali, rimane possibile indagare le relazioni tra variabili in modo più approfondito rispetto agli studi di correlazione semplice: è necessario per questo fare riferimento al concetto di *indipendenza condizionata*. Come si è visto, in un network di co-espressione le due variabili tra cui si

vuole valutare l'esistenza di un legame sono considerate come slegate dal resto della rete. Introdurre il concetto di indipendenza condizionata, invece, vuol dire tenere conto dei valori assunti da tutte le altre variabili del network nello studio di ogni singola relazione. Due variabili saranno allora dette indipendenti, condizionatamente al valore assunto dalle altre, se esse si rivelano indipendenti qualsiasi sia il comportamento del resto del network (e non "ignorando" il comportamento del network, come accade nello studio dell'indipendenza marginale).

La definizione probabilistica di indipendenza condizionata è la seguente: se X e Y sono variabili aleatorie con distribuzione congiunta $f_{X,Y}(x,y)$ e Z è un insieme di q altre variabili, scrivere che $X \perp Y \mid Z$ equivale a dire che

$$f_{X,Y|Z}(x,y \mid z) = f_X(x \mid z)\, f_Y(y \mid z),$$

ovvero che la distribuzione condizionata di X dato Y e Z è, di fatto, completamente determinata dal solo valore di Z, essendo Y superflua se Z è nota. In questi casi, il legame marginale eventualmente presente tra le due variabili si rivela spurio, ovvero esistente solo grazie alla mediazione di altre variabili (tutte o parte di quelle contenute in Z). I modelli basati sull'indipendenza condizionata sono dunque uno strumento di studio più affidabile rispetto a quelli di co-espressione basati sulla semplice correlazione marginale.

A seconda di come è strutturato l'insieme Z, la classe di modelli a indipendenza condizionata può essere suddividesa in sotto-classi: se Z contiene una singola variabile si ha un modello basato sull'indipendenza di primo ordine; se Z coincide con $V \setminus \{X,Y\}$ si ha l'indipendenza "completa" (di ordine massimo); se Z è un qualsiasi altro sottoinsieme di variabili, si ha un modello a indipendenza di ordine intermedio. I network di co-espressione possono essere visti come il caso particolare in cui $Z = \varnothing$.

2.4. Indipendenza completamente condizionata

I *modelli a indipendenza completamente condizionata* (o network markoviani, o modelli grafici non orientati) rappresentano il caso in cui la correlazione tra due variabili è calcolata condizionandosi al valore assunto da *tutte* le altre variabili nel modello. Nei grafi corrispondenti a modelli di questo tipo, due generiche variabili X_i e X_j non sono collegate da un arco se e solo se X_i è indipendente da X_j dato $V \setminus \{X_i, X_j\}$.

Nel caso di un GGM con matrice di covarianza Σ e *matrice di precisione* definita come $\Omega = \Sigma^{-1}$, individuare la struttura di un modello è particolarmente semplice, in quanto il valore dei singoli elementi ω_{ij} di Ω è collegato al coefficiente di correlazione parziale ρ_{ij} tra X_i e X_j. Si ha infatti che

$$\rho_{ij} = -\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}} ,$$

e dunque che, semplicemente:

$$X_i \perp X_j \mid X_{rest} \Leftrightarrow \omega_{ij} = 0 ,$$

dove $X_{rest} = V \setminus \{X_i, X_j\}$. Secondo questa relazione allora, un GGM è un grafo non orientato costruito su un insieme di vertici *V*, in cui a ogni vertice corrisponde una variabile casuale distribuita in modo normale, e l'insieme degli archi *E* è definito dalle correlazioni parziali non nulle, ovvero dagli elementi non nulli di Ω. Per stimare il GGM che meglio si adatta a un campione di dati è allora sufficiente calcolare la matrice di precisione campionaria $\hat{\Omega}$ e identificarne gli elementi che valgono zero. Questo risultato può essere raggiunto testando ipotesi per la significatività di singoli archi (Drton e Perlman, 2004), oppure sull'insieme complessivo degli archi possibili, utilizzando ad esempio tecniche di regressione a passi (Smith e Whittaker, 1999).

E' stato detto che l'uso della correlazione parziale, a differenza della correlazione semplice, consente di tenere conto dell'influenza di tutti i geni del network nello studio delle singole relazioni. Questa proprietà permette – oltre che di ridimensionare l'importanza di correlazioni marginali alte, nel caso queste fossero in realtà da attribuirsi all'azione di variabili terze – di scoprire che invece alcuni geni, solo debolmente correlati ad un gene di interesse, rivelano in realtà un legame più forte se si tiene conto dell'influenza degli altri geni nel loro "vicinato" (situazioni che sarebbero facilmente ignorate in un contesto di correlazione semplice).

Un ulteriore vantaggio dell'uso della correlazione parziale è il seguente: dato che – in network non troppo estesi – tutti i geni finiscono per essere in qualche misura correlati, il coefficiente di correlazione è un criterio forte per scoprire le indipendenze, ma debole per scoprire le dipendenze. Il coefficiente di correlazione parziale invece, tende di base ad attestarsi su valori quasi nulli, e a discostarsi dallo zero solo nei casi in cui il legame tra le variabili sia sufficientemente forte: dato che l'obiettivo principale degli studi di nostro interesse è di individuare l'insieme degli archi da tracciare nel grafo (le dipendenze significative, piuttosto che le indipendenze), la correlazione parziale appare uno strumento più appropriato di quella semplice.

Le procedure di stima e di selezione di modelli grafici in generale (e di GGM in particolare) presentano problemi nelle situazioni (frequenti nella pratica) in cui $N << p$, ovvero se il numero di geni è molto alto rispetto al numero di osservazioni disponibili: in questi casi, infatti, non è possibile calcolare l'inversa della matrice di covarianza, in quanto questa non è di rango pieno. Le usuali soluzioni, basate ad esempio sull'uso della matrice pseudo-inversa di Moore-Penrose o sul ri-campionamento di tipo *bootstrap*, possono nel nostro contesto essere affiancate o sostituite da considerazioni di tipo sostanziale sulla biologia cellulare. L'idea è che l'ampio insieme di geni che sembra regolare un determinato

gene di interesse sia in realtà riconducibile ad un insieme più limitato, se ci si spinge a considerare come nulli i coefficienti di correlazione parziali meno significativi. Nonostante sia possibile in questo modo (eliminando via via i geni meno legati agli altri) ricondursi a una situazione in cui $N > p$, il metodo presenta alcune componenti di arbitrarietà, che fanno sì che le stime con esso calcolate rimangano instabili (si veda Kishino e Waddell, 2000).

I network a indipendenza completamente condizionata sono legati ad un'altra classe di modelli, chiamati *network di dipendenza*: questi sono costruiti sulla base di numerose regressioni, in ognuna delle quali la variabilità di un gene è spiegata tramite la variabilità di tutti gli altri (il che aiuta a ridurre il numero di dimensioni in gioco, nei contesti in cui $N << p$). I modelli utilizzabili per le regressioni locali possono essere lineari (con eventuale penalizzazione per il numero di parametri), non lineari e bayesiani. Il principale svantaggio di tale approccio sta nel rischio di incoerenze tra il modello complessivo stimato e le singole regressioni: per questo motivo, il modello finale deve essere trattato come una semplice approssimazione del modello vero. Nonostante i suoi difetti, questa classe di modelli è ampiamente utilizzata per la sua flessibilità, e per vantaggi di tipo computazionale rispetto ai modelli a indipendenza completamente condizionata.

2.5. Indipendenza condizionata di primo ordine

I casi di studio in cui $N << p$, come visto, sono difficili da affrontare. Oltre a cercare di aggiustare la procedura tramite le tecniche sopra descritte, può essere utile adottare un approccio diverso da quello della piena indipendenza condizionata: si può analizzare la dipendenza tra *X* e *Y* dato un singolo gene *Z*, e ripetere questo procedimento

condizionandosi ogni volta a un singolo gene in $V \setminus \{X, Y\}$. Secondo questo criterio, detto *indipendenza condizionata di primo ordine*, tra *X* e *Y* è tracciato un arco se e solo se nessuno degli altri geni è singolarmente in grado di spiegare la correlazione tra di esse (posto naturalmente che questa sia diversa da zero), ovvero se *X* è dipendente da *Y* dato *Z* per ogni $Z \in V \setminus \{X, Y\}$. Dato che, seguendo questo approccio, ogni test coinvolge solo tre variabili per volta, sono notevolmente ridimensionati i problemi connessi a un numero di geni elevato rispetto alla numerosità campionaria.

E' naturalmente possibile (anche se di minore utilità pratica) estendere questo metodo all'*indipendenza di secondo, terzo, k-esimo ordine*: è sufficiente scegliere come *Z* , invece di singoli variabili, tutte le possibili coppie, triple, *k*-uple di variabili contenute in $V \setminus \{X, Y\}$.

2.6. Network bayesiani

Le definizioni viste nel paragrafo precedente possono essere generalizzate tramite il concetto di *indipendenza di ordine qualsiasi*. Secondo questo approccio, tra due nodi appartenenti a un network è tracciato un arco se e solo se nessun sottoinsieme (di qualsiasi numerosità) delle altre variabili può spiegare la correlazione tra loro. In altre parole, un arco è tracciato tra *X* e *Y* se e solo se *X* dipende da *Y* dato *Z* per ogni *Z*, dove $Z \subseteq V \setminus \{X, Y\}$ è un insieme senza restrizioni di cardinalità. Il concetto di indipendenza di ordine qualsiasi include quello di indipendenza marginale se $Z = \varnothing$, quello di indipendenza completamente condizionata se $Z = V \setminus \{X, Y\}$, quello di indipendenza di primo ordine se *Z* è composto di un solo elemento: di conseguenza, il numero totale di archi tracciati seguendo questo approccio risulterà piuttosto piccolo, se confrontato con i singoli metodi precedenti.

Il modello probabilistico che risulta dall'indipendenza di ordine qualsiasi è detto *network bayesiano*: la struttura di dipendenza è descritta sia da distribuzioni di probabilità locali relative a sottoinsiemi dei vertici del network, sia dalla distribuzione congiunta su tutti i vertici del grafo. Una proprietà importante di questi modelli riguarda il modo in cui è possibile di fattorizzare la distribuzione congiunta: questo infatti corrisponde a una segmentazione del network in *famiglie*, che possono essere trattate individualmente. Una conseguenza di questa proprietà è che – per quanto tecnicamente il grafo rimanga non orientato – la sua struttura può implicare un ordinamento tra alcune delle variabili. Infatti, i nodi possono possedere dei *genitori* – variabili che li rendono indipendente da tutti gli altri predecessori – ed essere a loro volta genitori di altri nodi. Nel caso dei GGM, ogni famiglia seguirà una distribuzione di probabilità di tipo normale, la cui media sarà una combinazione lineare dei valori assunti dai nodi genitori. In casi più generali, le distribuzioni di probabilità delle famiglie possono essere costruite ad esempio tramite alberi di regressione binari (dove le foglie sono associate a distribuzioni normali univariate, come mostrato da Segal et al., 2005), o tramite approcci non parametrici (si veda ad esempio Imoto, Goto e Miyano, 2002).

2.7. Benchmarking

La ricostruzione delle caratteristiche biologicamente rilevanti di un network cellulare è un compito laborioso. In alcuni contesti la stessa valutazione degli effetti sull'inferenza del variare della numerosità campionaria, o della validità delle assunzioni su cui si fonda il modello, o della tecnica di campionamento, è difficile da effettuare. Un recente studio comparativo (Werhli et al., 2006) mostra differenze non significative fra network di co-

espressione, GGM e network bayesiani, in un'applicazione su dati simulati (anche non lineari), e su dati reali. In generale, dunque, non sempre è possibile affermare che i più alti costi computazionali richiesti dai network bayesiani trovino giustificazione in risultati più accurati. Studi come quello citato mostrano inoltre che l'utilizzo di modelli grafici per l'analisi di dati provenienti da microarray rischia di essere in generale poco soddisfacente: la parte biologicamente rilevante dei network che si riesce a cogliere è spesso piuttosto limitata, e sono necessarie numerosità campionarie sostanziose anche per raggiungere risultati modesti. Nella consapevolezza dei limiti di cui l'applicazione di modelli grafici può soffrire, si rimanda al capitolo introduttivo per una giustificazione dell'uso di queste tecniche, almeno come tecnica esplorativa, da approfondire eventualmente in seguito tramite studi di tipo perturbativo.

3. Combinazione di GGM

3.1. Modelli grafici

Come già visto in precedenza, un grafo (non orientato) $G = (V, E)$ è una struttura costituita da un insieme finito di nodi *V* – rappresentanti le variabili – e da un insieme finito di archi *E*, che collegando due nodi rappresentano l'esistenza di una relazione tra le corrispondenti variabili. Il criterio scelto per misurare la relazione tra le variabili è quello della correlazione parziale, introdotto nel capitolo precedente. Il riferimento principale per questo capitolo è l'opera di Edwards (2000).

Se si assume che le variabili *X* e *Y* abbiano densità congiunta $f_{X,Y}(x, y)$, un arco tra i loro rispettivi nodi è allora tracciato se e solo se

$$f_{X,Y|Z}(x, y \mid z) = f_X(x \mid z) f_Y(y \mid z),$$

dove $Z = V \setminus \{X, Y\}$.

Quando due vertici sono collegati da un arco, essi si dicono *adiacenti*, e nel caso in cui tutti i vertici di un grafo siano adiacenti (ovvero esista un arco tra tutte le possibili coppie di vertici), esso si dice *completo*. Ad esempio, nella Figura 3.1 il grafo sulla sinistra non è completo, in quanto i vertici X e Z non sono adiacenti. Si verifica immediatamente che invece, nella stessa figura, il grafo sulla destra è completo.

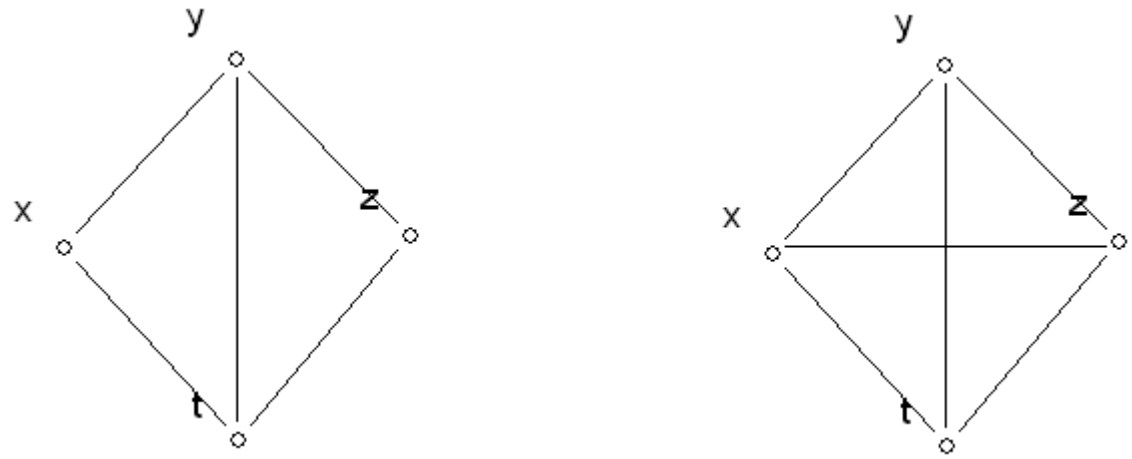

Figura 3.1: rappresentazione dei grafi //XYT YZT e //XYZT

Un qualsiasi sottoinsieme dei vertici di partenza induce un sotto-grafo di quello originario; questo sottoinsieme è detto *completo* se il sotto-grafo da esso indotto è completo. Un sottoinsieme è *massimamente completo* se, aggiungendo ad esso un qualsiasi altro vertice del grafo di partenza, il nuovo sottoinsieme non risulta completo; in tal caso, il sottoinsieme è anche chiamato *clique*, e può essere scritto tramite la semplice elencazione dei nomi dei vertici. Nella figura precedente, a partire dal grafo sulla sinistra possono essere indotti diversi sotto-grafi completi: sono quelli composti dai vertici $\{X, Y\}$, $\{Y, Z\}$, $\{Z, T\}$, $\{X, T\}$, $\{Y, T\}$, $\{X, Y, T\}$ e $\{Y, Z, T\}$ (gli ultimi due dei quali sono delle *cliques* e possono dunque essere scritte come *XYT* e *YZT*); molteplici sotto-grafi sono presenti nel

grafico sulla destra (da cui però, essendo esso stesso completo, non possono essere indotti sotto-grafi massimamente completi).

Si dice *percorso* (di lunghezza *k*) una qualsiasi sequenza di *k* vertici adiacenti all'interno di un grafo. In riferimento a tre sottoinsiemi di vertici del grafo, si dice che il primo *separa* gli altri due se qualsiasi percorso tra il secondo e il terzo deve necessariamente intersecarlo. Si può notare (e si può verificare nel prossimo esempio) che affermare che due variabili non sono adiacenti equivale ad affermare che tra di esse esiste almeno una variabile che le separa. Ad esempio, nel grafo in Figura 3.2 (in cui è possibile individuare le *cliques XYZT*, *ZW* e *ZV*), si può dire ad esempio che $\{Z\}$ separa $\{W\}$ da $\{X,Y,T\}$, o che $\{Z,T\}$ separa $\{V\}$ da $\{X,Y\}$.

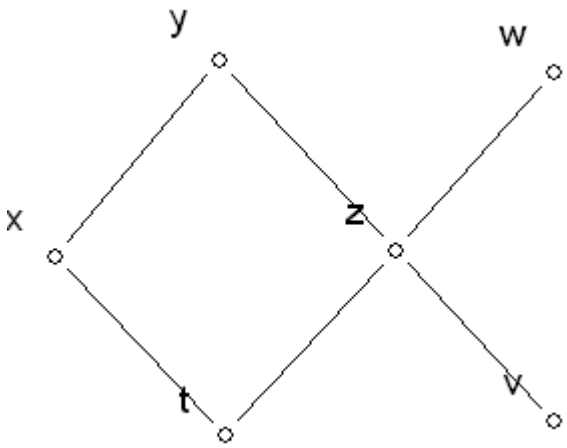

Figura 3.2: rappresentazione del grafo //XYZT ZV ZW.

L'identificazione di un modello grafico avviene generalmente per elencazione delle sue *cliques*, precedute convenzionalmente da una doppia barra, come nella scrittura *//XYT YTZ* (che si riferisce al grafo sulla sinistra nella Figura 3.1) o in *//XYZT* (che èil grafo sulla destra in Figura 3.1) o in *//XYZT ZV ZW* (che è il grafo in Figura 3.2). Per questo motivo le *cliques* sono anche dette *generatori* del modello.

Osservando un grafo, è possibile individuare immediatamente quali variabili sono incorrelate date le altre: esse sono (unicamente) quelle rappresentate da vertici non adiacenti. Questa fondamentale caratteristica è detta *proprietà di Markov per coppie di variabili* (*pairwise Markov property*), estendibile per insiemi generali di variabili alla *proprietà globale di Markov* (*global Markov property*), la quale afferma che se due insiemi di variabili X_1 e X_2 sono separati da un terzo insieme di variabili X_3, allora $X_1 \perp X_2 \mid X_3$ (Edwards, 2000). Ad esempio, nel grafo sulla destra in Figura 3.1, poiché l'insieme $\{Y,T\}$ separa X da Z, si ha che $X \perp Z \mid \{Y,T\}$. Sotto condizioni piuttosto generali, è possibile dimostrare l'equivalenza tra la proprietà di Markov riferita a coppie di variabili e quella globale.

3.2. Modelli grafici gaussiani (GGM)

Le variabili coinvolte in un modello grafico possono essere sia di tipo discreto che continuo. I modelli di cui ci si occupa in questa tesi si limitano a considerare variabili di questo secondo tipo, in particolare aventi distribuzione normale multivariata. In riferimento a un grafo G con p variabili, un *modello grafico gaussiano* è definito come la famiglia di distribuzioni normali multivariate

$$Y \sim N_p(\mu, \Sigma),$$

per le quali si assume, senza perdere di generalità, che $\mu = 0$, e si assume inoltre che la *matrice di precisione* Ω sia tale che $\Omega = \Sigma^{-1} \in S^+(G)$, dove $S^+(G)$ rappresenta l'insieme delle matrici simmetriche definite positive i cui elementi sono nulli nel caso non sia presente un arco tra i corrispondenti elementi di G. Questa ipotesi è legata, come già visto, ad una proprietà del coefficiente di correlazione parziale tra l'i-esima e la j-esima

variabile (date tutte le altre): esso ha infatti valore nullo se e solo se è nullo l'elemento di posto (*i,j*) nella matrice di precisione.

La funzione di densità di *Y* si può scrivere come

$$f(y) = (2\pi)^{-p/2}\left|\Sigma\right|^{-1/2} - \exp(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu))$$

e dunque il suo logaritmo come

$$\ln(f(y)) = \frac{1}{2}(-p\ln(2\pi) - \ln\left|\Sigma\right| - (y-\mu)'\Sigma^{-1}(y-\mu)).$$

Di conseguenza la funzione di log-verosimiglianza, indicata con *N* la numerosità campionaria e con Ω la matrice di precisione, risulterà pari a

$$l(\mu,\Omega) = \frac{1}{2}(-Np\ln(2\pi) - N\ln\left|\Omega^{-1}\right| - \sum_{i=1}^{N}(y_i-\mu)'\Omega(y_i-\mu)),$$

dove i pedici relativi alla *y* si riferiscono alle diverse unità campionarie. Questa espressione, riscrivendo opportunamente l'ultimo termine (sostituendo allo scarto tra unità e valore atteso la somma degli scarti tra le unità campionarie e la media campionaria e tra la media campionaria e il valore atteso), si semplifica in

$$l(\mu,\Omega) = \frac{1}{2}(-Np\ln(2\pi) - N\ln\left|\Omega^{-1}\right| - Ntr(\Omega S) - N(\bar{y}-\mu)'\Omega(\bar{y}-\mu)),$$

dove *S* rappresenta la matrice di covarianza campionaria. Dato che – per un modello che abbia come generatori le *cliques* $p_1,\ldots,p_t$ – una statistica sufficiente minimale è costituita dalla media campionaria e da tutte le sottomatrici di *S* relative a righe e colonne corrispondenti ai generatori, le equazioni di verosimiglianza si ottengono uguagliando queste quantità ai loro rispettivi valori attesi sotto il modello, ottenendo che

$$\hat{\mu} = \bar{y}$$

e

$$\hat{\Sigma}_{aa} = S_{aa}$$

per $a = p_1,..., p_t$. Dunque, la stima di massima verosimiglianza di μ è data dalla media campionaria, mentre quella di Σ è costituita da una matrice che equivale alla matrice di covarianza campionaria, tranne per gli elementi nei posti corrispondenti a variabili tra loro incorrelate (indipendenti): in quelle posizioni saranno presenti valori tali per cui, nella matrice inversa, risultino valori nulli nelle stesse posizioni. Ad esempio, per l'insieme di variabili *V*={*X*,*Y*,*Z*,*T*}, sotto il modello //*XYT YZT* (rappresentato dal grafo sulla sinistra in Figura 3.1) l'elemento di posto (*1*,*3*) di $\hat{\Sigma}$ sarà l'unico a discostarsi dal suo corrispondente campionario, e avrà un valore tale che $\hat{\omega}_{1,3} = 0$, dove $\hat{\omega}_{1,3}$ è l'elemento di posto (*1*,*3*) di $\hat{\Omega}$.

Sostituendo nella funzione di log-verosimiglianza le espressioni delle stime così trovate, l'ultimo termine si semplifica (in quanto $\hat{\mu} = \bar{y}$) e, poiché *S* si differenzia da $\hat{\Sigma}$ solo negli elementi che danno luogo a zeri in $\hat{\Omega}$, allora $tr(\hat{\Omega}S) = tr(\hat{\Omega}\hat{\Sigma}) = p$. Dunque nel suo massimo la log-verosimiglianza per un modello generico vale

$$\hat{l}_m = l(\hat{\mu}, \hat{\Sigma}) = \frac{1}{2}(-Np\ln(2\pi) - N\ln\left|\hat{\Sigma}\right| - Np),$$

e per il modello completo, nel quale $\hat{\Sigma} = S$, vale dunque

$$\hat{l}_f = l(\hat{\mu}, S) = \frac{1}{2}(-Nq\ln(2\pi) - N\ln\left|S\right| - Np).$$

Dunque, la devianza di un generico modello rispetto a quello completo sarà data da

$$G^2 = 2(\hat{l}_f - \hat{l}_m) = N\ln(\left|\hat{\Sigma}\right| / \left|S\right|),$$

e in generale quella tra due modelli annidati, $M_0 \subseteq M_1$, varrà

$$d = 2(\hat{l}_1 - \hat{l}_0) = N\ln(\left|\hat{\Sigma}_0\right| / \left|\hat{\Sigma}_1\right|)$$

dove $\hat{\Sigma}_0$ e $\hat{\Sigma}_1$ rappresentano, rispettivamente, le stime di Σ sotto i modelli M_0 e M_1. Sotto il modello ridotto (ovvero sotto l'ipotesi nulla che i parametri presenti solo nel modello più

complesso siano uguali a zero), la devianza avrà distribuzione asintotica χ_d^2, con gradi di libertà *d* pari alla differenza nel numero di parametri tra i due modelli.

Dunque, per confrontare due modelli grafici riferiti allo stesso insieme di nodi ma a due insiemi di archi differenti (uno sottoinsieme dell'altro), sarà sufficiente calcolare la devianza tra di essi (basandosi le stime delle rispettive matrici di covarianza), e confrontarla con un χ_d^2 con gradi di libertà *d*, pari al numero di archi presenti nel più complesso, ma non nel meno complesso tra i due modelli.

Per campioni particolarmente piccoli, per i quali non sia possibile applicare il test asintotico, sono in alcuni casi disponibili dei test esatti. In particolare, se i due modelli differiscono solamente per un arco, è possibile utilizzare il test

$$F = \frac{(e^{d/n} - 1)}{N - k}$$

dove *k* è il numero di vertici presenti nella *clique* che contiene il vertice di interesse nel modello più complesso. Sotto H_0, questa quantità ha una distribuzione di tipo *F*, con *1* grado di libertà al numeratore e *N-1* al denominatore (Edwards, 2000).

E' interessante notare la connessione tra la devianza *d* associata alla rimozione di un arco – poniamo sia quello tra X_i e X_j – e il coefficiente di correlazione parziale tra le due variabili all'arco associate, date tutte le altre, così come mostrato da Whittaker (1990):

$$d = -N \ln\left\{1 - (\hat{\rho}_{X_i X_j \cdot X_{rest}})^2\right\},$$

dove $X_{rest} = V \setminus \{X_i, X_j\}$. Sfruttando questo risultato, si può adattare un modello a un determinato insieme di dati tramite un criterio di selezione per passi (a partire dal modello saturo o da quello con la sola intercetta), valutando ad ogni passo, tramite l'opportuno test χ^2, la significatività della riduzione (o incremento) della devianza conseguente all'eliminazione (inserimento) di una variabile nel modello.

3.3. Combinazione di GGM

Sono state esplicitate nel capitolo introduttivo le motivazioni per uno studio delle tecniche per combinare modelli grafici differenti. Si vogliono ora approfondire i meccanismi tecnici e le eventuali condizioni che rendono la combinazione possibile nei diversi casi di interesse.

E' necessario per prima cosa distinguere il concetto di *grafo* (che come visto rappresenta la struttura di indipendenza condizionata del modello), da quello di *famiglia di distribuzioni* (che si riferisce all'insieme di un grafo e delle distribuzioni di probabilità con esso compatibili). Per combinare due grafi c'è bisogno di assumere che le famiglie ad essi relative siano in qualche senso *consistenti* (Dawid e Lauritzen, 1993), ovvero che la legge di probabilità che regola le variabili in comune tra i due grafi sia "coerente" nei due casi. Per definire questo concetto di coerenza, ci si basa sul concetto di *meta-consistenza* (Dawid e Lauritzen, 1993) se la distribuzione sulle variabili in comune è la stessa per le due famiglie – come succede ad esempio per due distribuzioni Normali bivariate di stessa media e rispettive matrici di covarianza Σ_1 e Σ_2 – e su quello (meno restrittivo) di *quasi-consistenza* (Massa, 2008), secondo cui la distribuzione sulle variabili in comune è la stessa in almeno un caso, ovvero le assunzioni dei due studi non sono identiche, ma non entrano in conflitto. Sono esempio di questa ultima condizione due distribuzioni Normali bivariate di stessa media, e con matrici di covarianza rispettive: una generica Σ_1 per la prima distribuzione, e una matrice fissata, ad esempio la matrice identità I_2, per la seconda: le distribuzioni non sono identiche, ma coincidono nel caso in cui gli elementi diagonali di Σ_1 siano pari a 1.

Solo nel caso in cui sia soddisfatta almeno la condizione di quasi-consistenza è possibile trovare una famiglia congiunta di distribuzioni tra i due modelli: in tutti gli altri casi occorre concludere che una combinazione di tipo grafico non è possibile. La combinazione può essere effettuata tramite una *meta-combinazione di Markov* (Dawid e Lauritzen, 1993), la quale combina tutte le coppie di distribuzioni consistenti, mantenendo le famiglie marginali originali se applicata a famiglie meta-consistenti, o riducendole (le marginali derivate dalla famiglia congiunta sono contenute nelle marginali originali) se applicata a famiglie quasi-consistenti. In alternativa è possibile utilizzare una *quasi-combinazione di Markov* (Massa, 2008), che combina le distribuzioni delle due famiglie mantenendo una delle due marginali e modificando l'altra (e dunque generalmente estendendo le famiglie marginali, dato che non associa solo le distribuzioni consistenti). La scelta tra le due possibilità dipende dalle informazioni di cui in principio si dispone riguardo alle famiglie iniziali. Se ci si trova nel caso in cui le famiglie sono meta-consistenti, le due combinazioni sono equivalenti.

Esistono ad ogni modo casi piuttosto semplici da trattare, come quelli in cui le famiglie da combinare sono una sottoinsieme dell'altra. Una famiglia definita su $G_1 = (V_1, E_1)$ si dice *sottoinsieme* di una definita su $G_2 = (V_2, E_2)$ se vale che $V_1 = V_2$ ed $E_1 \subseteq E_2$. Questo accade ad esempio tra le famiglie rappresentate nella Figura 3.3 (in cui la famiglia sulla destra è un sottoinsieme di quella sulla sinistra). In situazioni in cui le famiglie da combinare sono una sottoinsieme dell'altra, è semplice definire il grafo congiunto a partire da quelli marginali: esso è in generale $G_3 = (V_3, E_3)$, dove $V_3 = V_1 = V_2$ ed $E_3 = E_2$ nell'ipotesi che $E_1 \subseteq E_2$. Nell'esempio rappresentato, il grafo sulla sinistra è esso stesso rappresentazione del grafo congiunto.

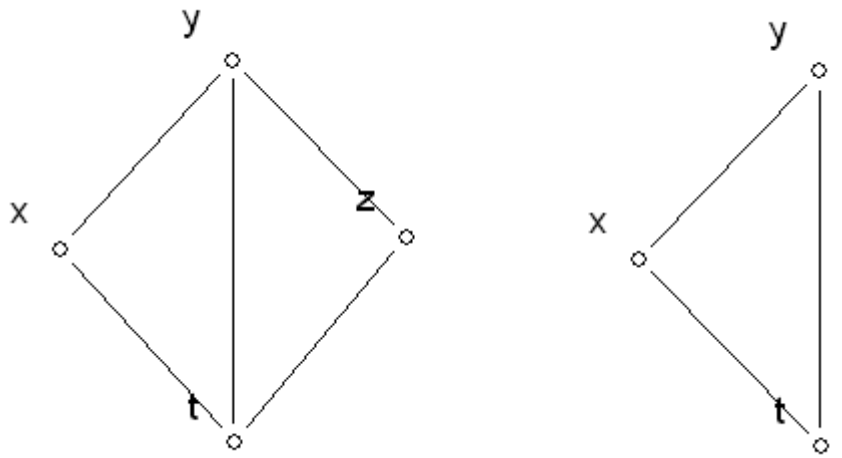

Figura 3.3: la famiglia definita sul grafo a destra è sottoinsieme della famiglia definita sul grafo a sinistra.

Situazioni meno banali – e più realistiche rispetto ai casi di applicazione – sono quelle in cui l'insieme dei vertici non è lo stesso per le due famiglie, bensì presenta solo alcuni elementi in comune. Un esempio di questa situazione lo si ha nella Figura 3.4:

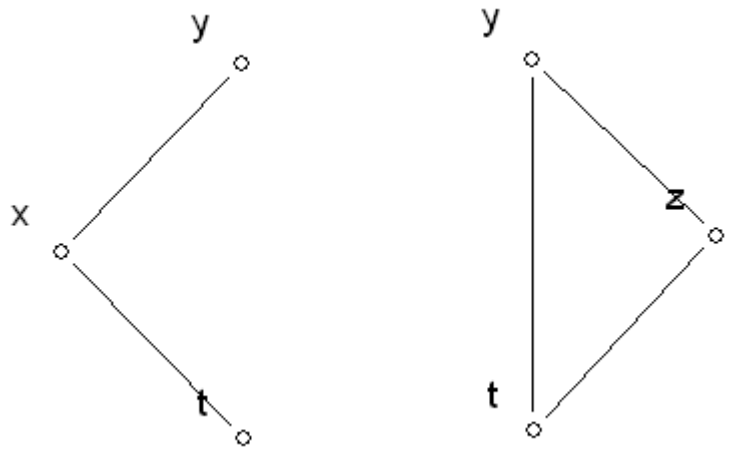

Figura 3.4: le due famiglie rappresentate sono definiti su insiemi di vertici differenti.

In questi casi la definizione del grafo congiunto non è immediata, e può portare ad errori nella definizione della struttura di indipendenza condizionata se effettuata in modo automatico aggregando i due modelli basandosi sulla rappresentazione grafica. Nel caso precedente, ad esempio, il grafo *//XYT YZT* potrebbe apparire una soluzione intuitiva per combinare i due grafi, ma si rivelerebbe una soluzione sbagliata: infatti, il grafo sulla sinistra in Figura 3.4 non ne rappresenterebbe una famiglia marginale.

Anche se in alcuni casi può accadere che una soluzione di tipo grafico non esista, nella maggior parte delle circostanze il problema sta nel fatto che il possibile modello congiunto non è unico. In questi casi, al metodo consistente nell'esaminare *tutte* le possibilità compatibili con i grafi marginali, si preferisce in questa tesi utilizzare l'approccio che tra tutti i grafi congiunti possibili privilegia quello dalla struttura più semplice.

3.4. Combinazione di GGM studiata tramite simulazione

Negli studi su dati provenienti da *microarray* sono frequenti situazioni in cui si dispone di risultati relativi a uno stesso argomento, ma condotti da laboratori diversi, che pertanto utilizzano insiemi di variabili solo parzialmente coincidenti: in questi casi può essere utile cercare un metodo per combinare i due risultati dei due studi (sfruttando le variabili in comune) piuttosto che condurre uno studio *ex novo* che includa tutte le variabili di interesse. In termini statistici, ci si domanda sotto quali condizioni sia possibile stimare un modello congiunto a partire da modelli marginali che hanno in comune tra loro solo una parte delle variabili, e come effettuare questa operazione combinando l'informazione fornita dalle fonti marginali nel miglior modo possibile.

Questa classe di problemi può essere affrontata attraverso studi di simulazione, che riproducano nel modo più fedele possibile la struttura reale del problema. Ad esempio, può essere utilizzato il seguente algoritmo:

1. si simulano N osservazioni p-variate da un certo GGM M_0, ottenendo un campione C_0 (che rappresenta il risultato dello studio complessivo);
2. si ripartiscono in modo casuale le osservazioni in due sottoinsiemi di numerosità N_1 e N_2, tali che $N_1 + N_2 = N$, così da ottenere due sotto-campioni C_1 e C_2 (che riproducono i risultati dei due diversi esperimenti). Inoltre si decompone il modello di partenza in due sotto-modelli M_1 e M_2, non necessariamente completi ma tali che M_0 rappresenti un loro possibile modello congiunto;
3. si effettua una regressione a passi sui dati campionari, calcolando la frequenza con cui il risultato ottenuto è il modello M_0 originario. Questo viene effettuato nelle due situazioni:
    3.1. a partire dal campione completo C_0 (il contatore relativo a M_0 rappresenta le prestazioni di uno studio che include tutte le variabili);
    3.2. a partire dai sotto-campioni C_1 e C_2 (effettuando due regressioni parallele) e contando quante volte i risultati delle regressioni sono contemporaneamente proprio i modelli M_1 e M_2, ovvero il modello congiunto è M_0 (questo contatore rappresenta le prestazioni della combinazione di due studi che hanno in comune una parte delle variabili);
4. si confrontano i valori dei contatori ottenuti nelle due situazioni e si determina quale tra i due metodi sia preferibile, ed eventualmente sotto quali condizioni.

Molti sono i parametri in gioco in studi di questo tipo, a partire dalla scelta dei modelli in gioco fino ad arrivare alle caratteristiche delle tecniche di regressione utilizzate (in avanti o all'indietro, con eventuali restrizioni ed eventuali penalizzazioni), e alle possibili soluzioni per i problemi computazionali che si incontrano nel corso dell'algoritmo: ad esempio, esistono tecniche diverse (e di efficacia diversa) per costruire una matrice di covarianza, da cui simulare i dati, che sia in grado di soddisfare i vincoli di interesse. Questa flessibilità dell'algoritmo rende naturale l'idea di applicarlo inizialmente in situazioni più "protette" (modelli composti da un insieme limitato di variabili, numerosità bilanciate tra i sotto-campioni e così via), per poi allentare via via i vincoli e cercare di estendere i risultati ad una situazione il più generale possibile.

3.5. Un esempio di studio di simulazione

Un recente studio di simulazione (Massa, 2008) ha applicato l'algoritmo di cui sopra al modello grafico congiunto *G=//XYT YZT*, prendendo come grafi marginali G_1*=//XYT* e G_2 *=//YZT*. I network generati da questi modelli sono rappresentati nella Figura 3.5.

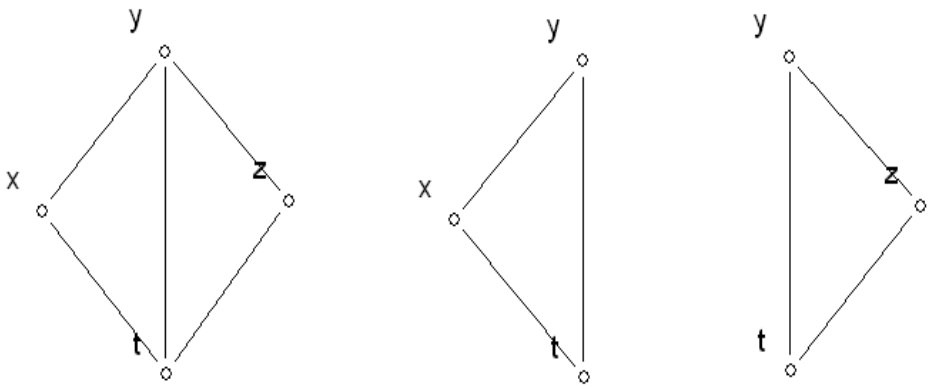

Figura 3.5: da sinistra a destra, i modelli G, G_1 e G_2, così come definiti nel testo.

Lo studio, basato su 10000 ripetizioni dell'algoritmo, impiega un campione di base di numerosità variabile (da *N*=40 a *N*=160) e sotto-campioni di numerosità ogni volta bilanciata ($N_1=N_2=N/2$); la tecnica di regressione utilizzata è di tipo a passi all'indietro. Per la generazione della matrice di covarianza si è utilizzato l'algoritmo IPS, il quale a partire da una certa matrice di base Σ_0 e da una *matrice di adiacenza* (che indica in quali posizioni è richiesto uno zero nella matrice di precisione), stima in modo iterativo la matrice $\hat{\Sigma}$. Per la scelta della matrice di partenza esistono diverse soluzioni, che possono portare a risultati anche piuttosto diversi. Nello studio è stata utilizzata una matrice del tipo $\Sigma_0 = \sigma \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$, dove σ=2 e ρ=0.6. L'algoritmo risultante – applicato utilizzando i software `R 2.6.0` e il programma per l'analisi di modelli grafici `MIM 3.2.0.6` – è dunque il seguente:

1. a partire da Σ_0, definita come sopra, si genera una matrice di covarianza $\hat{\Sigma}$ coerente con la struttura di *G*, tramite l'algoritmo IPS;
2. si genera un campione di *N=100* osservazioni pseudo-casuali da $Y \sim N_4(0, \hat{\Sigma})$;
3. il campione ottenuto è suddiviso pseudo-casualmente in due sotto-campioni di numerosità $N_1=50$ e $N_2=50$;
4. si effettuano, separatamente per il campione complessivo e per i due sotto-campioni, regressioni di tipo a passi all'indietro, nel primo caso relativamente all'insieme di variabili $V=\{X,Y,Z,T\}$, nel secondo all'insieme $V_1=\{X,Y,T\}$ e nel terzo all'insieme $V_2=\{Y,Z,T\}$;
5. si ripetono (10000 volte) i passi dal *2* al *4* e si conta quante volte la procedura per passi dà come risultato – rispettivamente, nei tre casi – il modello *G* (contatore

cont1), il modello G_1 (*cont2*), il modello G_2 (*cont3*). Il numero di volte in cui sono stati trovati contemporaneamente il modello G_1 e il modello G_2 è misurato dal contatore *cont4*.

La Tabella 3.1 sintetizza i risultati ottenuti.

n	n_A	n_B	cont1	cont2	cont3	cont4
40	20	20	337	42	46	0
50	25	25	1097	445	455	27
60	30	30	2194	1305	1315	160
70	35	35	3367	2519	2587	620
80	40	40	4230	3852	3779	1479
90	45	45	5019	5027	4962	2476
100	50	50	5712	6032	6020	3601
120	60	60	6540	7606	7654	5809
140	70	70	7251	8552	8576	7314
160	80	80	7808	9189	9199	8462

Tabella 3.1: risultati dello studio di simulazione.

Si può notare inizialmente che, come era logico aspettarsi, al crescere della numerosità campionaria il vero modello generatore dei dati viene trovato sempre più spesso dalla regressione a passi (sia per il modello congiunto che per i modelli marginali). Per confrontare i comportamenti dei diversi contatori, invece, deve essere tenuto presente il fatto che solo risultati ottenuti con lo stesso numero di osservazioni sono comparabili: per una più semplice lettura si rimanda alla Tabella 3.2. A parità di osservazioni, la combinazione dei sotto-grafi trova il vero modello molto più spesso di quanto non faccia il modello congiunto. Questo risultato è incoraggiante: può significare che combinare i risultati provenienti da studi diversi sia più conveniente – nelle condizioni di questo studio – rispetto a condurre un nuovo studio che includa tutte le variabili di interesse. E'

necessario d'altra parte notare che gli alti valori di *cont4* possono essere dovuti al minor numero di variabili contenuto nei grafi marginali, che rende più facile ottenere il modello vero tramite la regressione a passi.

n	cont1	cont4
40	337	1479
50	1097	3601
60	2194	5809
70	3367	7314
80	4230	8462

Tabella 3.2: confronti tra il modello congiunto

e i modelli marginali a parità di numerosità.

I risultati visti possono rivelarsi utili in un contesto in cui $N < p$: giustificano infatti – senza che le prestazioni complessive ne risentano – una suddivisione delle variabili in tanti sotto-gruppi, tali da contenere ciascuno un numero di variabili inferiore alla numerosità campionaria, e, in un secondo momento, una combinazione dei risultati ottenuti per i singoli sotto-gruppi.

3.6. Ulteriori possibilità

I risultati ottenuti nello studio citato sono promettenti, e potenziali di essere estesi in molteplici direzioni, pur rimanendo nell'ambito dei GGM. L'obiettivo di questa tesi è proprio di esplorare alcune delle strade aperte da studi come quello appena visto, cercando di allentare alcune assunzioni in modo da generalizzare i risultati a casi più ampi. In particolare ci si propone di:

1. considerare modelli diversi, e leggermente più complessi di quello utilizzato nello studio citato;
2. utilizzare tecniche diverse per la simulazione di dati, in particolare per la costruzione della matrice di covarianza da cui essi vengono generati (applicando lo stesso algoritmo ma partendo da una matrice di correlazione casuale invece che scelta arbitrariamente);
3. esplorare i casi in cui le numerosità dei sotto-campioni non siano bilanciate.

Quelli elencati rappresentano naturalmente solo alcuni dei passi possibili, in un'ottica di alleggerimento sempre maggiore delle ipotesi sotto le quali la combinazione di GGM possa risultare conveniente rispetto ad intraprendere studi *ex novo*. La direzione complessiva rimane quella di estendersi a modelli sempre più articolati, a tecniche di simulazione più generali, a modalità di campionamento più complesse. Nel prossimo capitolo saranno illustrati i risultati di due studi di simulazione originali, che perseguono gli obiettivi appena elencati.

4. Studi di simulazione

4.1. Scelte effettuate

L'algoritmo degli studi di simulazione descritti nel seguito segue da vicino la struttura di quello visto nel precedente capitolo. Il primo studio affronta lo stesso modello congiunto *//XYT YZT*, mentre il secondo si occupa del più complesso modello *//XY YZ ZT TX ZW ZV*. Per costruire la matrice di covarianza da cui generare i dati si utilizza come Σ_0 una matrice di correlazione casuale (costruita tramite il comando `rcorr` della libreria `ggm` di `R`), moltiplicata per un fattore $\sigma > 0$ (solitamente $\sigma = 2$). Alla matrice così generata si applica l'algoritmo IPS (tramite il comando `fitConGraph`, sempre nella libreria `ggm` di `R`), e sulla matrice $\hat{\Sigma}$ così ottenuta si svolgono i seguenti controlli:

- che i suoi autovalori siano non nulli (per motivi numerici, si richiedono maggiori di 0.05), così da assicurarsi che essa sia definita positiva;
- che sia, sempre per motivi numerici, perfettamente simmetrica, e provvedendo a mediarla con la sua trasposta nel caso di piccole incongruenze;
- che i valori della sua inversa (matrice di precisione), nelle posizioni dove *non* è previsto un valore nullo, siano superiori a un certo ε (pari a 0.5 nel primo studio e a 0.1 nel secondo), affinché, sempre per motivi numerici, non siano confusi dal programma con degli zeri.

I dati sono simulati attraverso il comando `rmvnorm` (appartenente alla libreria `mvtnorm`), e la regressione è effettuata tramite il comando `stepwise` della libreria `mimR` (che interfaccia `R` con il software `MIM` per l'analisi di modelli grafici). Una delle differenze fondamentali rispetto allo studio citato nel precedente capitolo sta nel fatto che i sotto-campioni non sono vincolati ad avere la stessa numerosità.

4.2. Studio 1

Il modello utilizzato per il primo studio di simulazione è quello generato dalle *cliques* *XYT* e *YZT*, e i sotto-grafi presi in considerazione sono quelli associati ai due singoli generatori. La Figura 4.1 mostra sulla sinistra il grafo di partenza, e al centro e a destra i due sotto-grafi considerati.

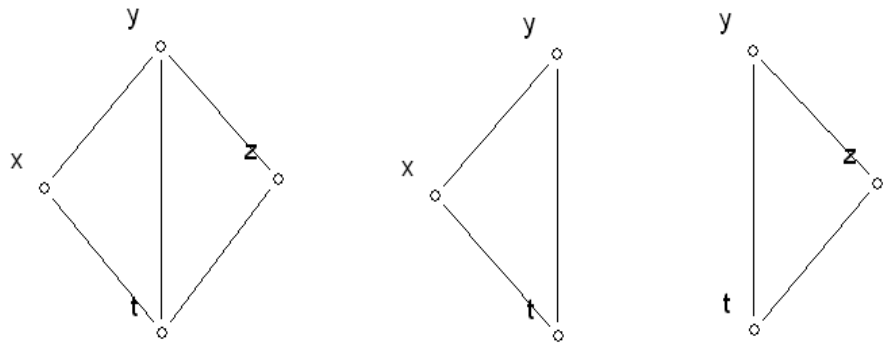

Figura 4.1: modello congiunto (a sinistra) e modelli marginali (al centro e a destra) per il primo studio di simulazione.

Tramite l'algoritmo IPS, basato su una matrice di correlazione casuale moltiplicata per σ=2, sono state ottenute tre differenti matrici di covarianza da cui simulare le osservazioni. Queste sono riportate, con le relative matrici inverse, nella Tabella 4.1:

matrice di covarianza (A)					**matrice di precisione (A)**				
	x	**y**	**z**	**t**		**x**	**y**	**z**	**t**
x	2.0000	-1.7807	-0.9121	1.3406	**x**	3.4827	4.7702	0.0000	1.8864
y	-1.7807	2.0000	0.6424	-1.7697	**y**	4.7702	11.7329	-1.8082	7.2382
z	-0.9121	0.6424	2.0000	0.0593	**z**	0.0000	-1.8082	1.1293	-1.6335
t	1.3406	-1.7697	0.0593	2.0000	**t**	1.8864	7.2382	-1.6335	5.6888

matrice di covarianza (B)					**matrice di precisione (B)**				
	x	**y**	**z**	**t**		**x**	**y**	**z**	**t**
x	2.0000	1.4445	0.0032	1.1479	**x**	3.4406	-2.4936	0.0000	-1.9857
y	1.4445	2.0000	0.9234	-0.0087	**y**	-2.4936	2.5383	-0.5033	1.1519
z	0.0032	0.9234	2.0000	-1.1540	**z**	0.0000	-0.5033	1.0960	0.6302
t	1.1479	-0.0087	-1.1540	2.0000	**t**	-1.9857	1.1519	0.6302	2.0084

matrice di covarianza (C)					**matrice di precisione (C)**				
	x	**y**	**z**	**t**		**x**	**y**	**z**	**t**
x	2.0000	-1.8689	-1.0958	-0.3287	**x**	5.7483	5.4146	0.0000	1.1484
y	-1.8689	2.0000	0.9283	-0.0753	**y**	5.4146	5.8588	-0.5309	1.4044
z	-1.0958	0.9283	2.0000	1.1081	**z**	0.0000	-0.5309	1.0930	-0.6256
t	-0.3287	-0.0753	1.1081	2.0000	**t**	1.1484	1.4044	-0.6256	1.0882

Tabella 4.1: matrici di covarianza e di precisione utilizzate per generare le simulazioni (studio 1).

Si può verificare che le matrici di precisione presentano i valori nulli nella cella opportuna: quella che si riferisce alla relazione tra *X* e *Z*, i cui rispettivi nodi sono gli unici a non essere collegati da un arco. In generale, si può notare una certa variabilità nei valori delle tre matrici, il che permetterà una più ampia varietà di risultati. La Tabella 4.2 riporta i valori dei contatori riferiti a 10000 simulazioni per le tre matrici: *cont1* rappresenta il numero di volte che il modello //*XYT YZT* è stato trovato con la regressione a passi a partire dall'intero insieme di dati simulati (numerosità campionaria pari a *N*); *cont2* e *cont3* rappresentano, rispettivamente, il numero di volte che i modelli //*XYT* e //*YZT* sono stati trovati con la regressione a passi sui sotto-campioni dei dati simulati (numerosità rispettive pari a *N1* e *N2*); *cont4* rappresenta il numero di volte che i //*XYT* e //*YZT* sono stati trovati contemporaneamente per la stessa simulazione, il che implica che il modello //*XYT YZT*, che ne rappresenta la combinazione grafica, è stato individuato.

			matrice A			
N	**N1**	**N2**	**cont1**	**cont2**	**cont3**	**cont4**
100	50	50	9405	9911	10000	9911
100	60	40	9474	9967	9996	9963
100	70	30	9425	9993	9977	9970
100	80	20	9422	9998	9651	9649

			matrice B			
N	**N1**	**N2**	**cont1**	**cont2**	**cont3**	**cont4**
100	50	50	9109	10000	7593	7593
100	60	40	9159	10000	6666	6666
100	70	30	9123	10000	5280	5280
100	80	20	9118	10000	3849	3849

			matrice C			
N	**N1**	**N2**	**cont1**	**cont2**	**cont3**	**cont4**
100	50	50	8315	9861	8447	8328
100	60	40	8299	9955	7620	7591
100	70	30	8372	9988	6191	6183
100	80	20	8301	9993	4557	4553

Tabella 4.2: esito delle simulazioni per le tre matrici (studio 1)

Il comportamento di *cont1*, per quanto sia difficile attribuire le differenze trovate a specifiche differenze tra le diverse matrici da cui i dati sono stati generati, è sostanzialmente "buono": il vero modello generatore dei dati viene trovato con una frequenza che va dall'83 al 95% (sul totale delle simulazioni). Questa frequenza è relativamente variabile tra i diversi casi ma essenzialmente stabile al loro interno (come era plausibile aspettarsi, rimanendo la numerosità campionaria la stessa). In *cont2* e *cont3*, già relativamente al caso bilanciato si nota una maggiore variabilità nei risultati (dovuta alle numerosità campionarie più basse: 50 simulazioni rispetto alle 100 del modello di partenza). Si sottolinea che i due sotto-modelli hanno un ruolo del tutto simmetrico, avendo struttura analoga e riferendosi alla stessa numerosità campionaria.

Al variare della numerosità dei sotto-campioni, le conseguenze sono simili nei tre casi considerati: come era ragionevole attendersi, al diminuire della numerosità, diminuisce anche il numero di volte che l'algoritmo è in grado di individuare il sotto-grafo in questione. Quello che differisce nei tre casi considerati è la "velocità" con cui questo calo avviene: quanto più basso è il valore iniziale del contatore (ovvero, quello relativo a un campione da 50 simulazioni), tanto più rapida sembra essere la caduta. Questo fenomeno può essere giustificato dal fatto che un basso valore iniziale è già esso indice di correlazioni parziali deboli tra le variabili. Concordemente con quanto detto, quando la numerosità cresce, anche il valore assunto dal contatore tende a crescere; è però difficile in questo caso dare una valutazione della "velocità" della crescita, in quanto i valori iniziali (quelli per una numerosità di 50 osservazioni) sono generalmente già molto vicini al massimo raggiungibile di 10000 già dal principio. Addirittura, in uno dei tre casi considerati, già con una numerosità di 50 il sotto-grafo *//XYT* è trovato nel 100% dei casi: in conformità con quanto ci si poteva aspettare, la frequenza rimane poi massima al

crescere della numerosità. E' il caso di specificare che questi andamenti si basano su osservazioni empiriche e non sottostanno ad alcun vincolo matematico: sarebbe senz'altro possibile incontrare situazioni in cui, al crescere della numerosità campionaria, il valore di almeno uno dei contatori diminuisce, o viceversa. L'andamento di *cont4*, infine – rappresentando l'intersezione tra due contatori di cui uno (*cont2*) si attesta quasi sempre su valori massimi – sarà strettamente legato (in alcuni casi coincidente) al comportamento dell'algoritmo per il modello riferito alla numerosità minore (ovvero, all'andamento di *cont3*).

Un'analisi delle matrici di correlazione parziale (derivate a partire dalle rispettive matrici di precisione e riportate nella Tabella 4.3) è utile ad approfondire i risultati. I valori di correlazione sono in media più alti per la matrice A, e minimi per la matrice C: questo è in accordo con i valori di *cont1*, che infatti decrescono (in media) passando dal caso A al caso B, e dal caso B al caso C. I valori di correlazione parziale maggiori (in modulo) relativamente ai legami che coinvolgono la variabile X (che appartiene al primo sotto-grafo, ma non al secondo) sono quelli trovati nel caso B: a questo fenomeno si può attribuire il fatto che i valori di *cont2* (che si riferisce al primo sotto-grafo) sono massimi proprio per il caso B. Allo stesso modo la matrice A, che è quella per cui i valori di *cont3* sono più elevati, mostra le correlazioni parziali più forti relativamente alla variabile Z (che appartiene al secondo sotto-grafo, ma non al primo).

matrice di correlazione parziale (A)				
	x	**y**	**z**	**t**
x	1,0000	-0,7465	0,0003	-0,4243
y	-0,7465	1,0000	0,4969	-0,8861
z	0,0003	0,4969	1,0000	0,6446
t	-0,4243	-0,8861	0,6446	1,0000

matrice di correlazione parziale (B)				
	x	**y**	**z**	**t**
x	1,0000	0,8440	-0,0001	0,7556
y	0,8440	1,0000	0,3016	-0,5107
z	-0,0001	0,3016	1,0000	-0,4245
t	0,7556	-0,5107	-0,4245	1,0000

matrice di correlazione parziale (C)				
	x	**y**	**z**	**t**
x	1,0000	-0,9332	0,0008	-0,4601
y	-0,9332	1,0000	0,2104	-0,5570
z	0,0008	0,2104	1,0000	0,5738
t	-0,4601	-0,5570	0,5738	1,0000

Tabella 4.3: matrici di correlazione parziale (studio 1).

4.3. Studio 2

Il secondo studio è leggermente più complesso del precedente. Il modello congiunto considerato è *//XY YZ ZT TX ZW ZV*, costituito da sei vertici e sei archi (contro i quattro vertici e cinque archi del precedente). Rispetto al primo studio, in cui i modelli marginali avevano struttura sostanzialmente simmetrica, in questo caso le differenze tra i due sono più marcate: il primo sotto-grafo è *//XY YZ ZT TX* (quattro nodi, quattro archi) e il secondo è *//ZW ZV* (tre nodi, due archi): hanno una sola variabile in comune e nessuno dei due è un sotto-modello completo. Una rappresentazione grafica dei tre modelli è mostrata nella Figura 4.2.

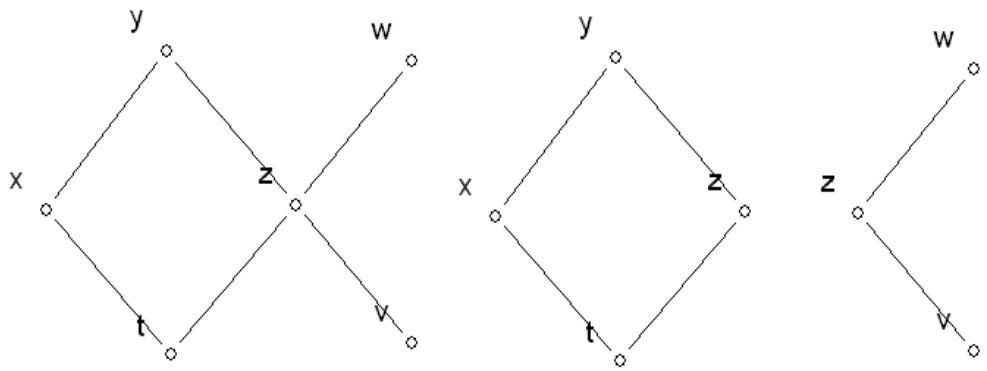

Figura 4.2: modello congiunto (a sinistra) e modelli marginali (al centro e a destra) per il secondo studio di simulazione.

In analogia con lo studio precedente, sono state costruite tre diverse matrici da cui generare i dati, al fine di studiare il comportamento delle simulazioni in situazioni differenti. Le matrici sono state costruite tramite l'algoritmo IPS, basato anche in questo caso su matrici di correlazione casuale moltiplicate per σ=2. Le matrici di covarianza – riportate nella Tabella 4.4 con le relative matrici di precisione – presentano una buona variabilità tra loro, e si può facilmente verificare che le matrici di precisione rispettano i vincoli dovuti all'assenza di alcuni archi, mostrando valori nulli nelle posizioni opportune.

matrice di covarianza (A)							matrice di precisione (A)						
	x	y	z	t	v	w		x	y	z	t	v	w
x	2.0000	0.8807	-0.4713	-1.0343	-0.2183	0.3713	x	0.8325	-0.3062	0.0000	0.3822	0.0000	0.0000
y	0.8807	2.0000	0.0984	-0.3159	0.0456	-0.0775	y	-0.3062	0.6399	-0.1036	0.0000	0.0000	0.0000
z	-0.4713	0.0984	2.0000	1.1055	0.9263	-1.5757	z	0.0000	-0.1036	1.6951	-0.4256	-0.2948	1.0385
t	-1.0343	-0.3159	1.1055	2.0000	0.5121	-0.8710	t	0.3822	0.0000	-0.4256	0.9329	0.0000	0.0000
v	-0.2183	0.0456	0.9263	0.5121	2.0000	-0.7298	v	0.0000	0.0000	-0.2948	0.0000	0.6366	0.0000
w	0.3713	-0.0775	-1.5757	-0.8710	-0.7298	2.0000	w	0.0000	0.0000	1.0385	0.0000	0.0000	1.3181

matrice di covarianza (B)							matrice di precisione (B)						
	x	y	z	t	v	w		x	y	z	t	v	w
x	2.0000	0.3337	-0.0200	-1.0340	0.0080	-0.0150	x	0.7056	-0.1096	0.0000	0.3624	0.0000	0.0000
y	0.3337	2.0000	0.8712	-0.0447	-0.3506	0.6511	y	-0.1096	0.6403	-0.2800	0.0000	0.0000	0.0000
z	-0.0200	0.8712	2.0000	0.3026	-0.8050	1.4948	z	0.0000	-0.2800	1.3670	-0.1027	0.2401	-0.8467
t	-1.0340	-0.0447	0.3026	2.0000	-0.1218	0.2262	t	0.3624	0.0000	-0.1027	0.7029	0.0000	0.0000
v	0.0080	-0.3506	-0.8050	-0.1218	2.0000	-0.6017	v	0.0000	0.0000	0.2401	0.0000	0.5967	0.0000
w	-0.0150	0.6511	1.4948	0.2262	-0.6017	2.0000	w	0.0000	0.0000	-0.8467	0.0000	0.0000	1.1328

matrice di covarianza (C)							matrice di precisione (C)						
	x	y	z	t	v	w		x	y	z	t	v	w
x	2.0000	1.9185	-0.3215	0.6645	-0.1832	-0.1944	x	6.4615	-6.0984	0.0000	-0.3361	0.0000	0.0000
y	1.9185	2.0000	-0.3912	0.5939	-0.2230	-0.2366	y	-6.0984	6.4031	0.2722	0.0000	0.0000	0.0000
z	-0.3215	-0.3912	2.0000	0.9179	1.1401	1.2095	z	0.0000	0.2722	1.2769	-0.42410	-0.4222	-0.4767
t	0.6645	0.5939	0.9179	2.0000	0.5233	0.5551	t	-0.3361	0.0000	-0.42410	0.8063	0.0000	0.0000
v	-0.1832	-0.2230	1.1401	0.5233	2.0000	0.6895	v	0.0000	0.0000	-0.4222	0.0000	0.7407	0.0000
w	-0.1944	-0.2366	1.2095	0.5551	0.6895	2.0000	w	0.0000	0.0000	-0.4767	0.0000	0.0000	0.7883

Tabella 4.4: matrici di covarianza e di precisione utilizzate per generare le simulazioni (studio 2).

Nella Tabella 4.5 sono riportati i risultati relativi a 10000 simulazioni, con l'usuale significato attribuito ai diversi contatori (*cont1* per il modello congiunto, *cont2* per il sotto-grafo *//XY YZ ZT TX*, *cont3* per il sotto-grafo e *//ZW ZV* e *cont4* per l'intersezione di *cont2* e *cont3*). Avendo i sotto-grafi una struttura tra loro differente, si è ritenuto opportuno

esaminare separatamente i casi in cui il campione di numerosità maggiore è relativo all'uno o all'altro modello marginale.

			matrice A			
N	**N1**	**N2**	**cont1**	**cont2**	**cont3**	**cont4**
100	20	80	1874	852	9062	772
100	30	70	1923	1501	8806	1308
100	40	60	1919	2060	8531	1761
100	50	50	1997	2568	8149	2086
100	60	40	1859	2964	7521	2228
100	70	30	1953	3470	6542	2284
100	80	20	1885	3773	5099	1924

			matrice B			
N	**N1**	**N2**	**cont1**	**cont2**	**cont3**	**cont4**
100	20	80	1274	265	8637	228
100	30	70	1253	497	8317	409
100	40	60	1256	765	7898	587
100	50	50	1281	1018	7382	753
100	60	40	1284	1359	6679	920
100	70	30	1192	1635	5638	911
100	80	20	1234	1958	4177	823

			matrice C			
N	**N1**	**N2**	**cont1**	**cont2**	**cont3**	**cont4**
100	20	80	2244	1145	9417	1078
100	30	70	2325	1919	9399	1788
100	40	60	2256	2594	9365	2441
100	50	50	2261	3158	9278	2927
100	60	40	2216	3756	8987	3364
100	70	30	2304	3994	8281	3309
100	80	20	2244	4341	6362	2744

Tabella 4.5: esito delle simulazioni per le tre matrici (studio 2).

I risultati possono essere analizzati separatamente rispetto ai diversi contatori e rispetto alle diverse relazioni tra di essi:

1. il valore di *cont1*, così come accadeva nello studio precedente, subisce oscillazioni lievi al ripetersi dell'esperimento con una stessa matrice, mentre subisce dei cambiamenti piuttosto evidenti (in media) al cambiare della matrice da cui i dati sono simulati. Il modello congiunto è trovato con una frequenza che va da circa 1200/10000 (per la matrice B) a circa 2300/10000 (per la matrice C). In generale, i valori trovati non sono indice di buone prestazioni: la capacità di ricostruire il modello stesso da cui i dati sono stati simulati appare nettamente inferiore per il modello *//XY YZ ZT TX ZW ZV* di quanto non apparisse per il modello *//XYT YZT* utilizzato nel primo studio (che era trovato con frequenze tra l'80 e il 90%). Questo risultato è ancora più significativo se si tiene conto del fatto che tra i due modelli c'è una differenza di due sole variabili, e di un solo arco. Si potrebbe, a questo punto, avanzare l'ipotesi generale che, al crescere del numero di geni in un network, la ricostruzione della sua struttura a partire dalle osservazioni sia sempre più difficile, anche per incrementi di complessità molto piccoli;

2. i comportamenti di *cont2* e di *cont3* presentano una somiglianza nel fatto di dipendere dalle numerosità a cui i modelli marginali fanno riferimento: il modello atteso – come ci si poteva aspettare dati anche i risultati dello studio precedente – è individuato più spesso quando le numerosità sono maggiori. I valori di *cont2* e *cont3* appaiono anche legati al valore di *cont1* (ovvero, in realtà, alla matrice da cui i dati sono simulati). Il sotto-grafo *//ZW ZV* è trovato con maggiore facilità rispetto al sotto-grafo *//XY YZ ZT TX* in tutti i casi, arrivando a superare la frequenza del 90% (di volte in cui viene individuato sul totale delle simulazioni) in due dei tre casi considerati, quando la numerosità è

sufficientemente alta, e calando con velocità diverse al calare della numerosità relativamente alle diverse matrici (per *N2*=20, nel caso B *cont3* supera di poco i 4000, mentre nel caso C rimane superiore ai 6000). Il sotto-grafo *//XY YZ ZT TX* è trovato molto più faticosamente dall'algoritmo: nel caso di numerosità massima di 80, presenta valori "migliori" rispetto a *cont1*, ma comunque insoddisfacenti come valore assoluto (nel caso B, il modello viene trovato meno del 20% delle volte anche quando la numerosità è pari a 80). Al decrescere della numerosità – e tenendo presente che casi di modelli di 4 variabili per 20 unità campionarie non sono infrequenti nelle applicazioni reali – i valori di *cont2* precipitano, fino a un minimo di 265/10000 riscontrato nel caso di *N1*=20, per la matrice B;

3. il comportamento di *cont4* dipende dagli andamenti di *cont2* e *cont3*, ma mostra particolarità diverse nei diversi casi considerati. Ad esempio, nella situazione di sotto-numerosità bilanciate, *cont4* assume valori "migliori" di *cont1* nei casi A e C, mentre non si discosta da valori minimi (inferiori al 10% per qualsiasi combinazione di numerosità) nel caso B, apparendo pertanto sempre meno affidabile di *cont1* rispetto a questa matrice di simulazione. Si ricorda che il confronto tra *cont1* e *cont4* indica se sia preferibile, nella ricostruzione del network, uno studio effettuato su 100 unità campionarie contemporaneamente, oppure due diversi studi marginali, effettuati sulle rispettive sotto-numerosità e poi combinati. Per tutte e tre le matrici considerate, è possibile notare che i valori di *cont4* vanno sostanzialmente migliorando al crescere di *N2* (nonostante parallelamente *N1* decresca), arrivando ad essere leggermente "migliori" di quelli di *cont1* nei casi A e C quando *N2* è massima. In pratica, *cont4* assume

valori maggiori quando la numerosità più alta corrisponde al modello marginale più complesso: infatti, anche per numerosità minime, il modello *//ZW ZV* viene trovato in almeno il 40% dei casi, mentre la frequenza con cui viene individuato il modello *//XY YZ ZT TX* , quando *N1* scende sotto una certa soglia, arriva anche sotto il 10%. Si può concludere che la ricostruzione del modello funziona meglio quando ci si trova di fronte ad una situazione più "equilibrata", in cui al modello più complesso corrisponde una numerosità campionaria più elevata, anche se per quello più semplice ci si "accontenta" di un campione più ristretto. Come ultima osservazione, si può notare che la crescita di *cont4* al crescere di *N2* subisce un'inversione di tendenza per *N2*=20 in tutti e tre i casi considerati: a quel punto è infatti il valore di *cont3* a subire un calo netto (per quanto il modello sia semplice, se la numerosità diventa troppo bassa esso stenta ad essere individuato), ed evidentemente questa caduta trascina con sé il valore di *cont4*.

La Tabella 4.6 riporta i valori di correlazione parziale ricavati a partire dalle matrici di precisione. Si tenta, in analogia con il primo studio, di trovare una connessione tra i valori di correlazione parziale e la frequenza con cui i diversi modelli sono individuati tramite la regressione a passi. Avendo i grafi considerati una struttura più complessa rispetto allo studio precedente, è più difficile trovare legami altrettanto evidenti. Sembra però, con riferimento al modello */XY YZ ZT TX*, che il valore di correlazione parziale particolarmente alto tra *X* e *Y* riscontrato rispetto alla matrice C sia da solo sufficiente a migliorare i risultati (infatti le altre correlazioni parziali sono sostanzialmente simili nei diversi casi, e questo è l'unico valore che subisce un cambio netto). Con riferimento al modello *//ZW ZV*, la connessione con la correlazione parziale è meno chiara, anche se si può verificare che i

casi A e B, che mostrano valori simili di correlazione parziale, danno anche luogo a valori simili dei relativi contatori.

matrice di correlazione parziale (A)						
	x	**y**	**z**	**t**	**v**	**w**
x	1.0000	0.4195	0.0000	-0.4337	0.0000	0.0000
y	0.4195	1.0000	0.0995	0.0000	0.0000	0.0000
z	0.0000	0.0995	1.0000	0.3384	0.2838	-0.6948
t	-0.4337	0.0000	0.3384	1.0000	0.0000	0.0000
v	0.0000	0.0000	0.2838	0.0000	1.0000	0.0000
w	0.0000	0.0000	-0.6948	0.0000	0.0000	1.0000

matrice di correlazione parziale (B)						
	x	**y**	**z**	**t**	**v**	**w**
x	1.0000	0.1631	0.0000	-0.5146	0.0000	0.0000
y	0.1631	1.0000	0.2993	0.0000	0.0000	0.0000
z	0.0000	0.2993	1.0000	0.1048	-0.2658	0.6804
t	-0.5146	0.0000	0.1048	1.0000	0.0000	0.0000
v	0.0000	0.0000	-0.2658	0.0000	1.0000	0.0000
w	0.0000	0.0000	0.6804	0.0000	0.0000	1.0000

matrice di correlazione parziale (C)						
	x	**y**	**z**	**t**	**v**	**w**
x	1.0000	0.9481	0.0000	0.1472	0.0000	0.0000
y	0.9481	1.0000	-0.0952	0.0000	0.0000	0.0000
z	0.0000	-0.0952	1.0000	0.4180	0.4341	0.4751
t	0.1472	0.0000	0.4180	1.0000	0.0000	0.0000
v	0.0000	0.0000	0.4341	0.0000	1.0000	0.0000
w	0.0000	0.0000	0.4751	0.0000	0.0000	1.0000

Tabella 4.6: matrici di correlazione parziale (studio 2).

La Tabella 4.7 riporta un'ulteriore analisi, in cui si approfondisce il ruolo della numerosità campionaria: si è ripetuta l'analisi effettuata sulla matrice B, ma con riferimento a un campione di 1000 unità invece che di 100, e sotto-campioni di numerosità proporzionali a quelle citate in precedenza (il caso bilanciato è $N1=N2=500$).

			matrice B			
N	**N1**	**N2**	**cont1**	**cont2**	**cont3**	**cont4**
1000	200	800	4935	5785	9483	5474
1000	300	700	4979	7467	9504	7081
1000	400	600	5073	8598	9528	8177
1000	500	500	4939	8667	9496	8227
1000	600	400	4984	8813	9486	8362
1000	700	300	4936	8843	9474	8372
1000	800	200	4955	8764	9470	8301

Tabella 4.7: esito di simulazioni su una diversa numerosità campionaria (studio 2, caso B).

Sono evidenti netti progressi rispetto al caso con numerosità campionaria pari a 100: in particolare, i due grafi marginali arrivano ad essere trovati contemporaneamente in oltre l'80% dei casi, anche in alcune delle situazioni in cui le sotto-numerosità sono maggiormente sbilanciate. In particolare, i miglioramenti più evidenti li mostra il sotto-grafo *//XY YZ ZT TX*, che nelle precedenti analisi mostrava i risultati meno soddisfacenti. Per quanto appaiano piuttosto migliorati anche i valori relativi a *cont1*, invece, essi rimangono ben lontani dall'essere accettabili in senso assoluto: il modello congiunto viene individuato (in media) solo nel 50% dei casi: questo vuol dire che un campione di 1000 unità, generato dal modello *//XY YZ ZT TX ZW ZV*, in un caso su due non è in grado di ricostruire il modello stesso da cui proviene, tramite le tecniche in analisi. Questo risultato invita a una particolare cautela nell'applicazione dei modelli grafici a insiemi di geni composti da un numero di variabili già pari a cinque o sei, a maggior ragione se si tiene conto del fatto che in applicazioni reali le numerosità campionarie rientrano più facilmente nell'ordine di grandezza delle decine che in quello delle centinaia.

5. Studio di dati reali

5.1. Uno studio sul rabdomiosarcoma

I dati utilizzati in questa analisi sono tratti da un ampio studio riguardante il rabdomiosarcoma, un tumore maligno della muscolatura striata che colpisce prevalentemente bambini e adolescenti. Dallo studio, riguardante circa 1200 geni e una numerosità campionaria di 138 unità, è stato estratto, per le analisi di interesse, un piccolo insieme di 28 geni, mentre è stato mantenuto il riferimento all'intero campione. Se questa selezione iniziale può apparire drastica (e naturalmente ridimensiona la rilevanza biologica dei risultati ottenuti), si anticipa che una volta effettuate le prime analisi descrittive il numero di variabili con cui ci si è trovati a trattare è apparso persino troppo elevato rispetto alla complessità del network sottostante, e ci si è pertanto, in fase di selezione del

modello, ricondotti a ulteriori sotto-campioni. D'altra parte, si è visto nei precedenti capitoli come la complessità dei legami biologici comporti difficoltà nel ricostruirne la struttura già per network composti da cinque o sei variabili: si è ritenuto dunque opportuno cercare di applicare con più precisione i metodi e i risultati proposti nei precedenti capitoli a un piccolo numero di variabili, piuttosto che tentare la difficile impresa di adattare un modello a un network molto esteso, con dubbi risultati dal punto di vista delle prestazioni.

5.2. Analisi descrittive

Una prima analisi di tipo descrittivo consiste nella costruzione delle matrici di correlazione e di correlazione parziale, riportate per intero nell'Appendice B (i geni sono identificati da sigle convenzionali). Per quanto riguarda la correlazione semplice, i valori più alti (che da ora in avanti si intenderanno sempre in modulo) non superano il valore di 0.7; il 50% dei valori non supera lo 0.13, e il 90% non supera lo 0.42. Le coppie di geni che appaiono più strettamente legati sono X4085 e X4089, X4607 e X4735, X4089 e X4735, X4887 e X4735, tutte con valori di correlazione compresi tra 0.66 e 0.7. Basandosi su queste misure si può avere una prima idea di come sia strutturato il network (Figura 5.1), tenendo conto delle dovute avvertenze:

1. per quanto appaia graficamente analogo, quello riportato non è assolutamente da interpretarsi come un modello grafico, dato che non costituisce un modello stimato sulla base di una regressione, bensì una semplice rappresentazione dei massimi valori di correlazione semplice riscontrati nel campione;

2. la figura è costruita sulla base di correlazioni marginali e non di correlazioni parziali (condizionate). Si è ritenuto tuttavia utile proporla lo stesso, proprio per analizzare, nel seguito, le differenze con quella costruita sulla base di queste ultime.

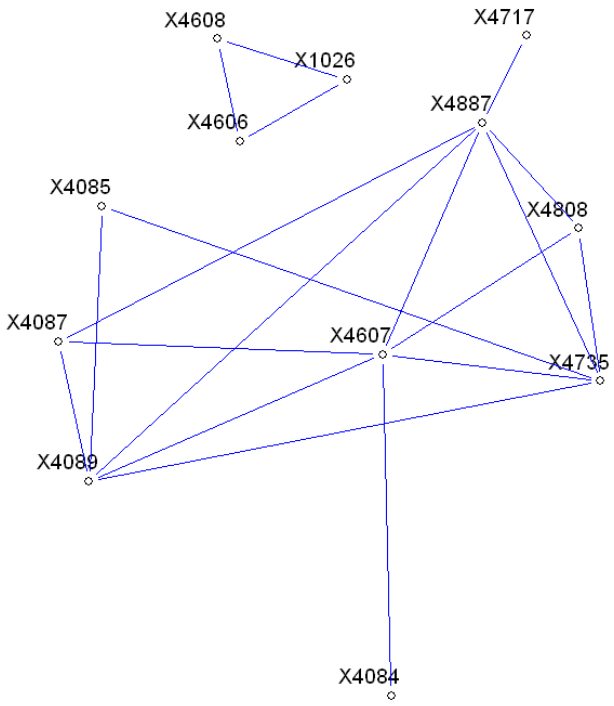

Figura 5.1: rappresentazione del 5% delle correlazioni semplici massime (superiori in modulo a 0.41).

Nella Figura 5.1, sono considerate "legate" le variabili con un valore di correlazione superiore in modulo a 0.41 (che rappresentano il 5% dei legami complessivi). I geni che sembrano avere con un maggior numero di legami sono X4887, X4089, X4735, X4607. Tre geni (X1026, X4608, X4606) appaiono correlati fortemente tra loro, ma non rispetto agli altri considerati.

E' senza dubbio più significativa la parallela analisi effettuata sulla matrice di correlazione parziale, dalla quale emergono sostanziali differenze rispetto ai risultati precedenti. Il valore di correlazione parziale più alto trovato vale in modulo 0.54; il 50%

dei valori (in modulo) è inferiore allo 0.10 e il 90% allo 0.22. Queste osservazioni sono coerenti con il comportamento generale della correlazione parziale, che non essendo influenzata dall'insieme delle altre variabili del network si discosta da zero più difficilmente, e pertanto indica con maggior precisione quali sono i legami "veri" tra le variabili. La matrice completa è riportata nell'Appendice B, e mostra che le coppie di geni maggiormente legati sono X4087 e X4089, X4606 e X4608, X4610 e X4885, e infine (con il valore in modulo massimo di -0.54) X3720 e X4609. Dunque, alcuni dei legami marginali più forti rimangono tali anche condizionandosi al comportamento delle altre variabili nel network; altri invece (come ad esempio quelli che coinvolgono il gene X4735) si rivelano almeno in parte di natura spuria, dovuti alla mediazione di altri geni. Tenendo sempre presente che si tratta di analisi descrittive, e che dunque la figura non rappresenta un modello grafico, si riporta nella Figura 5.2 il network costruito sulla base del 5% dei legami più forti (ovvero delle correlazioni parziali superiori in modulo a 0.27).

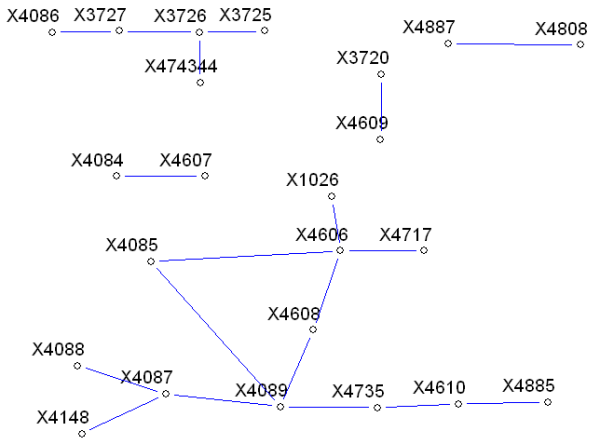

Figura 5.2: rappresentazione del 5% delle correlazioni parziali massime (superiori in modulo a 0.27).

La figura relativa alla correlazione parziale lascia intravedere scenari più complessi (e più realistici) di quelli a cui ci si sarebbe trovati di fronte limitandosi a un'analisi della correlazione semplice: le variabili tra loro "altamente" legate secondo la correlazione parziale costituiscono un insieme più numeroso, e danno l'idea di un network più esteso e composito, probabilmente rappresentando in modo più preciso la realtà. I geni indicati con X4089 e X4085 si confermano centrali alla struttura, mentre il gruppo composto da X4606, X4608 e X1026, che sembrava slegato dal resto dei geni, appare qui inserito nel "blocco" principale di relazioni.

5.3. Selezione di modelli

Si può a questo punto passare a una selezione per passi del modello, analoga a quella usata negli studi di simulazione proposti nel capitolo precedente. Il risultato mostrato nella Figura 5.3 (che rappresenta in questo caso un "vero" modello grafico), però, è di ardua interpretazione. Molti geni si rivelano legati a tutti o quasi tutti gli altri, rendendo impossibile cogliere una struttura semplice, composta dai soli legami principali.

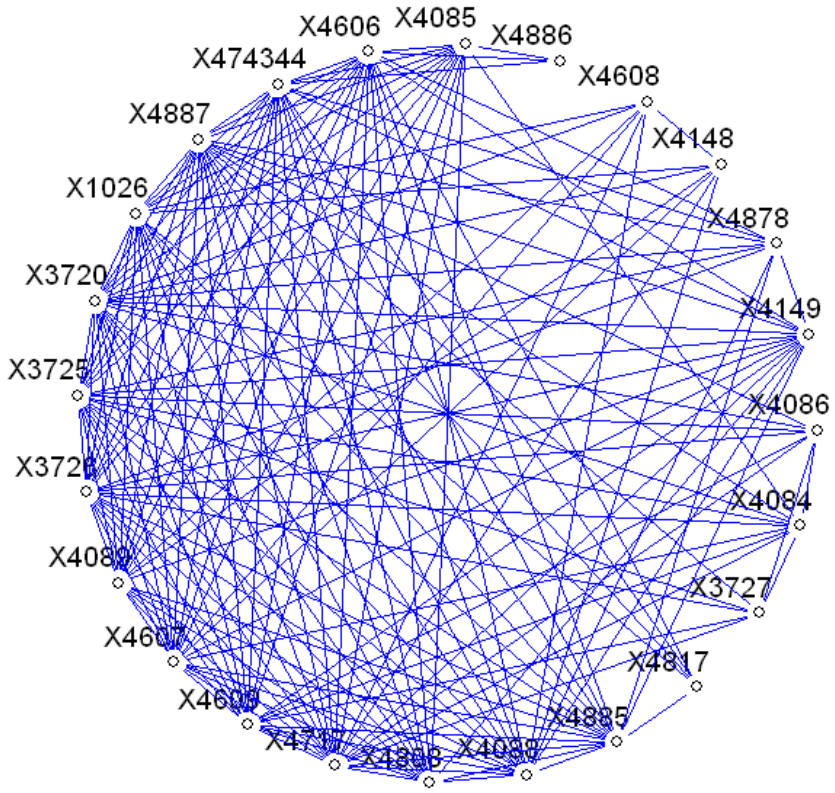

Figura 5.3: rappresentazione del modello grafico per i dati reali, selezionato tramite una regressione a passi a partire dal modello saturo.

Un simile risultato è probabilmente dovuto a un'effettiva moltitudine di legami biologici tra i geni selezionati per lo studio, e – per quanto questo rappresenti un risultato interessante di per sé – d'altra parte rende difficile le analisi di nostro primario interesse sul confronto tra modelli congiunti e combinazioni di modelli marginali. L'idea di effettuare di nuovo la selezione a passi, imponendo un criterio di significatività più rigido per l'ingresso di nuovi archi nel modello, non ha portato a risultati che si discostassero particolarmente dal precedente, né per una soglia di significatività pari a 0.01, né per una

soglia pari a 0.001, portando a concludere che il network sia effettivamente costituito da geni quasi tutti legati tra di loro.

Una diversa possibilità per indagare la combinazione di modelli marginali sta nel concentrarsi su una parte più piccola del grafo, limitandosi a un gruppetto di geni che appare correlato in base alle analisi descrittive sulla correlazione parziale. I geni considerati sono X4086, X3725, X3726, X3727, X474344 (che nella Figura 5.2 sono rappresentati in alto a sinistra), con l'aggiunta di un sesto gene (X3720): si cerca in pratica di riportarsi vicino alla struttura degli studi di simulazione. Si sono analizzate due diverse possibilità per la scelta dei sotto-grafi. In un primo momento la selezione del modello è effettuata su quello congiunto e relativamente ai due sotto-grafi composti uno da X3725, X3726, X474344 e X3727, e l'altro da X3726, X3720 e X4086. Il risultato, riportato graficamente nella Figura 5.4, mostra come la combinazione di modelli marginali sia sostanzialmente in accordo con il modello trovato su tutte le variabili.

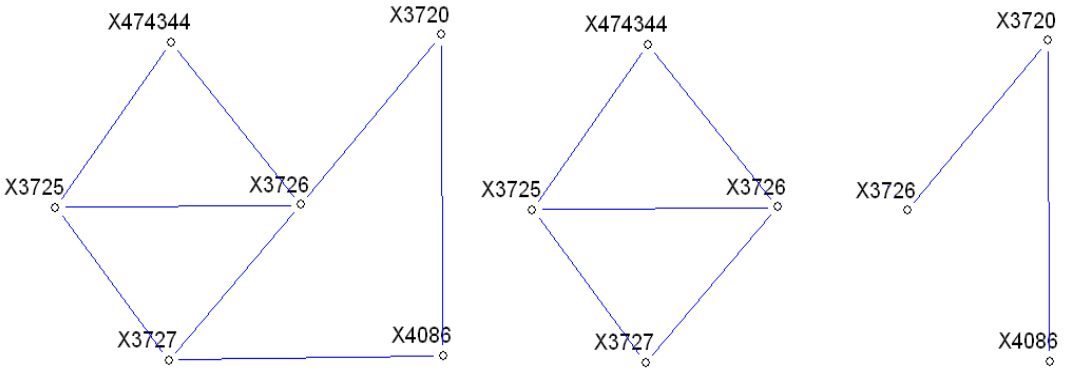

Figura 5.4: rappresentazione grafica dei modelli selezionati per il caso congiunto (a sinistra) e per i due casi marginali (esempio 1a).

Leggermente diverse sono le conclusioni se si considerano il sottoinsieme di variabili composto da X4086, X3725, X3726 e X3720 e quello costituito da X3726, X3727 e

X474344. I risultati riportati nella Figura 5.5 (il grafo a sinistra è lo stesso della Figura 5.4, solo con una diversa disposizione delle variabili) mostrano infatti una sostanziale coerenza tra modello congiunto e modelli marginali, ma anche qualche differenza, sia nel senso che alcuni archi trovati nei modelli marginali non sono presenti nel modello congiunto, sia nel senso opposto.

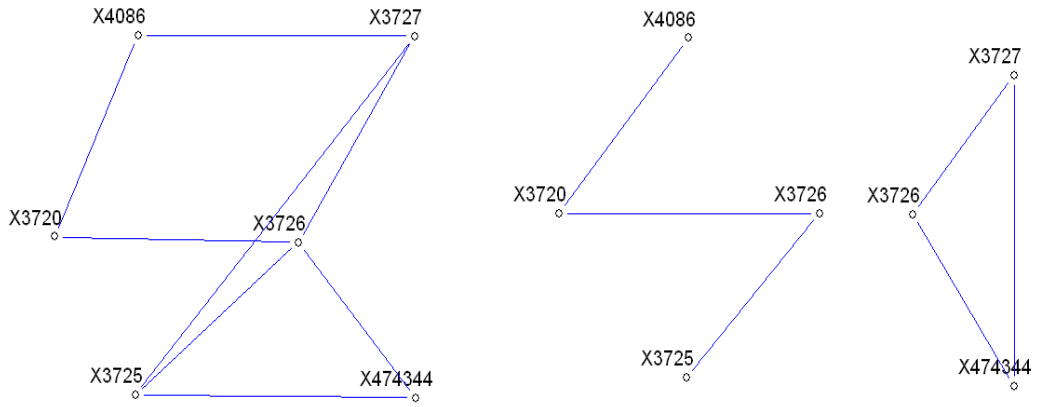

Figura 5.5: rappresentazione grafica dei modelli selezionati per il caso congiunto (a sinistra) e per i due casi marginali (esempio 1b).

Si può mostrare un'altra applicazione su un diverso sottoinsieme di variabili: si prendono come esempio i geni X4088, X4089, X4148, X4087, X4735, X4608, con sotto-grafi relativi agli insiemi X4088, X4089, X4148 contro X4089, X4087, X4735, X4608. In questo caso i modelli generati nelle due situazioni sono piuttosto differenti: la quantità di legami che non può essere individuata semplicemente combinando i sotto-grafi marginali è elevata, come viene evidenziato dalla Figura 5.6.

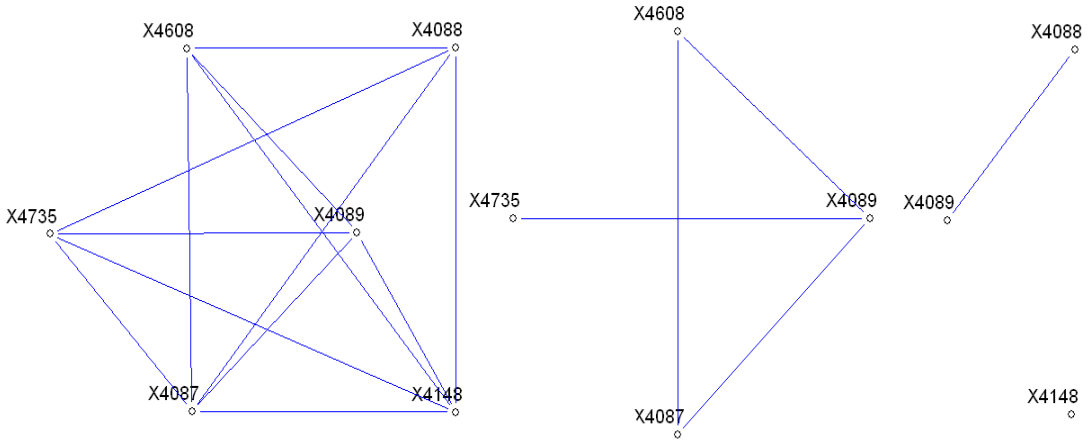

Figura 5.6: rappresentazione grafica dei modelli selezionati per il caso congiunto (a sinistra) e per i due casi marginali (esempio 2).

6. Conclusioni

6.1. Studi di simulazione

Gli studi di simulazione effettuati hanno prodotto risultati che potranno essere utili a indirizzare ulteriori ricerche. Per prima cosa hanno evidenziato le difficoltà generali di applicazione di modelli grafici a network biologici, ricordando di utilizzare le dovute cautele nell'interpretazione dei risultati di questo tipo di analisi. Tentare di inferire da un punto di vista statistico strutture di enorme complessità come i network cellulari è un'impresa difficile, che richiede strumenti adeguati e non può limitarsi a un'applicazione automatica di tecniche generali per la selezione di modelli: un uso poco consapevole di queste tecniche può portare a risultati di difficile interpretazione, o a modelli che non rispecchiano la vera struttura del network di interesse.

Nello specifico, si è visto che un network composto da quattro variabili può essere ricostruito con buona affidabilità basandosi su un unico campione di numerosità pari a 100, e che lo stesso network può essere ricostruito con relativa facilità anche a partire da due studi indipendenti (ognuno dei quali riguardante tre variabili, di cui due in comune con l'altro studio), nei casi in cui la numerosità campionaria relativa ai due studi (pari in totale a 100) sia sufficientemente equilibrata. Nei casi in cui le numerosità campionarie risultino troppo sbilanciate (70-80 unità contro 30-30 unità), invece, la ricostruzione del modello è apparsa molto più complicata.

Queste considerazioni sono confermate con maggiore forza dallo studio riguardante un network di 6 variabili: si sono trovati valori insoddisfacenti persino relativamente alla frequenza con cui le osservazioni simulate erano in grado di ricostruire il modello stesso da cui provenivano. La situazione migliorava leggermente utilizzando combinazioni di modelli marginali, ma solo a condizione che al sotto-modello più complesso fosse relativa una numerosità campionaria sufficientemente sostanziosa (altrimenti le prestazioni finivano con il peggiorare sensibilmente). Risultati migliori si sono ottenuti con numerosità campionarie nell'ordine di grandezza delle centinaia (invece che delle decine), anche se queste rappresentano situazioni che difficilmente è possibile incontrare nella realtà. In queste circostanze, tuttavia, la combinazione dei modelli marginali mostrava prestazioni molto migliori rispetto all'uso di un modello congiunto con numerosità campionaria comparabile (per quanto ancora non molto soddisfacenti in senso assoluto). Risultati come questo ultimo citato possono aprire la strada a ulteriori esperimenti, che potranno tenere conto delle seguenti indicazioni:

1. la ricostruzione di network biologici tramite l'uso di modelli grafici è una materia difficile, e al crescere della complessità del modello le prestazioni di questa tecnica diminuiscono sensibilmente;
2. quanto appena affermato si riferisce anche a network composti da un numero di variabili molto limitato, e vistosi cambiamenti (peggioramenti nelle prestazioni) possono verificarsi anche con l'inserimento di solo una o due variabili in più;
3. diversi sono i parametri che determinano quale – tra uno studio congiunto sul totale delle variabili di interesse o la combinazione di due studi marginali – offra le prestazioni migliori. Tra questi, la numerosità campionaria totale, le numerosità campionarie dei singoli studi, l'effettiva struttura del network;
4. le prestazioni di combinazioni di modelli marginali migliorano al crescere della numerosità più rapidamente di quanto facciano quelle relative a modelli congiunti;
5. nel caso di combinazioni di modelli marginali, è preferibile che sia il sotto-grafo più complesso ad essere ricostruito, se possibile, tramite un campione una numerosità più elevata;
6. se il modello "vero" supera una certa soglia di complessità, incrementi anche sostanziosi della numerosità campionaria non servono che a ottenere modesti vantaggi nelle prestazioni, soprattutto negli studi basati su di un unico modello congiunto.

6.2. Studio di dati reali

L'applicazione ai dati provenienti dallo studio sul rabdomiosarcoma ha in effetti mostrato i limiti che le tecniche prese in considerazione – già difficili da applicare nel contesto relativamente controllato degli studi di simulazione – presentano una volta introdotte nel "mondo reale". In particolare, si sono dimostrate poco potenti nel distinguere le relazioni più forti contenute in un network da quelle più labili: anche se probabilmente questo è in parte dovuto all'effettiva esistenza di molte relazioni forti tra i geni, il risultato dell'analisi è un network molto esteso e complesso, in cui tutti i geni appaiono legati a quasi tutti gli altri.

6.3. Ulteriori possibilità

In definitiva, non crediamo che i limiti mostrati dalle tecniche utilizzate debbano scoraggiare dall'approfondire le ricerche relative alla ricostruzione di network biologici tramite tecniche statistiche, contesto che al contrario appare fertile di potenzialità. Ci sembra che tali limiti vadano piuttosto pensati come strumenti per un utilizzo sempre più consapevole e non "automatico" delle metodologie considerate. Banalmente, avere la consapevolezza, attraverso studi di simulazione, che un certo modello, in un determinato contesto, viene ricostruito fedelmente solo nel 10% dei casi, è fondamentale per trattare con la dovuta cautela eventuali applicazioni di tale modello a problemi reali.

Infine, le difficoltà che ci si è trovati ad affrontare, solo in alcuni casi superabili, hanno al tempo stesso indicato una possibile nuova via da sperimentare per lo studio statistico dei network biologici. Dato che il contesto in cui la combinazione di modelli grafici sembra

funzionare bene è quello che coinvolge modelli molto piccoli, composti da non più di tre o quattro variabili (posto che le numerosità campionarie siano sufficientemente alte nei diversi campioni considerati), un nuovo approccio potrebbe consistere nel considerare uno studio di interesse come combinazione di tanti piccoli studi diversi. L'idea potrebbe essere quella di suddividere, magari sulla base di analisi descrittive di correlazione parziale, l'insieme totale delle variabili in piccoli sottoinsiemi, e cercare di combinarli non semplicemente a coppie, bensì in insiemi sempre più ampi, fino a ricostruire in qualche modo il network originario (o parte di esso). Questa tecnica potrebbe forse consentire di evidenziare le sole relazioni principali interne a un network, invece di mostrare una struttura troppo ricca di legami e pertanto difficile da interpretare. I metodi e i risultati proposti in questa tesi potrebbero essere un utile punto di partenza per muovere in questa direzione, o in altre simili legate a una maggiore interpretabilità dei risultati, che rimane uno dei principali punti di forza delle tecniche connesse all'uso di modelli grafici.

Bibliografia

Dawid, A. P. e Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of statistics*, **21**: 1272-1317.

Drton, M. e Perlman, M. D. (2004). Model Selection for Gaussian Concentration Graphs. *Biometrika 2004*, **91(3)**.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer-Verlag, New York, 2000.

Imoto, S., Goto, T. e Miyano, S (2002). Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. *Pac Symp Biocomput 2002*:175-186.

Kishino, H. e Waddell, P.J.: *Genome Informatics*. Universal Academy Press, Tokyo, 2000.

Kostka, D. e Spang, R. (2004): Finding disease specific alterations in the co-expression of genes. *Bioinformatics 2004*, **20 (Suppl. 1)**: 1194-1199.

Markowetz, F. e Spang, R. (2007). Inferring cellular networks – a review. *BMC Bioinformatics 2007*, **8**.

Massa, M. S. (2008). Combining information from Gaussian graphical models, 2008.

Nickel, D. R. (2005). Probabilities of spurious connections in gene networks: application to expression time series. *Bioinformatics 2005*, **21(7)**:1121-8.

Segal, E., Pe'er, D., Regev, A., Koller, D. e Friedman, N. (2005). Learning Module Networks. *Journal of Machine Learning Research 2005*, **6 (Apr)**:557-588.

Smith, P. W. F. e Whittaker, J. (1999). *Learning in Graphical Models*. Jordan M. MIT Press, 1999: 555-574.

Werhli, A. V., Grzegorczyk, M. e Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics 2006*, **22**:2523-2531.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, 1990.

Yamanishi, Y., Vert, J. P. e Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. Bioinformatics 2004, **20 (Suppl. 1)**:1363-1370.

Appendice A

A.1 Funzione utilizzata per lo studio 1

```
library(ggm)
library(mvtnorm)
library(mimR)

#per generare la matrice di covarianza

sim0=function(sigma){

inversa=matrix(rep(0,16),4,4)

a=0

while (a<0.05 | abs(inversa[1,2])<0.5 | abs(inversa[1,4])<0.5 |
abs(inversa[2,3])<0.5 | abs(inversa[2,4])<0.5 |
abs(inversa[3,4])<0.5){
mcorr=rcorr(4)
sigma0=matrix(sigma*mcorr,4,4,byrow=T,dimnames=list(c("x","y","z",
"t"),c("x","y","z","t")))
adj=matrix(c(0,1,0,1,1,0,1,1,0,1,0,1,1,1,1,0),4,4,byrow=T,dimnames
=list(c("x","y","z","t"),c("x","y","z","t")))
sigmahat=fitConGraph(adj,sigma0,n=nobs)$Shat
a=eigen(sigmahat)$values[4]
inversa=solve(sigmahat)}
```

```
return(sigmahat)
}

#per generare le osservazioni e effettuare la ricerca del modello

sim=function(sigmahat,nobs,nobs1,nobs2,narray){

cont1=0
cont2=0
cont3=0
cont4=0

for (j in 1:narray){
data=as.data.frame(rmvnorm(n=nobs,mean=rep(0,4),sigma=sigmahat,met
hod="chol"))
names(data)=c("x","y","z","t")

model=mim("..",as.gmData(data))
step1=stepwise(model,"u")

rand1=sample(1:nobs,nobs1,replace=F)
left=(1:nobs)[-rand1]
data_1=data[rand1,c("x","y","t")]
data_2=data[left,c("y","z","t")]

data_a=as.gmData(data_1)
data_b=as.gmData(data_2)
model_a=mim("..",data_a)
model_b=mim("..",data_b)
step2=stepwise(model_a,"u")
step3=stepwise(model_b, "u")

if (step1$mimFormula == "//t:x:y + t:y:z"  ){cont1=cont1+1}
if (step2$mimFormula == "//t:x:y"){cont2=cont2+1}
if (step3$mimFormula == "//t:y:z") {cont3=cont3+1}
if (step2$mimFormula == "//t:x:y" & step3$mimFormula ==
"//t:y:z")(cont4=cont4+1)

}

cont=c(cont1,cont2,cont3,cont4)
return(cont)

}
```

A.2 Funzione utilizzata per lo studio 2

```
library(ggm)
```

```
library(mvtnorm)
library(mimR)

nobs=100

#per generare la matrice di covarianza

sim0=function(sigma){

inversa=matrix(rep(0,36),6,6)

a=0

while (a<0.05 | abs(inversa[1,2])<0.1 | abs(inversa[1,4])<0.1 |
abs(inversa[2,3])<0.1 | abs(inversa[3,4])<0.1 |
abs(inversa[3,5])<0.1 | abs(inversa[3,6])<0.1){
mcorr=rcorr(6)
sigma0=matrix(sigma*mcorr,6,6,byrow=T,dimnames=list(c("x","y","z",
"t","v","w"),c("x","y","z","t","v","w")))
adj=matrix(c(0,1,0,1,0,0,1,0,1,0,0,0,0,1,0,1,1,1,1,0,1,0,0,0,0,0,1
,0,0,0,0,0,1,0,0,0),6,6,byrow=T,dimnames=list(c("x","y","z","t","v
","w"),c("x","y","z","t","v","w")))
sigmahat=fitConGraph(adj,sigma0,n=nobs)$Shat
a=eigen(sigmahat)$values[6]
inversa=solve(sigmahat)}

write.table(sigmahat,"varcovmatrix.txt")
write.table(solve(sigmahat),"inversa.txt")

return(sigmahat)
}

#per generare le osservazioni e effettuare la ricerca del modello

sim=function(sigmahat,nobs,nobs1,nobs2,narray){

cont1=0
cont2=0
cont3=0
cont4=0

for (j in 1:narray){
data=as.data.frame(rmvnorm(n=nobs,mean=rep(0,6),sigma=sigmahat,met
hod="chol"))
names(data)=c("x","y","z","t","v","w")

model=mim("..",as.gmData(data))
step1=stepwise(model,"u")

rand1=sample(1:nobs,nobs1,replace=F)
left=(1:nobs)[-rand1]
data_1=data[rand1,c("x","y","z","t")]
```

```
data_2=data[left,c("z","v","w")]

data_a=as.gmData(data_1)
data_b=as.gmData(data_2)
model_a=mim("..",data_a)
model_b=mim("..",data_b)
step2=stepwise(model_a,"u")
step3=stepwise(model_b,"u")

if (step1$mimFormula == "//t:x + t:z + v:z + w:z + x:y + y:z"
){cont1=cont1+1}
if (step2$mimFormula == "//t:x + t:z + x:y + y:z"){cont2=cont2+1}
if (step3$mimFormula == "//v:w + v:z") {cont3=cont3+1}
if (step2$mimFormula == "//t:x + t:z + x:y + y:z" &
step3$mimFormula == "//v:w + v:z") {cont4=cont4+1}

}

cont=c(cont1,cont2,cont3,cont4)
return(cont)

}
```

Appendice B

B.1 Matrice di correlazione semplice

Sono evidenziate in grigio scuro il 5% delle correlazioni massime (in modulo); in grigio chiaro il 5% delle correlazioni minime (in modulo).

	X4887	X4886	X4885	X4878	X4817	X4808	X474344	X4735	X4717	X4610	X4609	X4608	X4607	X4606	X4149	X4148	X4089	X4088	X4087	X4086	X4085	X4084	X3728	X3727	X3726	X3725	X3720	X1026
X1026	-0.2182	-0.1278	0.1057	-0.1723	0.2566	-0.0387	-0.1187	0.1165	0.2691	-0.0681	0.1006	0.5165	-0.2257	0.6099	-0.0463	-0.1848	-0.0320	0.0400	0.1332	0.1451	-0.2396	-0.1278	-0.0233	0.1206	0.1147	0.2599	-0.1707	1.0000
X3720	-0.1989	-0.0877	-0.2804	0.0937	0.0397	-0.1917	0.0175	0.2213	0.1098	-0.0619	-0.4589	-0.2204	-0.2626	-0.1186	-0.1507	-0.2785	0.2982	-0.2955	0.1704	0.3344	0.4324	-0.1980	0.0049	-0.0086	-0.2521	-0.0464	1.0000	
X3725	-0.0849	-0.0165	-0.0314	-0.0913	0.0274	-0.0139	-0.1194	0.1510	0.1091	-0.1550	0.0862	0.0348	-0.0843	0.1765	0.0602	-0.0079	0.0106	0.0390	0.1345	0.0583	-0.2258	-0.0539	0.0345	0.3771	0.3924	1.0000		
X3726	-0.0169	0.1709	0.0671	0.1564	0.0912	0.0345	0.2897	-0.0971	0.0488	-0.2672	0.2010	0.0550	-0.0255	0.1610	0.0779	-0.0240	-0.3276	0.0244	-0.1884	-0.0330	-0.4583	0.0031	0.2523	0.3939	1.0000			
X3727	-0.1475	0.0115	0.0550	-0.0446	-0.0100	-0.1375	-0.0582	0.1533	-0.0758	-0.1355	0.2596	0.0150	-0.1169	0.0035	0.0895	-0.0566	-0.0182	0.1306	-0.0887	0.2674	-0.1283	0.0294	0.2217	1.0000				
X3728	0.1232	0.1183	-0.0764	0.1163	-0.0275	0.0101	0.0909	-0.0964	-0.1225	-0.0840	-0.0777	0.0310	0.0331	-0.0445	-0.1255	-0.0835	-0.1425	0.1460	-0.2198	0.0724	-0.2020	-0.0407	1.0000					
X4084	0.4495	-0.0393	-0.1007	0.2365	-0.1825	0.4075	-0.1099	-0.4934	-0.1911	0.1690	-0.0970	0.1334	0.6252	0.0183	-0.0920	0.0844	-0.4541	0.1719	-0.4048	-0.2202	-0.4140	1.0000						
X4085	-0.4479	-0.1444	0.0071	-0.2987	0.1372	-0.3487	-0.1127	0.5390	0.2858	-0.0907	0.0731	-0.4485	-0.5012	-0.3701	-0.1041	-0.0428	0.6651	-0.3145	0.4339	0.2632	1.0000							
X4086	-0.3454	-0.0286	0.1179	-0.2109	0.2055	-0.3003	-0.0833	0.4054	0.3186	-0.2311	0.0629	-0.1874	-0.4280	-0.0280	-0.1905	-0.2148	0.2726	-0.1392	0.1745	1.0000								
X4087	-0.5272	-0.0249	0.0441	-0.4037	0.1634	-0.4112	0.0336	0.5093	0.3946	-0.1544	-0.0330	-0.1632	-0.5724	0.0019	0.0320	0.1289	0.6612	-0.4102	1.0000									
X4088	0.2292	0.0931	0.1922	-0.0725	-0.0862	0.1430	-0.1689	-0.1448	-0.2279	-0.0230	0.1425	0.2038	0.3279	0.1608	0.0309	0.1065	-0.2654	1.0000										
X4089	-0.5778	-0.0149	0.0660	-0.3891	0.1718	-0.4343	0.0188	0.6952	0.3151	-0.1023	0.0524	-0.4306	-0.5347	-0.1449	0.1196	-0.1225	1.0000											
X4148	0.1728	-0.0060	0.1246	-0.1009	-0.2009	0.1448	-0.0490	-0.1516	-0.1872	-0.0565	0.1159	-0.0525	0.0649	-0.1299	0.0065	1.0000												
X4149	0.0080	-0.0468	0.0005	0.1554	0.0556	-0.0052	0.1808	0.1193	-0.2121	-0.0324	0.0284	-0.0748	0.0506	-0.0530	1.0000													
X4606	-0.1279	-0.1677	-0.0576	0.0793	0.0543	0.1037	-0.0927	-0.0320	0.2763	0.0529	-0.1127	0.6038	0.0079	1.0000														
X4607	0.6367	-0.0799	-0.0893	0.4278	-0.3015	0.5539	-0.1351	-0.6760	-0.4804	0.3500	-0.1467	0.1410	1.0000															
X4608	0.1653	-0.1309	-0.1005	0.0893	0.0015	0.2209	-0.0890	-0.3153	-0.0613	0.1757	-0.1643	1.0000																
X4609	-0.1971	-0.0200	0.3548	-0.2032	0.0736	-0.1708	-0.0548	0.2565	0.0087	-0.1704	1.0000																	
X4610	0.2834	-0.1588	-0.3799	0.2286	-0.0740	0.1709	-0.0995	-0.3667	-0.3282	1.0000																		
X4717	-0.5515	-0.0263	0.0594	-0.3473	0.2798	-0.3674	0.0236	0.4521	1.0000																			
X4735	-0.6842	-0.0319	0.1023	-0.3809	0.3241	-0.5933	-0.0918	1.0000																				
X474344	-0.0957	0.3361	-0.1180	-0.0751	-0.0750	-0.0928	1.0000																					
X4808	0.6498	-0.0811	-0.0058	0.4175	-0.2549	1.0000																						
X4817	-0.3251	-0.1165	-0.0971	-0.0332	1.0000																							
X4878	0.4693	-0.1993	-0.3074	1.0000																								
X4885	-0.1060	0.0502	1.0000																									
X4886	-0.0196	1.0000																										
X4887	1.0000																											

B.2 Matrice di correlazione parziale

Sono evidenziate in grigio scuro il 5% delle correlazioni parziali massime (in modulo); in grigio chiaro il 5% delle correlazioni parziali minime (in modulo).

	X4887	X4886	X4885	X4878	X4817	X4808	X47434	X4735	X4717	X4610	X4609	X4608	X4607	X4606	X4149	X4148	X4089	X4088	X4087	X4086	X4085	X4084	X3728	X3727	X3726	X3725	X3720	X1026
X1026	-0.0014	-0.0594	0.0806	-0.1691	0.2399	0.0932	-0.0629	-0.0627	0.0042	-0.0479	0.1526	0.2327	-0.1548	0.3455	0.0119	-0.1833	0.0029	-0.1325	0.0382	0.0719	-0.1808	-0.0772	0.0278	0.0539	-0.1629	0.1447	-0.0479	1.0000
X3720	-0.0884	-0.1128	-0.1385	0.2288	-0.0615	-0.0997	0.1219	-0.0128	-0.2203	-0.2209	-0.5354	-0.1477	-0.1092	0.0687	-0.2437	-0.2150	-0.0257	-0.0162	-0.0496	0.2000	0.2437	-0.0241	-0.1018	0.1276	-0.1581	0.1310	1.0000	
X3725	0.0715	-0.0710	-0.2239	-0.2065	-0.1108	0.1413	-0.2679	0.1069	0.0736	-0.0314	0.0201	-0.1072	0.0577	-0.0607	0.0177	-0.0211	0.0435	0.0402	0.1695	-0.0801	-0.2374	-0.1128	-0.0971	0.2034	0.3227	1.0000		
X3726	-0.1014	0.0424	0.0959	0.1912	0.2132	0.0397	0.2976	-0.0691	0.0823	-0.1713	0.0744	-0.1634	-0.1223	0.1486	-0.0745	-0.0409	-0.1895	-0.1409	-0.0576	-0.1194	-0.2019	-0.1311	0.1152	0.3358	1.0000			
X3727	-0.1495	-0.0177	-0.0418	-0.0181	-0.1902	-0.0582	-0.1556	0.0000	-0.1649	0.0730	0.1680	0.1544	-0.0933	-0.1233	0.1739	0.0253	0.0837	0.1083	-0.0811	0.2708	0.0116	0.2504	0.1610	1.0000				
X3728	0.0556	-0.0468	-0.1333	0.0727	-0.0300	-0.0554	0.0210	-0.0152	-0.0562	-0.1392	-0.1714	-0.0057	-0.0301	-0.1153	-0.2494	-0.0487	0.1281	0.1198	-0.1069	0.0457	-0.1299	-0.1666	1.0000					
X4084	0.0448	-0.0379	-0.1374	-0.0634	0.0606	0.1075	-0.0490	-0.0409	0.2043	-0.1005	0.0064	-0.0311	0.3401	-0.0530	-0.1269	0.0288	-0.0163	-0.0634	-0.0395	-0.0176	-0.2203	1.0000						
X4085	-0.1546	-0.2313	-0.1270	-0.1010	-0.0181	0.2059	-0.2069	0.0219	0.2000	-0.0081	0.1772	-0.0375	-0.2111	-0.2857	-0.1296	0.0205	0.2877	-0.0521	-0.0901	-0.1181	1.0000							
X4086	0.0315	-0.0437	0.1211	-0.0551	0.0992	0.0303	-0.0360	0.0958	0.1146	-0.0030	0.0103	-0.1578	-0.1475	-0.0050	-0.1521	-0.0794	-0.0304	-0.0682	-0.0516	1.0000								
X4087	-0.0699	-0.0195	0.0122	-0.0592	0.0126	-0.1255	-0.0424	-0.0536	0.0743	-0.0548	-0.1333	0.1092	-0.1469	-0.0238	-0.0135	0.3440	0.4496	-0.3155	1.0000									
X4088	0.1238	0.1341	0.1177	-0.2657	0.1455	-0.0755	-0.0897	0.1029	-0.0883	-0.1009	0.0643	0.0718	0.1914	0.2535	0.0384	0.1270	0.0420	1.0000										
X4089	-0.0912	0.0521	0.1197	-0.0646	0.0111	0.1302	0.1562	0.3439	-0.1031	0.1989	-0.0807	-0.3198	0.0385	0.1601	0.1248	-0.2052	1.0000											
X4148	0.0294	-0.1214	-0.0548	-0.1076	-0.0842	0.0974	-0.0594	-0.0764	-0.1825	-0.1333	0.0046	-0.1422	-0.1604	0.0618	-0.0397	1.0000												
X4149	0.0002	-0.1580	0.0112	0.2093	0.0770	-0.0189	0.2623	0.1988	-0.2027	-0.1242	-0.1990	-0.0603	0.0521	-0.0527	1.0000													
X4606	-0.2675	-0.1348	-0.0875	0.2166	-0.2421	0.1604	-0.0758	0.0659	0.3268	0.1160	-0.0641	0.3630	0.0181	1.0000														
X4607	-0.0624	-0.1563	-0.0162	0.0904	-0.0987	0.1102	-0.1573	-0.2566	-0.1043	0.0700	-0.0507	-0.1742	1.0000															
X4608	-0.0009	-0.1123	-0.0873	-0.0981	0.0256	0.0416	-0.0030	-0.0680	-0.1675	0.0057	-0.1836	1.0000																
X4609	-0.0981	-0.1311	0.1344	0.0912	-0.0381	-0.0836	0.0872	0.1870	-0.1870	-0.0203	1.0000																	
X4610	0.0313	-0.1060	-0.3647	0.0002	0.0503	-0.1365	-0.0733	-0.2844	-0.2238	1.0000																		
X4717	-0.1275	-0.0594	-0.0337	-0.1105	0.0812	-0.0898	0.0459	0.0714	1.0000																			
X4735	-0.1438	-0.0144	-0.1483	0.0342	0.0720	-0.1969	-0.2468	1.0000																				
X47434	-0.1075	0.1518	-0.2467	-0.2113	-0.1525	0.0283	1.0000																					
X4808	0.3390	0.0233	0.1123	0.1691	-0.0106	1.0000																						
X4817	-0.1596	-0.1041	-0.1864	0.1186	1.0000																							
X4878	0.1873	-0.1379	-0.1931	1.0000																								
X4885	-0.0688	-0.0840	1.0000																									
X4886	-0.0506	1.0000																										
X4887	1.0000																											