



Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXII ciclo

**Metodi statistici a variabili latenti
per lo studio di fenomeni finanziari**

Luca De Angelis

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2010



Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXII ciclo

Metodi statistici a variabili latenti per lo studio di fenomeni finanziari

Luca De Angelis

Coordinatore:
Prof.ssa Daniela Cocchi

Tutor:
Prof. Michele Costa

Co-Tutor:
Dott. Leonard J. Paas

Settore Disciplinare:
SECS-S/01

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2010

“ Theory without empirics is empty. Empirics without theory is blind. ”
Immanuel Kant (1724 – 1804)

Indice

Introduzione	7
1. Il ruolo delle variabili latenti nello studio dei fenomeni finanziari	11
1.1 La misura delle variabili	14
1.2 La classificazione	16
1.2.1 <i>La classificazione nei mercati finanziari</i>	20
1.3 La misura e la valutazione del profilo rischio-rendimento	22
1.4 L'analisi della covarianza e dell'andamento di una serie storica finanziaria	24
1.5 L'importanza dei modelli a struttura latente nello studio dei fenomeni finanziari	28
2. Metodi statistici per le variabili latenti	29
2.1 Gli aspetti teorici	29
2.1.1 <i>Il modello a struttura latente</i>	32
2.1.2 <i>La stima del modello a struttura latente: l'algoritmo EM</i>	33
2.2 L'analisi a classi latenti	35
2.2.1 <i>Il modello a classi latenti con variabili manifeste dicotomiche e una variabile latente dicotomica</i>	35
2.2.2 <i>Il modello a classi latenti con variabili manifeste dicotomiche</i>	36
2.2.3 <i>La stima dei parametri</i>	37

2.2.4	<i>La stima degli errori standard</i>	40
2.2.5	<i>L'analisi della bontà del modello</i>	42
2.2.6	<i>L'analisi a posteriori</i>	44
2.2.7	<i>L'interpretazione delle classi</i>	44
2.2.8	<i>Il modello a classi latenti con variabili manifeste politomiche</i>	45
2.2.9	<i>La stima dei parametri</i>	46
2.2.10	<i>L'analisi a posteriori</i>	48
2.2.11	<i>Il modello a classi latenti con variabili manifeste politomiche ordinabili</i>	48
2.2.12	<i>L'identificabilità del modello</i>	49
2.3	<i>I modelli a classi latenti: ulteriori impostazioni ed estensioni</i>	51
2.3.1	<i>I modelli a classi latenti vincolati</i>	53
2.3.2	<i>I modelli con più di una variabile latente</i>	55
2.3.3	<i>I modelli con variabili esterne</i>	56
2.3.4	<i>I modelli per diverse sottopopolazioni</i>	57
2.3.5	<i>I modelli di dipendenza locale</i>	59
2.4	<i>I modelli a profilo latente</i>	60
2.4.1	<i>La stima di massima verosimiglianza</i>	61
2.4.2	<i>Ulteriori metodi di stima</i>	62
2.4.3	<i>L'allocazione nelle classi latenti</i>	63
2.5	<i>I modelli a tratto latente</i>	64
2.5.1	<i>I modelli a tratto latente per variabili dicotomiche</i>	66
2.5.2	<i>I modelli a tratto latente per variabili politomiche</i>	68
2.5.3	<i>Le principali procedure di stima</i>	69
2.6	<i>Il latent Markov model</i>	71
2.6.1	<i>La stima dei parametri</i>	74
2.6.2	<i>Le estensioni del LMM</i>	77

3. Metodi statistici per le variabili latenti: gli ultimi sviluppi	79
3.1 Il modello a classi latenti fattoriali	79
3.1.1 <i>L'interpretazione dei fattori</i>	83
3.1.2 <i>Le estensioni al modello a classi latenti fattoriale base</i>	85
3.2 Il latent Markov model di mistura	87
3.2.1 <i>I casi speciali del modello MLMM vincolato</i>	90
3.2.2 <i>La stima dei parametri</i>	91
3.2.3 <i>La scelta del numero di classi latenti e stati latenti</i>	92
3.3 Il latent Markov model misto	93
3.3.1 <i>Il MHMM con effetti casuali per il processo osservato</i>	93
3.3.2 <i>Il MHMM con effetti casuali per il processo latente</i>	96
4. Soluzioni metodologiche per l'analisi dei fenomeni finanziari:	
il caso del mercato azionario italiano	99
4.1 Il mercato azionario italiano	101
4.2 I dati	105
4.3 La classificazione dei titoli finanziari	110
4.3.1 <i>I metodi a struttura latente per la classificazione</i>	110
4.3.2 <i>Le implicazioni dell'analisi: il confronto tra la nuova</i> <i>classificazione e quella settoriale tradizionale</i>	114
4.3.3 <i>I dati per le analisi</i>	115
4.3.4 <i>La stima dei modelli</i>	118
4.3.5 <i>La nuova classificazione dei titoli azionari</i>	123
4.4 La misura e la valutazione del profilo rischio-rendimento	125
4.4.1 <i>Il profilo rischio-rendimento nell'analisi a classi latenti</i>	125
4.4.2 <i>Il confronto tra l'analisi a classi latenti e</i> <i>la tecnica delle k-medie</i>	126
4.4.3 <i>Il profilo rischio-rendimento nell'analisi a profilo latente</i>	132

4.4.4	<i>Le implicazioni per l'analisi del portafoglio: la diversificazione settoriale</i>	135
4.4.5	<i>Le implicazioni per l'analisi del portafoglio: la scelta di un portafoglio efficiente</i>	137
4.5	L'analisi e lo sviluppo dei numeri indici finanziari	139
4.5.1	<i>Le questioni metodologiche affrontate</i>	140
4.5.2	<i>I numeri indici dei prezzi settoriali</i>	140
4.5.3	<i>I risultati</i>	144
4.5.4	<i>Le implicazioni per l'analisi del portafoglio</i>	150
4.5.5	<i>Osservazioni conclusive</i>	152
4.6	L'analisi della dinamica e la previsione del mercato azionario	156
4.6.1	<i>Lo scopo dell'analisi</i>	156
4.6.2	<i>Il modello e i dati</i>	158
4.6.3	<i>La stima del modello</i>	161
4.6.4	<i>Le implicazioni per l'analisi finanziaria</i>	164
4.6.5	<i>Osservazioni conclusive e sviluppi futuri</i>	169
	Conclusioni	173
	Bibliografia	177
	Appendice A	187
	Appendice B	196
	Appendice C	203

Introduzione

Negli ultimi decenni il concetto di variabile latente ha riscosso un enorme successo nelle discipline statistiche, dando luogo ad una vasta letteratura sia di indirizzo teorico, sia in campo applicativo. In particolare, nelle scienze sociali e in psicomatria, l'uso del concetto di variabile latente è stato largamente adottato per far fronte al problema di misurare quantità che, in natura, non possono essere direttamente osservate. Il primo autore ad introdurre il concetto di variabile latente è stato Charles Spearman nel suo articolo del 1904 sul *American Journal of Psychology* per definire il concetto di intelligenza generale. Nonostante il lavoro seminale di Spearman sia considerato come il punto di partenza per l'analisi fattoriale, fu durante la seconda guerra mondiale che la metodologia statistica per lo studio delle variabili latenti venne formalizzata teoricamente. Il contributo di Paul F. Lazarsfeld al team multidisciplinare impiegato dal Dipartimento della Guerra del governo americano per gli studi sociali e psicologici del personale militare consistette nella formulazione della teoria e la dimostrazione dell'uso dei modelli a struttura latente nel quarto volume del *The American Soldier: Studies in Social Psychology in WW II* (Stouffer, 1949-50), intitolato *Measurement and Prediction*. Più tardi, Lazarsfeld contribuì ad un capitolo sull'analisi a struttura latente nel monumentale lavoro *Psychology: A Study of A Science* (1959) fino alla stesura del libro con Neil W. Henry, *Latent Structure Analysis* (1968), che colleziona e raffina i progressi fatti in questa metodologia statistica in venticinque anni. Negli ultimi quaranta anni, numerosi ricercatori di statistica, psicologia e sociologia hanno contribuito allo studio dei modelli riconducibili all'analisi a struttura latente.

La letteratura relativa ai metodi statistici per le variabili latenti ha interessato, seppure in maniera più limitata, anche il campo della ricerca economica. Nonostante esistano studi che sviluppano modelli a variabili latenti applicati a

fenomeni di tipo economico, i lavori che considerano fenomeni finanziari sono attualmente ad uno stadio soltanto preliminare.

L'obiettivo della tesi è quello di ricorrere alle potenzialità esplicative ed investigative dei metodi statistici a variabili latenti per l'analisi dei fenomeni finanziari. Si fa riferimento, in particolare, ai modelli a classe latente che consentono di sviluppare soluzioni metodologicamente corrette per importanti problemi ancora aperti in campo finanziario.

In primo luogo, la natura stessa delle variabili finanziarie è riconducibile al paradigma delle variabili latenti. Infatti, variabili come il rischio ed il rendimento atteso non possono essere misurate direttamente e necessitano di approssimazioni per valutarne l'entità. Tuttavia, trascurare la natura non osservabile delle variabili finanziarie può portare a decisioni di investimento inopportune o, talvolta, addirittura disastrose.

Secondariamente, vengono prese in considerazione le capacità dei modelli a classi latenti nel contesto della classificazione. Per i prodotti finanziari, infatti, una corretta classificazione sulla base del profilo (latente) di rischio e rendimento rappresenta il presupposto indispensabile per poter sviluppare efficaci strategie di investimento. Ci si propone, inoltre, di sviluppare un collegamento, finora mancante, tra uno dei principali riferimenti della finanza moderna, la teoria classica del portafoglio di Markowitz (1952), e la metodologia statistica dei modelli a variabili latenti. In questo contesto, si vogliono investigare, in particolare, i benefici che i modelli a variabili latenti possono dare allo studio di ottimizzazione del profilo rischio - rendimento atteso di un portafoglio di attività finanziarie.

Lo sviluppo di numeri indici dei prezzi delle attività finanziarie caratterizzati da una solida base metodologica rappresenta un ulteriore aspetto nel quale i modelli a classe latente possono svolgere un ruolo di fondamentale importanza. In particolare, si propone di analizzare il contesto dei numeri indici dei prezzi

settoriali, che costituiscono uno dei riferimenti più importanti nelle strategie di diversificazione del rischio.

Il passaggio da una specificazione statica ad una analisi dinamica coglie aspetti metodologici di frontiera che possono essere investigati nell'ambito dei modelli markoviani a classi latenti. Il profilo latente di rischio – rendimento può essere, così, investigato in riferimento alle diverse fasi dei mercati finanziari, per le quali le probabilità di transizione consentono valutazioni di tipo previsivo di forte interesse.

1. Il ruolo delle variabili latenti nello studio dei fenomeni finanziari

Nelle scienze economiche e finanziarie, diverse variabili che rivestono un notevole interesse non possono essere rilevate precisamente, in quanto non sono direttamente osservabili oppure la loro misura è soggetta ad errore. La pluralità di differenti situazioni, di natura sia teorica che empirica, può essere ricondotta a tre distinte tipologie (Trivellato, 1990): il caso di errori di misura nelle variabili, la presenza di variabili latenti o di variabili di aspettativa.

Si incorre nel primo caso quando il processo di misura della variabile oggetto di studio può essere non accurato e di conseguenza si dispone di un indicatore che misura la variabile con errore. Questo è il caso di numerose variabili, soprattutto economiche, ottenute tramite indagine campionaria e, perciò, soggette ad errore campionario, oppure perché ricavate seguendo procedure di aggregazione di informazioni provenienti da diverse fonti.

Nel secondo caso si fa riferimento a costrutti ipotetici per i quali non esiste un corrispettivo empirico. Si ricorre, allora, a variabili osservabili, di solito più di una, per approssimare il costrutto di interesse. Le variabili non osservabili, definite latenti, sono già presenti nei lavori di Spearman (1904) per definire la *general intelligence* o “fattore g”.

Anche il terzo caso riguarda delle variabili non direttamente osservabili, in quanto costituiscono delle aspettative riferite ad eventi o a grandezze future.

Negli studi finanziari e, in particolar modo, nella teoria classica del portafoglio (Markowitz, 1952), le analisi si basano essenzialmente su due variabili che non possono essere direttamente rilevate: il rischio e il rendimento atteso di un'attività finanziaria. Tali variabili possono essere ricondotte alle tre situazioni descritte in precedenza. Sia il rendimento atteso, sia il rischio, infatti, non possono essere misurati con precisione in quanto non esiste un corrispettivo empirico

univoco per queste variabili. Di conseguenza, si è soliti considerare un insieme di indicatori osservabili come, ad esempio, la media dei rendimenti passati ed alcune misure di variabilità per ottenere una approssimazione delle due variabili rendimento atteso e rischio. In altre parole, il rendimento atteso ed il rischio sono le due variabili *latenti* sottostanti e caratterizzanti il fenomeno finanziario oggetto di indagine.

Inoltre, entrambe queste entità corrispondono ad un'aspettativa (futura e, in quanto tale, incognita) alla quale l'investitore si riconduce nel momento in cui decide di acquistare un'attività finanziaria o compie una scelta tra due o più potenziali opportunità di investimento.

In termini più rigorosi, si è soliti approssimare la variabile non osservabile rendimento atteso $E(R)$ e rischio V come

$$E(R) \approx \bar{R} \quad (1.1)$$

dove \bar{R} indica la media dei rendimenti passati dell'attività finanziaria.

Per la variabile non osservabile rischio V si fa, invece, riferimento alla relazione

$$V = f(\sigma, \tau_v). \quad (1.2)$$

L'equazione (1.2) definisce il rischio come funzione dello scarto quadratico medio dei rendimenti passati σ e di alcuni percentili τ_v (di solito, il primo τ_1 o il quinto τ_5) che esprimono, rispettivamente, la variabilità ed i valori estremi della distribuzione dei rendimenti del titolo finanziario.

Tra gli obiettivi di questa tesi vi è l'analisi e l'approfondimento della relazione tra le variabili latenti $X = [E(R), V]$ e le loro approssimazioni ricavate dagli indicatori osservabili $Z = [\bar{R}, \sigma, \tau_v]$. In particolare, si desidera illustrare come la relazione tra le variabili X e Z possa essere efficacemente descritta ricorrendo alla metodologia statistica per l'analisi delle variabili latenti sviluppata a partire dalla metà del secolo scorso (Lazarsfeld, 1950a; 1950b), principalmente per lo studio di

fenomeni sociologici e psicometrici. I modelli a struttura latente offrono, infatti, una valida impostazione metodologica per l'analisi dei fenomeni finanziari che, tuttavia, come attestano i pochissimi lavori esistenti in letteratura, finora è stata solo parzialmente sfruttata.

Questa tesi si prefigge l'obiettivo di trattare la non osservabilità del rendimento atteso e del rischio sfruttando le potenzialità che i modelli a struttura latente hanno nel definire questi costrutti e nel renderli facilmente interpretabili.

Nonostante il limitato interesse che queste metodologie hanno riscontrato negli studi finanziari rispetto ad altri campi delle scienze sociali, il ricorso ai modelli a variabili latenti offre spunti interessanti per affrontare, in un contesto metodologico rigoroso, alcune importanti questioni alle quali, finora, sono state proposte soluzioni soltanto parziali. Nel seguito del lavoro si valutano alcune delle principali problematiche metodologiche legate ai fenomeni finanziari: il problema della misura delle variabili, la classificazione, la misura e la valutazione del profilo rischio-rendimento e l'analisi dinamica dell'andamento di una serie storica dei rendimenti attraverso lo studio della struttura di covarianza sottostante un processo stocastico.

1.1 La misura delle variabili

Il passaggio dalle variabili osservate a quelle latenti non è mai un processo banale e richiede una particolare attenzione, considerando il fatto che gli indicatori osservabili sono solo approssimazioni dei costrutti latenti. Nel campo delle variabili finanziarie ed in particolare nella teoria del portafoglio, si è soliti ricorrere alle approssimazioni presentate nelle equazioni (1.1) e (1.2).

La teoria classica del portafoglio basata sull'analisi media-varianza è largamente riconosciuta come la pietra miliare della moderna teoria di investimento. Nonostante la sua semplicità ed il fatto che siano passati oltre cinque decenni e mezzo dal lavoro seminale di Markowitz (1952) sulla teoria dell'allocazione del portafoglio in regime di incertezza, essa rimane il metodo più usato per la scelta della composizione del portafoglio finanziario. Ciononostante, l'approccio media-varianza presenta tuttora la problematica riguardante la natura delle variabili adottate nell'analisi. Ad esempio, si consideri l'assunzione che la media dei rendimenti osservati di un titolo sia uguale al suo rendimento atteso. In realtà, quest'ultimo fa riferimento al prezzo che l'attività finanziaria avrà in futuro e, quindi, oggi risulta incognito. Nella pratica, si considera la performance che il titolo ha avuto nel passato e si assume che la medesima avvenga anche nel futuro. In maniera analoga al rendimento atteso, il rischio viene usualmente approssimato dalla varianza o dalla deviazione standard della distribuzione dei rendimenti del titolo. Inoltre, come mostrato nell'equazione (1.2), può essere utile valutare anche i primi percentili di tale distribuzione. Questi indicatori permettono di misurare l'entità dei valori estremi che, negli studi di gestione del rischio e nella mente dell'investitore avverso al rischio, assumono un ruolo di particolare rilievo in quanto rappresentano i forti ribassi nel prezzo dell'attività finanziaria. Queste assunzioni sulle variabili oggetto di studio comportano un approccio all'analisi del fenomeno piuttosto semplicistico e possono far incorrere in risultati errati.

Una proposta metodologica innovativa consiste nel trasformare le variabili \bar{R} , σ e π_v da continue in categoriche. Paradossalmente, nonostante la perdita di informazione, questo passaggio permette di ottenere alcuni benefici nella misurazione delle variabili osservate e, di conseguenza, nelle approssimazioni di quelle latenti. In particolare, la trasformazione dei valori puntuali in 2 o 3 categorie, che rappresentano i livelli basso, medio ed alto della variabile osservata, permette di ridurre l'importanza della bontà e dell'accuratezza di tali approssimazioni. Il ricorso a variabili categoriche ordinali consente di oltrepassare la rigidità dei valori puntuali e di evitare un possibile ordinamento inopportuno imposto dai valori osservati.

Come esempio, si consideri un titolo finanziario caratterizzato da un basso rendimento medio, poniamo $\bar{R} = 0,01$. Questa osservazione viene classificata nella categoria “basso” dell'indicatore categorico rendimento medio, M . È piuttosto intuitivo dedurre che il rendimento atteso di quel titolo assumerà un valore basso anche in futuro, mentre è più difficile affermare che $E(R)$ sarà esattamente uguale a 0,01.

Per queste ragioni, una classificazione basata su variabili categoriche ordinali suggerisce un ordinamento più flessibile ed una corrispondenza maggiore tra l'indicatore osservato del rendimento medio e quello atteso (latente).

Applicando una procedura analoga anche alle variabili σ e τ_v e raggruppando i valori puntuali in due (tre) categorie, si possono ottenere dei benefici interessanti nell'approssimazione della variabile V e nella misura del livello di rischio.

1.2 La classificazione

La classificazione è una delle operazioni essenziali nell'indagine scientifica ed è particolarmente importante nelle scienze sociali, dove una teoria esauriente è spesso mancante ed il primo passo nell'analisi è normalmente quello di rilevare qualche tipo di struttura esistente nei dati (Bartholomew et al., 2002).

I metodi di classificazione hanno una lunga tradizione in discipline come la biologia con la tassonomia (Sokal e Sneath, 1963), nel marketing attraverso la segmentazione della clientela (Smith, 1956) e nel campo dell'archeologia (Hodson et al., 1966), tuttavia solo recentemente si è sentita la necessità di proporre delle rigorose procedure basate su modelli statistici anche per l'analisi di variabili finanziarie.

La classificazione comporta l'investigazione di un insieme di oggetti volta a stabilire se questi possano essere riassunti in un piccolo numero di classi di oggetti simili tra loro (Gordon, 1999). Generalmente, i modelli di classificazione collegano un insieme di variabili indipendenti Z (input) ad una variabile dipendente W (output o variabile classificatoria) discreta. La classificazione consiste nel prevedere nel modo più accurato possibile la variabile W usando le informazioni fornite dalle variabili Z . Nell'approccio di classificazione non supervisionato questo obiettivo viene raggiunto specificando un modello statistico per la distribuzione condizionata $P(Z|W)$. Definendo la variabile categorica non osservabile X come variabile di intervento, la funzione congiunta $P(W, Z)$ può essere fattorizzata come

$$P(W, Z) = P(W)P(Z|W) = P(W) \sum_x P(X|W)P(Z|W, X). \quad (1.3)$$

In questo approccio non esiste una relazione tra l'adattamento del modello e la performance della classificazione in quanto la stima della funzione di verosimiglianza si basa sulla distribuzione condizionale $P(Z|W)$ o sulla funzione

congiunta $P(W, Z)$ e non direttamente su $P(W | Z)$ come nei metodi di classificazione supervisionata quali, ad esempio, i modelli per reti neurali *mixture-of-experts* (Bishop, 1995) o la regressione a classi latenti o regressione di mistura (Wedel e DeSarbo, 1994). Per questo motivo, i modelli specificati nell'equazione (1.3) si attribuiscono alla famiglia dei metodi di classificazione (o di apprendimento) non supervisionata.

La distribuzione predittiva di W dato Z , necessaria per ottenere la classificazione delle unità nei gruppi, può essere ottenuta ricorrendo al teorema di Bayes:

$$P(W | Z) = \frac{P(W)P(Z | W)}{\sum_w P(W)P(Z | W)}.$$

Dalla equazione (1.3) si possono ricavare numerosi classificatori noti che dipendono dalla forma di $P(Z | W)$. Il classificatore *Naive Bayes* (NB), ad esempio, assume che le variabili Z siano mutualmente indipendenti nei livelli di W : $P(Z | W) = \prod_i P(z_i | W)$. Ovviamente, la forma esatta della densità condizionale $P(z_i | W)$ dipende dalla scala delle variabili z_i . Alcune forme meno vincolate per $P(Z | W)$ sono usate nelle classificazioni bayesiane ad albero e nell'analisi discriminante.

Magidson e Vermunt (2001, 2003) propongono di ricorrere al modello a classi latenti come metodo per la classificazione. Questo si ottiene come caso particolare del modello definito nell'equazione (1.3) quando

$$P(Z | W, X) = \prod_i P(z_i | X).$$

Questo metodo di classificazione assume che tutte le variabili, comprese le W , siano indipendenti le une dalle altre all'interno delle diverse classi latenti K di cui è composta la variabile di intervento non osservabile (latente) X .

In maniera analoga all'analisi dei gruppi (*cluster analysis*) che permette di derivare una utile divisione in classi e dove sia il numero che le proprietà di tali

classi devono essere determinate (Everitt, 1993), anche l'analisi a classi latenti di tipo esplorativo, ovvero quando non è nota a priori nessuna informazione riguardante i gruppi che sia andranno a formare, ricopre la stessa funzione. In pratica, le unità vengono classificate in una delle K classi latenti, dove sia il valore (discreto) di K che la dimensione delle classi è oggetto di indagine da parte del ricercatore. Inoltre, le unità appartenenti alla stessa classe latente risultano simili rispetto alle variabili osservate, ovvero si assume che i valori osservati provengano dalla stessa distribuzione di probabilità in cui, però, i parametri sono quantità incognite che devono essere stimate.

Tuttavia, a differenza dei metodi di analisi dei cluster più noti (ad esempio, il metodo delle *k-medie*), la classificazione ottenuta da un modello a classi latenti si riferisce ad un processo di raggruppamento basato su di un modello probabilistico (Banfield e Raftery, 1993). In altri termini, viene postulato un modello statistico per la popolazione dalla quale proviene il campione oggetto di studio. Più precisamente, si assume che i dati siano generati da una mistura di distribuzioni di probabilità sottostanti (Di Zio et al., 2007). Usando il metodo della massima verosimiglianza per la stima dei parametri, la classificazione delle unità statistiche nei gruppi scaturisce dalla massimizzazione della funzione di log-verosimiglianza. In maniera analoga, anche le tecniche di clustering partizionali (non gerarchiche) si basano su un criterio specifico per ottimizzare la classificazione delle unità nei gruppi. Di solito, i criteri consistono nella minimizzazione della varianza entro e/o nella massimizzazione della varianza tra i cluster. Tuttavia, un ulteriore vantaggio nel ricorrere ad un modello statistico per ottenere la classificazione delle unità nei gruppi è la possibilità di scegliere il criterio di raggruppamento in maniera meno arbitraria verificando la bontà di adattamento del modello ai dati con l'ausilio di uno o più test statistici.

Magidson e Vermunt (2002b) individuano altri vantaggi nell'uso dei modelli a classi latenti per la classificazione dovuti alla loro maggiore flessibilità rispetto ai

tradizionali metodi partizionali di clustering come il k-medie. Il primo riguarda il fatto che le distribuzioni delle variabili osservate nei gruppi possono essere indifferentemente semplici o complicate. Un secondo punto di forza consiste nel poter imporre vincoli sui parametri per ottenere una maggiore parsimonia e ricorre a dei test per verificare la loro validità. Un terzo vantaggio riguardante l'approccio di raggruppamento basato su modelli a classi latenti è che, a differenza dei metodi di clustering partizionali, consente di non dover normalizzare le variabili per problemi legati alla scala: se, ad esempio, si considerano distribuzioni gaussiane con varianze incognite, i risultati della classificazione saranno gli stessi sia che le variabili osservate vengano o meno normalizzate. Un ulteriore vantaggio consiste nel fatto di poter considerare, anche congiuntamente, variabili con diverse tipologie di misura. Infine, è possibile includere nel modello delle variabili esogene (covariate) per meglio descrivere le diverse caratteristiche dei vari gruppi senza dover ricorrere ad analisi supplementari come, ad esempio, ad una analisi discriminante con lo scopo di individuare le differenze tra i cluster ricavati dalla tecnica delle k-medie.

La classificazione ottenuta ricorrendo ai modelli a classi latenti si basa su di un approccio probabilistico nell'assegnazione delle unità ai vari gruppi. Questo significa che, nonostante ogni oggetto sia assunto come appartenente ad una sola classe, bisogna tenere presente il fatto che l'allocazione avviene con un certo grado di incertezza. La classificazione tramite modelli a classi latenti risulta concettualmente simile alle tecniche di clustering *fuzzy* (Kaufman e Rousseeuw, 1990). Tuttavia, una importante differenza tra i due approcci consiste nella possibilità di classificare le unità appartenenti alla popolazione da cui il campione è stato estratto seguendo la regola di allocazione dettata dalle probabilità a posteriori stimate dal modello, che, al contrario, il metodo fuzzy non è in grado di fornire.

Infine, ricorrendo a dati simulati, Magidson e Vermunt (2002a) dimostrano come la classificazione ottenuta con i metodi a classi latenti sia più efficiente rispetto a quella definita dal k-medie riducendo sensibilmente l'errore di misclassificazione. Nel loro lavoro, gli autori generano diversi dataset composti da due variabili distribuite normalmente con medie $\mu_1 \neq \mu_2$ che rappresentano due diversi cluster. Nel caso in cui le varianze delle due variabili siano uguali, entrambi i metodi riescono ad individuare i due cluster correttamente. All'aumentare dello scarto tra le varianze delle due variabili o inserendo una struttura di covarianza entro i cluster, la classificazione ottenuta con il modello a classi latenti risulta più precisa di quella ricavata dal metodo delle k-medie anche nel caso in cui si considerino variabili standardizzate. Inoltre, è interessante notare come in tutti i casi analizzati dagli autori, il modello a classi latenti con due classi viene individuato ricorrendo al confronto tra i valori del criterio di informazione BIC, a differenza del metodo delle k-medie dove il numero dei cluster deve essere deciso in anticipo su basi soggettive o in merito a conoscenze a priori sul fenomeno.

1.2.1 La classificazione nei mercati finanziari

In questo lavoro, si sfrutta la provata superiorità dei modelli a classi latenti rispetto ai metodi di classificazione convenzionali al fine di definire gruppi di attività finanziarie con caratteristiche omogenee. In particolare, si propone di classificare i titoli basandosi sul loro profilo di rischio-rendimento che, negli studi finanziari, è il concetto alla base di ogni decisione di investimento (si veda paragrafo **1.3**).

Nei mercati azionari, i titoli vengono classificati in numeri indici o in altri prodotti finanziari principalmente in base a due caratteristiche, oltre a quella geografica. La prima riguarda il settore di riferimento in cui opera la società e la seconda è la capitalizzazione.

La classificazione riferita al settore economico raggruppa nella stessa classe le società che hanno business principale simile. Come analizzato più nel dettaglio nei paragrafi **4.3** e **4.4** per quanto riguarda la classificazione settoriale GICS, l'allocazione dei titoli rispetto al settore merceologico presenta alcune problematiche rilevanti che non permettono una corretta valutazione finanziaria. La classificazione settoriale dovrebbe infatti caratterizzare in maniera significativa la performance del titolo finanziario, tuttavia, nella realtà, i processi di classificazione adottati non sembrano in grado di formare dei gruppi tra loro eterogenei e composti da titoli che presentano caratteristiche finanziarie simili. Di conseguenza, il profilo di rischio-rendimento di un settore non viene adeguatamente definito e discriminato rispetto agli altri.

Il secondo metodo di classificazione usato nei mercati azionari è quello riferito alla capitalizzazione: un esempio è rappresentato dai numeri indici dei prezzi nei quali il paniere sottostante viene definito raggruppando i titoli in base alla loro rilevanza in termini di valore; la capitalizzazione viene infatti misurata dal prodotto tra il prezzo ed il numero di quote azionarie emesse della società. L'obiettivo di questi numeri indici è quello di rappresentare diversi segmenti del mercato e dell'economia del paese suddividendo le società per azioni in gruppi definiti in base al loro valore di mercato. Nonostante questo criterio di classificazione riesca a discriminare meglio del settore economico alcune delle caratteristiche finanziarie dei titoli, nel paragrafo **4.5** si dimostra come il metodo di classificazione ottenuto ricorrendo ai modelli a struttura latente comporti una nuova definizione ed un miglioramento della performance dei numeri indici azionari dei prezzi.

1.3 La misura e la valutazione del profilo rischio-rendimento

Il profilo rischio-rendimento di una attività finanziaria rappresenta il punto cruciale nella teoria standard del portafoglio, in quanto sintetizza le caratteristiche fondamentali alla base dei processi decisionali di investimento.

Nei paragrafi **1** ed **1.1** si è sottolineato il fatto che le due quantità che compongono il profilo, il rendimento atteso $E(R)$ ed il rischio V , siano variabili latenti e come sia possibile ottenerne una approssimazione ricorrendo ad un insieme di variabili osservate definito con Z (si vedano le equazioni (1.1) e (1.2)). Nell'estensione della rappresentazione vettoriale $X = [E(R), V]$ al caso di n titoli finanziari, l'elemento $X_h = [E(R_h), V_h]$, riga della matrice di dimensione $n \times 2$, sintetizza il profilo rischio-rendimento del titolo h -esimo, per $h = 1, 2, \dots, n$. Essendo composto da variabili non osservabili, anche il vettore X_h è, ovviamente, una variabile latente. Perciò, determinare il valore di X_h permette di misurare il profilo rischio-rendimento (latente) del titolo h -esimo e consente di svelare le caratteristiche finanziarie rilevanti dell'attività finanziaria oggetto di analisi.

Come per le singole variabili $E(R)$ e V , anche la natura latente del vettore X viene frequentemente trascurata, ma ricorrendo ai modelli a struttura latente è possibile trattare su un piano metodologico corretto ed avanzato la non osservabilità del profilo rischio-rendimento. Inoltre, una volta ottenuta la classificazione degli n titoli finanziari nelle K classi latenti (paragrafo **1.2**) è possibile interpretare le proprietà finanziarie di ogni classe, definite dal profilo di rischio-rendimento che le caratterizza, osservando i valori degli indicatori dei titoli appartenenti. In particolare, all'interno di ogni classe latente x (con $x = 1, \dots, K$), si possono valutare i valori medi degli indicatori, di cui

$$\bar{z}_{ix} = \sum_{s=1}^{c_i} z_i(s) p_{ix}(s) \quad (1.4)$$

rappresenta la media dell' i -esima variabile osservata appartenente alla classe x , ottenuta come media ponderata dei valori assunti dalla variabile i per le c_i categorie con pesi le probabilità condizionate di z_i data la classe x ¹. L'insieme dei valori medi di tutti gli indicatori considerati forniscono indicazioni fondamentali per definire un preciso profilo di rischio-rendimento dei titoli appartenenti alla classe x . In questo lavoro, la discriminazione delle K classi rispetto al profilo finanziario che le caratterizza è considerato come il punto di partenza per ulteriori analisi sui titoli classificati nei diversi gruppi (o sui gruppi stessi) e, per questo motivo, l'approccio metodologico che viene proposto risulta particolarmente rilevante sia a livello teorico che analitico.

¹ Nell'esempio riportato, si fa riferimento a variabili manifeste categoriche politomiche dove la probabilità della generica variabile z_i condizionata all'appartenenza alla classe x è data da

$$g(z_i(s) | x) = p_{ix}(s).$$

Si vedano i paragrafi **2.2.7** e **2.2.8**.

1.4 L'analisi della covarianza e dell'andamento di una serie storica finanziaria

L'analisi dell'andamento dinamico della serie storica di una variabile finanziaria (per esempio, il rendimento di un titolo) ricopre un ruolo di primissimo piano nell'econometria finanziaria.

Nell'impostazione media-varianza della teoria standard del portafoglio di Markowitz e sotto l'ipotesi di efficienza dei mercati (Fama, 1970), il processo generatore dei dati di una serie storica dei rendimenti di un titolo azionario $\{R_t\}_{t=1}^T$ è approssimabile ad un *random walk* e, per questa ragione, i valori futuri di R_t non sono prevedibili (Fama, 1965). Il valore atteso condizionato risulta essere

$$E(R_t | I_{t-1}) = \bar{R}$$

dove $I_{t-1} = \{R_{t-1}, R_{t-2}, \dots\}$ indica l'informazione disponibile fino al tempo $t - 1$. La volatilità, espressa in termini di varianza condizionata, è, anch'essa, ipotizzata come costante nel tempo:

$$Var(R_t | I_{t-1}) = \sigma^2.$$

Inoltre, in questa impostazione, i rendimenti sono assunti come indipendenti nel tempo, identicamente distribuiti e approssimativamente normali. Tuttavia, numerose analisi empiriche hanno evidenziato il fatto che queste ipotesi nella realtà fenomenica non vengono rispettate.

Per questa ragione, nella letteratura econometrica sono state introdotte alcune proposte metodologiche in grado di gestire la non stazionarietà e la non normalità delle serie storiche finanziarie, oltre alla dipendenza temporale esistente tra le osservazioni (modelli ARIMA, ARCH, ecc.). Più recentemente, è stato necessario anche affrontare altre due problematiche tipiche dei processi stocastici sottostanti le variabili finanziarie: il cambiamento dei regimi ed i break strutturali. Hamilton (1989) sottolinea come non curarsi di queste caratteristiche spesso presenti nelle

serie storiche finanziarie conduca ad uno scarso adattamento del modello ai dati osservati ed a una performance previsiva insoddisfacente.

Riprendendo il lavoro seminale di Goldfeld e Quandt (1973) sulla regressione con cambiamenti di regime, Hamilton propone una classe di modelli con parametri che variano nel tempo in base a processi stocastici di tipo markoviano (Hamilton e Raj, 2002). Questo approccio rappresenta una valida alternativa non lineare alle specificazioni ARIMA di Box e Jenkins o ai modelli di tipo ARCH e permette un miglioramento sostanziale nella precisione della previsione (Lamoureux e Lastrapes, 1993; Hamilton e Susmel, 1994).

Se l'andamento della serie dei rendimenti del titolo $\{R_t\}_{t=1}^T$ è soggetto a cambiamenti discreti nei regimi, ovvero a periodi in cui la dinamica della serie è marcatamente diversa, allora è utile considerare un modello che sfrutta il processo stocastico osservato per inferire su di un insieme di stati latenti discreti che caratterizzano un processo non osservabile sottostante in grado di spiegare le variazioni della serie osservata. Per esempio, i rendimenti dell'attività finanziaria possono essere caratterizzati da periodi di forte crescita, profondo declino o da fasi di stabilità ed i cambiamenti tra questi regimi possono essere modellati attraverso un processo markoviano che indaga la struttura di covarianza presente nella serie. Il ricorso ad una specificazione a catena di Markov offre non solo la possibilità di individuare le variazioni tra uno specifico stato ed un altro, ma permette anche di prevedere la dinamica futura della serie dei rendimenti.

Per queste ragioni, il *latent Markov model* (LMM) che rappresenta il modello a struttura latente particolarmente adatto per lo studio di serie storiche e dati longitudinali, comincia ad ottenere un discreto successo nell'analisi dei fenomeni finanziari (Rydén et al., 1998; Dias et al., 2008; Frühwirth-Schnatter e Kaufmann, 2008). Il LMM si basa sull'assunzione che gli stati latenti, interpretabili come i diversi regimi, non siano tra loro indipendenti ma seguano una struttura di tipo markoviano. In tal modo, il modello risulta piuttosto generale e flessibile e

permette di valutare la (auto)correlazione presente nei dati osservati attraverso una dipendenza che decresce a zero al crescere della loro distanza nel tempo.

La variabilità della serie storica osservata può quindi essere analizzata attraverso un LMM che ipotizza le osservazioni come appartenenti ad una di tante distribuzioni condizionali differenti, ognuna associata ad un diverso stato latente. In questo aspetto, il LMM è paragonabile all'analisi a classi latenti sia come modello di mistura di distribuzioni sia per l'efficiente metodo di classificazione che è in grado di fornire, raggruppando nello stesso stato latente le osservazioni temporali che presentano caratteristiche simili. Nel LMM, i gruppi, definiti stati latenti ed equivalenti alle classi dei modelli a classi latenti, sono interpretabili come i diversi regimi che caratterizzano l'andamento dinamico della serie oggetto di analisi. Inoltre, modellando le variazioni nei regimi tramite una mistura di distribuzioni (gaussiane, nel caso di $\{R_t\}_{t=1}^T$), il LMM permette di ovviare alla non normalità della serie finanziaria, riuscendo a catturare quasi ogni allontanamento dalla distribuzione normale (Dias et al., 2008).

Negli ultimi anni, nella letteratura econometrica, lo studio del cambiamento dei regimi di una o più serie storiche finanziarie ha riscosso parecchio successo. L'inclusione di una componente latente caratterizzata da un processo markoviano nei modelli autoregressivi o per l'analisi dell'eteroschedasticità condizionata ha permesso di ottenere risultati piuttosto interessanti. Tuttavia, l'attenzione dei ricercatori si è rivolta principalmente allo studio della componente di rischio e, quindi, all'estensione della specificazione dei modelli per la varianza condizionata. In questo ambito, i principali riferimenti sono il modello ARCH con cambiamenti di regime, detto *SWARCH* e proposto da Hamilton e Susmel (1994) o i modelli GARCH con cambiamenti markoviani, *MS-GARCH* proposti da Duekel (1997) e ripresi, ad esempio, da Haas et al. (2004). L'approfondimento dello studio dei modelli per l'eteroschedasticità condizionata con cambiamenti di regime ha permesso di stimare i parametri *time-variant* e di individuare i periodi

caratterizzati da crisi finanziarie. Tuttavia, questo maggiore interesse per la componente di volatilità del processo stocastico si è tradotto nel trascurare una indagine rigorosa delle medie condizionate, ovvero della componente relativa al rendimento atteso. Queste o non vengono di fatto analizzate oppure risultano non significativamente diverse tra loro nei vari regimi (si veda, ad esempio, Ang e Bekaert, 2002). Tali risultati ipotizzano implicitamente che il rendimento atteso durante i periodi caratterizzati da crisi finanziarie sia il medesimo rispetto ai periodi di stabilità.

Per ovviare a questa forte incongruenza rispetto alla teoria economica, si è deciso di ricorrere al LMM che, nonostante possa essere considerato come un caso particolare dei modelli sopracitati, permette una misura piuttosto accurata delle medie condizionate e del profilo di rischio-rendimento rispetto ai vari regimi dei mercati finanziari. In particolare, riesce a discriminare le osservazioni temporali in base al rendimento medio e, oltre a distinguere i periodi di alta volatilità da quelli più stabili, consente anche uno studio approfondito delle fluttuazioni che caratterizzano i periodi di turbolenza finanziaria.

1.5 L'importanza dei modelli a struttura latente nello studio dei fenomeni finanziari

I modelli a struttura latente permettono di ottenere i quattro scopi descritti nei paragrafi precedenti che possono essere così sintetizzati:

- (i) trattare diverse tipologie di variabili, comprese le trasformate categoriche introdotte nel paragrafo **1.1** e ricavare una misura del vettore X ricorrendo ad un insieme di variabili osservate Z ;
- (ii) classificare le unità statistiche in base alle loro caratteristiche riferendosi ad un approccio *model-based* che permette una procedura rigorosa nella determinazione del numero delle classi ed una migliore performance classificatoria;
- (iii) determinare il profilo (latente) di rischio-rendimento delle attività finanziarie;
- (iv) definire il processo stocastico non osservabile sottostante ad una o più serie storiche attraverso l'analisi della struttura di covarianza osservabile nei dati.

Per queste ragioni, il lavoro si prefigge lo scopo di sfruttare le ottime proprietà dei modelli a struttura latente per contribuire alla ricerca di soluzioni ai problemi metodologici legati ai fenomeni finanziari descritti in questo capitolo. L'introduzione di un rigoroso approccio statistico nell'analisi di questi fenomeni comporta un contributo sostanziale e, in taluni casi, una assoluta novità nello studio delle variabili finanziarie e nelle decisioni di investimento.

Il ricorso a modelli a struttura latente in campo finanziario è in letteratura particolarmente limitato nonostante gli ottimi requisiti di cui dispongono. Inoltre, nessun ricercatore ha mai messo in congiunzione questa metodologia statistica con la teoria standard del portafoglio.

2. Metodi statistici per le variabili latenti

Questo capitolo viene dedicato alla rassegna dei principali metodi statistici a variabili latenti presenti in letteratura. Si comincia la trattazione partendo dagli aspetti teorici generali presentati nel paragrafo **2.1**.

2.1 Gli aspetti teorici

Da un campione di n oggetti (o individui) vengono rilevati, per ogni oggetto, i valori di p variabili. In quanto direttamente osservabili, queste variabili prendono il nome di *manifeste* e possono essere sia continue che categoriche².

La singola osservazione riguardante l' i -esima variabile manifesta per il generico oggetto h -esimo appartenente al campione (con $i = 1, \dots, p$ e $h = 1, \dots, n$) viene indicata con z_{ih} . La rilevazione delle p variabili manifeste per l' h -esima unità del campione formano il vettore p -dimensionale di variabili casuali Z_h . In riferimento alla singola variabile manifesta i -esima, si ottengono p vettori n -dimensionali denotati con Z_i .

Pertanto, si definisce la matrice di dimensione $(n \times p)$ delle variabili manifeste Z come

$$Z = \begin{bmatrix} z_{11} & \dots & z_{i1} & \dots & z_{p1} \\ \dots & \dots & \dots & \dots & \dots \\ z_{1h} & \dots & z_{ih} & \dots & z_{ph} \\ \dots & \dots & \dots & \dots & \dots \\ z_{1n} & \dots & z_{in} & \dots & z_{pn} \end{bmatrix}.$$

² Nello studio dei fenomeni finanziari si è soliti considerare solo variabili continue e, spesso, l'importanza della informazione fornita da variabili qualitative sia in qualità di indicatori che come covariate (variabili esterne esogene) viene trascurata. Il ricorso ai metodi a variabili latenti permette di poter includere nell'analisi variabili con diversa tipologia sia tra le variabili esplicative sia tra quelle esterne.

Con X si indica una variabile latente composta, nel caso categorico, da K classi esaustive e mutuamente esclusive che raggruppano le osservazioni sulla base di caratteristiche comuni. Nel caso in cui la variabile latente sia continua, come ad esempio nel caso dell'analisi fattoriale, allora lo spazio latente non è più suddiviso in classi discrete ma viene rappresentato da un segmento sul quale si posizionano le unità statistiche che presentano il medesimo vettore osservato Z_i . Per questa ragione, oltre all'analisi fattoriale, i modelli statistici con variabili latente continue si definiscono a *tratto* latente.

Nella tabella 2.1 vengono riportati i diversi metodi a struttura latente a cui è possibile ricorrere in corrispondenza della natura continua o categorica delle variabili manifeste e di quelle latente. Ad esempio, quando entrambe le variabili sono continue, il metodo statistico a cui ci si riferisce è l'*analisi fattoriale*. Questa metodologia permette di ottenere una riduzione del numero dei dati osservati attraverso una sintesi degli stessi ottenuta con l'individuazione di uno o più fattori non osservabili che permettono di riassumere il fenomeno analizzato. Per l'enorme notorietà e diffusione di questi metodi, in questo lavoro, si evita la loro trattazione. Per un approfondimento sul tema si rimanda alla cospicua letteratura esistente, ad esempio Bartholomew et al. (2002).

Quando entrambe le variabili sono di natura categorica, l'*analisi a classi latente* risulta essere la metodologia più opportuna per valutare le associazioni esistenti tra le variabili osservate attraverso la variabile latente X ed ottenere una classificazione delle unità statistiche nelle K classi. Dell'analisi a classi latente viene data una trattazione accurata degli sviluppi esistenti in letteratura nei paragrafi **2.2** e **2.3**.

Il modello statistico a struttura latente che spiega la relazione tra variabili manifeste continue e variabili latente categoriche viene identificato con il nome di *analisi a profilo latente*. Un'analisi di questo tipo è particolarmente indicata per lo

studio di variabili finanziarie che, di norma, sono caratterizzate da valori reali. I modelli a profilo latente sono descritti nel paragrafo **2.4**.

Nel caso in cui le variabili manifeste siano categoriche e le variabili latenti continue, la metodologia statistica di riferimento, molto rinomata nel campo delle scienze sociali, è nota come *analisi a tratto latente*, altrimenti conosciuta con il nome di *Item Response Theory* (IRT). Questi modelli sono particolarmente indicati per analisi di tipo psicométrico ma risultano poco pratici agli scopi di questo lavoro. Ciononostante, se ne dà una breve introduzione nel paragrafo **2.5**.

Infine, quando si analizza una serie storica o dei dati longitudinali, al vettore delle variabili manifeste Z viene aggiunta la dimensione temporale $\{Z_t\}_{t=1}^T$. Per questa tipologia di dati, può essere particolarmente interessante analizzare la struttura di covarianza sottostante la serie storica attraverso un modello a struttura latente in grado di determinare un processo stocastico discreto non osservabile ottenuto sulla base dell'andamento della serie osservata. Il modello a cui si fa riferimento in queste situazioni è il *latent Markov model* di cui si parla nel paragrafo **2.6**.

		Variabili manifeste		
		<i>Continue</i>	<i>Categoriche</i>	<i>Temporali</i>
Variabili latenti	<i>Continue</i>	Analisi fattoriale	Analisi a tratto latente	
	<i>Categoriche</i>	Analisi a profilo latente	Analisi a classi latenti	Latent Markov model

Tabella 2.1: Metodi a variabili latenti

2.1.1 Il modello a struttura latente

In questo paragrafo si illustrano gli aspetti teorici principali dei modelli a struttura latente nel caso in cui sia Z che X siano variabili continue. Tuttavia, il passaggio a modelli per variabili categoriche o miste è pressoché immediato.

Essendo le sole direttamente osservabili, le variabili manifeste Z sono le uniche informazioni che si possiedono sul fenomeno oggetto di indagine, perciò qualsiasi inferenza deve basarsi sulla distribuzione di densità congiunta

$$f(Z) = \int_{D_X} h(X)g(Z|X)dX \quad (2.1)$$

dove $h(X)$ è la distribuzione di probabilità a priori di X , $g(Z|X)$ è la distribuzione condizionata di Z dato X e con D_X si indica il dominio della variabile latente X .

Il principale interesse risiede nel ricavare informazioni sulla variabile latente X dopo aver osservato le variabili Z . Per il teorema di Bayes, questa informazione si ottiene dalla distribuzione condizionata (a posteriori)

$$h(X|Z) = \frac{h(X)g(Z|X)}{f(Z)}.$$

Per ottenere una specificazione completa di $h(X|Z)$ è necessario imporre qualche assunzione, in quanto, nonostante si possa stimare $f(Z)$, le funzioni h e g non possono essere unicamente determinate dall'equazione (2.1).

Il punto cruciale di ogni modello a struttura latente consiste nell'assumere che tutte le associazioni esistenti tra le variabili manifeste Z siano spiegate dal loro dipendere dalle variabili latenti X ; in altri termini, le variabili Z risultano statisticamente incorrelate se le X sono tenute fisse. In letteratura, questa proprietà è nota come l'assunzione di indipendenza locale. È però fuorviante pensare che questa sia un'assunzione che possa essere testata empiricamente, in quanto non c'è modo di fissare le variabili X . Per questa ragione, sarebbe più corretto da un punto di vista formale riferirsi all'*assioma* di indipendenza locale piuttosto che

all'assunzione (Bartholomew e Knott, 1999). È compito del ricercatore quello di determinare il numero di variabili X sufficienti a spiegare le relazioni di dipendenza tra le Z . Perciò, nel caso di variabili latenti categoriche, il numero di classi latenti K di cui la variabile X è composta deve essere scelto in modo da rendere le variabili osservate indipendenti l'una dall'altra. L'assunzione di indipendenza locale è verificata quando si hanno

$$g(Z | X) = \prod_{i=1}^p g(z_i | X)$$

e

$$f(Z) = \int h(X) \prod_{i=1}^p g(z_i | X) dX . \quad (2.2)$$

Una parte importante dell'analisi riguarda la ricerca del valore di K più piccolo per cui la rappresentazione nell'equazione (2.2) risulta adeguata.

2.1.2 La stima del modello a struttura latente: l'algoritmo EM

La stima dei modelli a struttura latente avviene ricorrendo alla procedura iterativa dell'algoritmo *Expectation-Maximization* (EM) sviluppato da Dempster, Laird e Rubin (1977). L'algoritmo EM viene usato per trovare la stima di massima verosimiglianza dei parametri nei modelli probabilistici in cui sono presenti delle variabili latenti.

Il metodo iterativo dell'EM alterna l'esecuzione di un passo *expectation* (E), che calcola il valore atteso della funzione di log-verosimiglianza rispetto alla stima attuale della distribuzione delle variabili latenti, ed un passo *maximization* (M), che calcola i parametri che massimizzano la log-verosimiglianza attesa trovata al passo E. Questi parametri sono quindi usati per determinare la distribuzione delle variabili latenti nel passo E successivo.

In pratica, data la funzione di verosimiglianza $L(\theta; Z, X)$, dove θ è il vettore dei parametri, Z i dati osservati e X rappresenta la variabile latente non osservabile

(o i valori mancanti), la stima di massima verosimiglianza viene determinata dalla funzione marginale dei dati osservati $L(\theta; Z)$. Tuttavia, questa quantità è spesso intrattabile. L'algoritmo EM cerca di trovare la stima di massima verosimiglianza iterando i due passi nel modo seguente:

(i) passo E: calcolo del valore atteso della funzione di log-verosimiglianza rispetto alla distribuzione condizionata delle X date le variabili Z e alla stima attuale dei parametri $\theta^{(t)}$:

$$Q(\theta | \theta^{(t)}) = E_{X|Z, \theta^{(t)}} [\log L(\theta; Z, X)];$$

(ii) passo M: stabilire i parametri che massimizzano la seguente quantità:

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

Nonostante l'iterazione EM non faccia decrescere il valore della funzione di verosimiglianza dei dati osservati, non vi è garanzia che la sequenza converga ad uno stimatore di massima verosimiglianza. Perciò, per distribuzioni multimodali come i modelli a struttura latente, l'algoritmo EM può convergere ad un massimo locale della funzione $L(\theta; Z)$ piuttosto che al massimo globale. Questa convergenza dipende dai valori iniziali da cui si fa partire la procedura iterativa dell'algoritmo. Per ovviare a questo inconveniente, è necessario considerare diversi valori iniziali, facendo partire la procedura da più stime iniziali dei parametri $\theta^{(t)}$ differenti, scelte casualmente.

L'algoritmo EM è particolarmente utile quando la funzione di verosimiglianza appartiene alla famiglia esponenziale: il passo E diventa la somma dei valori attesi di statistiche sufficienti mentre il passo M comporta la massimizzazione di una funzione lineare. In questo caso, di solito, è possibile derivare degli aggiornamenti in forma chiusa per ogni passo della procedura.

2.2 L'analisi a classi latenti

2.2.1 Il modello a classi latenti con variabili manifeste dicotomiche e una variabile latente dicotomica

Il modello a classi latenti con variabili manifeste dicotomiche ed una variabile latente anch'essa dicotomica rappresenta il caso più semplice tra i vari metodi a struttura latente.

In questa specificazione, le p variabili manifeste possono assumere solo due valori: $z_i = 0$ o $z_i = 1$ per ogni $i = 1, \dots, p$. Lo scopo dell'analisi è di valutare se una variabile latente con due classi è in grado di spiegare le relazioni tra le variabili manifeste. In altri termini, se è possibile dividere la popolazione dalla quale il campione è stato idealmente estratto in due sottopopolazioni tali che le Z risultino mutualmente indipendenti in ogni gruppo. In pratica, ci si interroga se $K = 2$ permette di verificare l'assunzione di indipendenza locale.

La probabilità a priori della variabile latente, indicata da $h(X)$ nell'equazione (2.1), può essere scritta come

$$h(1) = P(X = 1) = \pi_1^X \text{ e } h(2) = P(X = 2) = \pi_2^X = 1 - \pi_1^X .$$

Nella notazione usata, l'apice indica la variabile mentre il pedice denota la categoria, quindi π_1^X rappresenta la probabilità (a priori) della prima classe della variabile latente X . Nel seguito, per maggiore semplicità e chiarezza, si tenderà ad omettere l'apice a meno che non risulti strettamente necessario.

La distribuzione condizionata delle z_i data X è la stessa di una variabile aleatoria di Bernoulli

$$g(z_i | X) = P(z_i | X) = \pi_{ix}^{Z|X} = p_{ix}^{z_i} (1 - p_{ix})^{1-z_i} \text{ con } z_i = 0, 1 \text{ e } x = 1, 2.$$

$\pi_{ix}^{Z|X}$ indica la probabilità delle variabili manifeste z_i condizionata a $X = x$ e p_{ix} è la probabilità che $z_i = 1$ quando la variabile latente è x . Le espressioni delle

probabilità condizionate denotate in questo modo ($\pi_{ilx}^{Z|X}$) saranno riprese nel paragrafo **2.3**.

Nell'analisi a classi latenti, considerando variabili di tipo categorico, l'integrale riportato in equazione (2.1) diventa una somma

$$f(Z) = \sum_{x=1}^K \pi_x^X \prod_{i=1}^p \pi_{ilx}^{Z|X} = \pi_1 \prod_{i=1}^p p_{i1}^{z_i} (1-p_{i1})^{1-z_i} + (1-\pi_1) \prod_{i=1}^p p_{i2}^{z_i} (1-p_{i2})^{1-z_i}. \quad (2.3)$$

La stima dei parametri (π_x e π_{ilx}) avviene con il metodo della massima verosimiglianza e, applicando un test sulla bontà del modello tra la distribuzione di probabilità riportata in equazione (2.3) e la distribuzione delle frequenze osservate delle Z , si ottiene una misura dell'adattamento ai dati osservati. Se la bontà del modello non risulta soddisfacente allora si può considerare la possibilità di aggiungere una o più classi.

Come passo conclusivo, si può classificare ogni oggetto (individuo) osservato ad una classe latente. Questa allocazione non può essere effettuata con certezza ma avviene in termini di probabilità, ovvero valutando quale sia la classe di appartenenza più probabile attraverso il confronto delle distribuzioni a posteriori

$$h(1|Z) = P(X=1|Z) = \frac{\pi_1^X \left(\prod_{i=1}^p p_{i1}^{z_i} (1-p_{i1})^{1-z_i} \right)}{f(Z)} = \left[1 + \left(\frac{1-\pi_1^X}{\pi_1^X} \right) \exp \sum_{i=1}^p \left\{ z_i \ln \frac{p_{i2}}{p_{i1}} + (1-z_i) \ln \frac{1-p_{i2}}{1-p_{i1}} \right\} \right]^{-1} \quad (2.4)$$

Si assegna un oggetto alla classe 1 se $h(1|Z) > h(2|Z)$.

2.2.2 Il modello a classi latenti con variabili manifeste dicotomiche

Il modello a classi latenti descritto nel paragrafo **2.2.1** può essere facilmente generalizzato al caso in cui vi siano $K > 2$ classi latenti.

Il numero di classi latenti può essere determinato in anticipo sulla base di informazioni preliminari che si dispongono circa le caratteristiche della popolazione oggetto di indagine ed, in particolare, su come possa essere suddivisa in K sottopopolazioni. Questo approccio è definito *confermativo* in quanto l'interesse del ricercatore ricade principalmente sull'analisi della bontà del modello stimato. Tuttavia, in molti casi K è ignoto ed una delle principali ragioni che spingono il ricercatore a ricorrere ad un modello a classi latenti è proprio la possibilità di suddividere i dati in K gruppi non noti e non osservabili. Questo approccio è chiamato *esplorativo* e, sulla base di alcuni criteri statistici in grado di aiutare il ricercatore nella scelta del numero di classi, si ricerca il modello migliore possibile.

La distribuzione congiunta nel caso di K classi latenti è

$$f(Z) = \sum_{x=1}^K \pi_x \prod_{i=1}^p \pi_{i|x} = \sum_{x=1}^K \pi_x \prod_{i=1}^p p_{ix}^{z_i} (1 - p_{ix})^{1-z_i} \quad (2.5)$$

con $\sum_{x=1}^K \pi_x = 1$.

La probabilità a posteriori che un oggetto appartenga alla classe generica x dato il vettore di osservazioni Z diventa

$$h(x|Z) = \pi_x \prod_{i=1}^p p_{ix}^{z_i} (1 - p_{ix})^{1-z_i} / f(Z) \quad (2.6)$$

per $x = 1, \dots, K$. Si può usare l'equazione (2.6) per assegnare ogni oggetto alla classe latente con probabilità a posteriori maggiore come descritto nel paragrafo **2.2.6**.

2.2.3 La stima dei parametri

La stima dei parametri si ottiene massimizzando la funzione di log-verosimiglianza derivata dalla equazione (2.5) usando routine di ottimizzazione standard. In particolare, la stima può avvenire ricorrendo alla tecnica Newton-

Raphson oppure all'algorithmo EM (Dempster et al. 1977). Questo ultimo metodo risulta meno dispendioso in termini di elaborazione e riesce a prevenire alcuni problemi sostanziali (Bartholomew et al., 2002). Per queste ragioni l'algorithmo EM (Expectation-Maximization) è spesso preferito per la stima dei modelli a classi latenti.

Dall'equazione (2.5) si ottiene la log-verosimiglianza per un campione casuale di dimensione n

$$LL = \sum_{h=1}^n \log \left\{ \sum_{x=1}^K \pi_x \prod_{i=1}^p p_{ix}^{z_{ih}} (1-p_{ix})^{1-z_{ih}} \right\}$$

che può essere massimizzata sotto il vincolo $\sum_{x=1}^K \pi_x = 1$, ottenendo così il massimo non vincolato di

$$\phi = LL + \theta \sum_{x=1}^K \pi_x$$

dove θ è un moltiplicatore indefinito. Le derivate parziali sono

$$\frac{\partial \phi}{\partial \pi_x} = \sum_{h=1}^n \left\{ \prod_{i=1}^p p_{ix}^{z_{ih}} (1-p_{ix})^{1-z_{ih}} / f(Z_h) \right\} + \theta = \sum_{h=1}^n \{g(Z_h | x) / f(Z_h)\} + \theta$$

per $x = 1, \dots, K$, dove $g(Z_h | x)$ è la probabilità congiunta di Z_h per un oggetto appartenente alla classe x , e

$$\frac{\partial \phi}{\partial p_{ix}} = \sum_{h=1}^n \pi_x \frac{\partial}{\partial p_{ix}} g(Z_h | x) / f(Z_h) \quad (2.7)$$

per $i = 1, \dots, p$ e $x = 1, \dots, K$, dove

$$\begin{aligned} \frac{\partial g(Z_h | x)}{\partial p_{ix}} &= \frac{\partial}{\partial p_{ix}} \exp \sum_{i=1}^p \{z_{ih} \log p_{ix} + (1-z_{ih}) \log(1-p_{ix})\} = \\ &= g(Z_h | x) \left\{ \frac{z_{ih}}{p_{ix}} - \frac{1-z_{ih}}{1-p_{ix}} \right\} = \frac{(z_{ih} - p_{ix}) g(Z_h | x)}{p_{ix}(1-p_{ix})} \end{aligned}$$

da cui si ricava che

$$\frac{\partial \phi}{\partial p_{ix}} = \frac{\pi_x}{p_{ix}(1-p_{ix})} \sum_{h=1}^n (z_{ih} - p_{ix}) g(Z_h | x) / f(Z_h). \quad (2.8)$$

Queste equazioni possono essere semplificate, esprimendo la (2.7) e la (2.8) rispetto alle probabilità a posteriori $h(x | Z)$. Per il teorema di Bayes,

$$h(x | Z_h) = \pi_x g(Z_h | x) / f(Z_h)$$

che sostituita nella (2.7) e posta uguale a zero, si ottiene

$$\sum_{h=1}^n h(x | Z_h) = -\theta \pi_x.$$

Sommando entrambi i lati rispetto a x e riprendendo il vincolo $\sum_{x=1}^K \pi_x = 1$ si ottiene che $\theta = -n$, e quindi la prima equazione stimata risulta essere

$$\hat{\pi}_x = \sum_{h=1}^n h(x | Z_h) / \theta. \quad (2.9)$$

La seconda è

$$\sum_{h=1}^n (z_{ih} - p_{ix}) h(x | Z_h) / p_{ix}(1-p_{ix}) = 0$$

da cui

$$\hat{p}_{ix} = \frac{\sum_{h=1}^n z_{ih} h(x | Z_h)}{\sum_{h=1}^n h(x | Z_h)} = \frac{\sum_{h=1}^n z_{ih} h(x | Z_h) / n \hat{\pi}_x}{\sum_{h=1}^n h(x | Z_h)}. \quad (2.10)$$

Nonostante le equazioni ricavate sembrano avere una forma semplice, in realtà bisogna ricordare che $h(x | Z_h)$ è una funzione piuttosto complicata data da

$$h(x | Z_h) = \pi_x \prod_{i=1}^p p_{ix}^{z_{ih}} (1-p_{ix})^{1-z_{ih}} / \sum_{x=1}^K \pi_x \prod_{i=1}^p p_{ix}^{z_{ih}} (1-p_{ix})^{1-z_{ih}}. \quad (2.11)$$

Ad ogni modo, se $h(x | Z_h)$ fosse nota sarebbe facile risolvere la (2.9) e la (2.10) rispetto a π_x e p_{ix} . L'algoritmo EM approfitta di questo fatto procedendo nel seguente modo:

- (i) sceglie un insieme di valori iniziali per le probabilità a posteriori $h(x|Z_h)$;
- (ii) usa la (2.9) e la (2.10) in modo da ottenere una prima approssimazione di $\hat{\pi}_x$ e \hat{p}_{ix} (passo E);
- (iii) sostituisce queste stime nella (2.11) per ottenere una nuova stima di $h(x|Z_h)$ (passo M);
- (iv) ritorna al punto (ii) per ottenere una seconda approssimazione dei parametri e continua il ciclo finché non si raggiunge la convergenza.

La soluzione che si raggiunge è identificabile come un punto di massimo locale. È noto che modelli di questo tipo possono avere più punti di massimo e che all'aumentare di K il numero di questi punti tende ad aumentare. Al contrario, aumentando l'ampiezza del campione n , i punti di massimo locale tendono a diminuire (Aitkin et al., 1981). Una soluzione adottata per ovviare al problema di interpretare un punto di massimo locale per un punto di massimo globale è quella di far convergere l'algoritmo partendo da valori iniziali diversi.

Un modo ragionevole di scegliere i valori iniziali al punto (i) è di classificare le osservazioni nelle classi latenti sulla base del loro punteggio totale $\sum_{i=1}^p z_i$, cioè imporre $h(x|Z_h) = 1$ se Z_h viene assegnata alla classe x e $h(x|Z_h) = 0$ altrimenti.

Oltre a fornire le stime dei parametri, il metodo EM fornisce anche le probabilità a posteriori che esprimono l'appartenenza di ogni oggetto ad una data classe latente.

2.2.4 La stima degli errori standard

Per ottenere una stima degli errori standard è necessario imporre un vincolo. Poiché la somma delle π_x è vincolata ad essere 1, si può eliminare una probabilità a priori, per esempio $\pi_{x=1}$, ottenendola per differenza:

$\pi_{x=1} = 1 - \pi_{x=2} - \pi_{x=3} - \dots - \pi_{x=K}$. In questo modo, le derivate seconde e le derivate miste di LL possono essere espresse in termini di distribuzione a posteriori:

$$\frac{\partial^2 LL}{\partial \pi_x \partial \pi_{x'}} = - \frac{\sum_{h=1}^n \{h(x|Z_h) - h(1|Z_h)\} \{h(x'|Z_h) - h(1|Z_h)\}}{\pi_x \pi_{x'}}$$

$$\frac{\partial^2 LL}{\partial p_{ix} \partial p_{i'x'}} = \frac{\sum_{h=1}^n (z_{ih} - p_{ix})(z_{i'h} - p_{i'x'}) h(x|Z_h) \{\delta_{xx'}(1 - \delta_{ii'}) - h(x'|Z_h)\}}{p_{ix}(1 - p_{ix}) p_{i'x'}(1 - p_{i'x'})}$$

dove

$$\delta_{rs} = \begin{cases} 1 & \text{se } r = s \\ 0 & \text{altrimenti} \end{cases}$$

$$\frac{\partial^2 LL}{\partial \pi_x \partial p_{ix'}} = \sum_{h=1}^n (z_{ih} - p_{ix}) h(x|Z_h) \{\delta_{xx'} - h(x'|Z_h)\}$$

per $x, x' = 2, \dots, K$ e $i, i' = 1, \dots, p$.

La matrice di varianza-covarianza asintotica delle stime è quindi l'inversa del valore atteso della matrice di dimensione $(K-1)(p+1) \times (K-1)(p+1)$ delle derivate ricavate sopra cambiate di segno. Il calcolo esatto dei valori attesi comporta la somma di 2^p possibili combinazioni dei punteggi osservabili di Z . Perciò, se il numero di variabili manifeste p è elevato, i termini da calcolare diventano estremamente numerosi e la grandezza di ogni termine così piccola che effettuare calcoli accurati risulta impossibile. In questo caso, il valore atteso può essere approssimato considerando l'inversa della matrice delle derivate seconde osservate. Inoltre, gli errori standard asintotici possono essere approssimazioni piuttosto scadenti soprattutto quando le stime dei parametri risultano vicine ai valori limite di 0 o 1 (De Menezes, 1999).

2.2.5 L'analisi della bontà del modello

Una volta stimati i parametri, un passo fondamentale nell'analisi a classi latenti consiste nel valutare la bontà di adattamento del modello che permette la scelta del numero di classi K per cui l'assunzione di indipendenza locale risulta soddisfatta.

Un modo per analizzare la bontà del modello a classi latenti è quello di confrontare le frequenze osservate dei valori individuali con quelle teoriche previste dal modello. In questo caso, si ricorre al test chi-quadro standard oppure al rapporto di verosimiglianza (Goodman, 1978) definito come

$$L^2 = 2 \sum_{i=1}^p f_i \log(\hat{F}_i / f_i)$$

dove f_i sono le frequenze osservate e \hat{F}_i le frequenze teoriche attese. Nel caso in cui $\hat{F}_i = f_i$ per ogni i , il modello ha un adattamento perfetto e il valore di L^2 risulta nullo. L'allontanamento da zero indica un peggioramento nella bontà del modello, in questo senso il valore di L^2 misura la mancanza di adattamento ai dati. La statistica L^2 si distribuisce asintoticamente come un chi-quadro con un numero di gradi di libertà (df) pari alla differenza tra il numero totale di combinazioni (celle nella tabella a multi-entrata) 2^p e il numero di parametri ($NPar = K(p+1)$) meno 1, da cui si ricava che

$$df = 2^p - K(p+1) + 1.$$

La bontà del modello a classi latenti stimato si misura accertando la non significatività del test utilizzando i valori teorici della distribuzione chi-quadro.

Quando p è grande, anche un campione con dimensione n elevata può far incorrere in celle con valori attesi prossimo allo zero e, di conseguenza, la distribuzione della statistica test non è più approssimabile ad un chi-quadro.

In presenza di dati sparsi è quindi preferibile ricorrere all'approccio *bootstrap* per stimare i livelli di soglia (*p-value*) (Langeheine et al., 1996) o valutare i residui bivariati (Bartholomew e Knott, 1999).

Un ulteriore metodo per ovviare a questo problema è quello di considerare un criterio di informazione che è anche un ottimo metodo per coniugare due aspetti fondamentali di ogni modello statistico: l'adattamento ai dati e la parsimonia. I criteri di informazione, come l' Akaike Information Criterion (AIC) o il Bayesian Information Criterion (BIC), sono molto utili per il confronto di modelli annidati. I metodi più usati nei modelli a classi latenti sono il BIC basato sul rapporto di verosimiglianza (Raftery, 1986)

$$BIC_{L^2} = L^2 - \log(n)df$$

di cui, una versione più generale è data dal BIC basato sulla funzione di log-verosimiglianza (*LL*) e sul numero di parametri (*NPar*) invece che sulla statistica L^2 e sui gradi di libertà (*df*):

$$BIC_{LL} = -2LL + \log(n)NPar ;$$

la statistica AIC basata sul rapporto di verosimiglianza (Akaike, 1974)

$$AIC_{L^2} = L^2 - 2df$$

di cui la versione basata sulla funzione di log-verosimiglianza risulta

$$AIC_{LL} = -2LL + 2NPar .$$

Bozdogan (1987) ha fornito una versione della statistica AIC che penalizza maggiormente la complessità del modello rispetto alla versione standard:

$$CAIC_{L^2} = L^2 - [(\log n) + 1]df$$

oppure

$$CAIC_{LL} = -2LL + [(\log n) + 1]NPar .$$

Il Consistent Akaike Information Criterion (CAIC) permette di ovviare al problema di sovrastima del numero delle classi latenti che notoriamente viene imputato alla statistica AIC (Dias e Vermunt, 2007).

2.2.6 L'analisi a posteriori

L'analisi a posteriori di un modello a classi latenti riguarda la classificazione degli oggetti nell'opportuna classe latente e si ricava ricorrendo alle osservazioni dei valori delle variabili manifeste: in letteratura l'insieme di questi valori relativi all'oggetto h -esimo è noto come *response pattern*. Tutte le informazioni rilevanti sono contenute nella distribuzione di probabilità a posteriori $h(x|Z)$ e, analogamente al caso di un modello con due classi presentato in equazione (2.4), l'allocatione nella classe x risulta più probabile rispetto alla classe x' se $h(x|Z) > h(x'|Z)$ ovvero se $h(x|Z)/h(x'|Z) > 1$, il che è vero se

$$\frac{\pi_x}{\pi_{x'}} \exp \sum_{i=1}^p \{ [z_i \log p_{ix} + (1 - z_i) \log(1 - p_{ix})] - [z_i \log p_{ix'} + (1 - z_i) \log(1 - p_{ix'})] \} > 1.$$

L'ultima diseguaglianza è equivalente a

$$\sum_{i=1}^p z_i \text{logit} p_{ix} + \sum_{i=1}^p \log(1 - p_{ix}) + \log \pi_x > \sum_{i=1}^p z_i \text{logit} p_{ix'} + \sum_{i=1}^p \log(1 - p_{ix'}) + \log \pi_{x'} \quad (2.12)$$

dove $\text{logit} p_{ix} = p_{ix} / (1 - p_{ix})$. Questo risultato sottolinea come la regola di allocatione dipenda da Z in modo lineare; in altri termini, la (2.12) contiene tutta l'informazione rilevante nei dati riguardo alla variabile latente X .

2.2.7 L'interpretazione delle classi latenti

Dopo aver asserito che un modello a classi latenti è in grado di spiegare le relazioni esistenti nei dati per un qualche valore K , può essere necessario interpretare, ovvero etichettare, le classi. Nel caso si stimi il modello a classi

latenti per scopi confermativi, di solito, le classi sono definite a priori e, quindi, sono già note le caratteristiche intrinseche che differenziano una determinata classe latente da un'altra. Un esempio classico è quello riferito a prove di abilità somministrate a studenti (modello con 2 classi latenti) nel quale un rispondente sarà assegnato alla classe dei *masters* nel caso egli dimostri una certa padronanza della materia e ai *non-masters* nel caso risponda in maniera errata alle domande sottoposte (si veda l'esempio di Goodman (1978)).

Quando lo scopo del modello a classi latenti è esplorativo, le K classi risultanti possono essere etichettate sulla base delle caratteristiche empiriche che emergono dalle stime. Essenzialmente, ci si chiede cosa abbiano in comune gli oggetti assegnati ad una classe che li differenziano dai membri di altre classi. Un modo per rispondere a questa domanda è quello di guardare le stime delle probabilità condizionate p_{ix} : per una determinata classe latente x , osservare per quali variabili i le \hat{p}_{ix} risultano vicine ad uno o a zero. Il primo caso rappresenta un attributo che molto probabilmente i membri della classe x possiedono, al contrario, il secondo indica una caratteristica che appartiene raramente ai membri della classe x . Basandosi sulle stime \hat{p}_{ix} , è possibile riconoscere le caratteristiche delle K classi latenti e darne un'interpretazione utile ai fini dell'analisi. In particolare, ci si può riferire alle medie delle variabili manifeste per ogni classe latente $\bar{z}_{i|x}$ proposte nell'equazione (1.4) dove, nel caso dicotomico, $c_i = 2$.

2.2.8 Il modello a classi latenti con variabili manifeste politomiche

L'estensione a modelli a classi latenti con variabili manifeste politomiche e non ordinabili si basa sugli spunti di Goodman (1978), in particolare per quanto riguarda la teoria sulla stima.

Nel caso in cui vi siano più di due categorie, la variabile z_i diventa un vettore Z_i con c_i elementi in cui il singolo elemento $z_i(s)$ risulta uguale a 1 se l'oggetto (la

risposta) è nella categoria s della variabile i e pari a zero altrimenti, per $s=1, 2, \dots, c_i$. Quindi, $\sum_s z_i(s)=1$ e il vettore dei valori delle p variabili manifeste per un determinato oggetto (il vettore di risposte per un individuo) è $Z'=(Z_1', Z_2', \dots, Z_p')$.

Le probabilità condizionate $p_{ix}(s)$ esprimono la probabilità che l'oggetto appartenente alla classe latente x sia nella categoria s della variabile manifesta i . La funzione di probabilità congiunta di Z è quindi

$$f(Z) = \sum_{x=1}^K \pi_x \prod_{i=1}^p \prod_{s=1}^{c_i} (p_{ix}(s))^{z_i(s)}, \quad (2.13)$$

e la distribuzione a posteriori è

$$h(x|Z) = \pi_x \prod_{i=1}^p \prod_{s=1}^{c_i} (p_{ix}(s))^{z_i(s)} / f(Z). \quad (2.14)$$

2.2.9 La stima dei parametri

Come nel caso con dati dicotomici, la funzione di log-verosimiglianza può essere scritta come

$$LL = \sum_{h=1}^n \log f(Z_h),$$

ma, in questo caso, la massimizzazione avviene sotto la restrizione di due vincoli: oltre al vincolo sulle probabilità a priori già imposto precedentemente $\sum \pi_x = 1$, si affianca quello riferito alle probabilità condizionate $\sum_{s=1}^{c_i} p_{ix}(s) = 1$ per $i=1, \dots, p$. Questo vincolo non appare nel caso di variabili dicotomiche in quanto una delle due probabilità può essere ottenuta come complemento a 1 dell'altra.

La funzione da massimizzare risulta quindi

$$\phi = LL + \theta \sum_{x=1}^K \pi_x + \sum_{x=1}^K \sum_{i=1}^p \beta_{ix} \sum_{s=1}^{c_i} p_{ix}(s),$$

dove θ e β_{ix} sono moltiplicatori indefiniti. Le derivate parziali rispetto a π_x sono

$$\frac{\partial \phi}{\partial \pi_x} = \sum_{h=1}^n \{g(Z_h | x) / f(Z_h)\} + \theta$$

che conducono, come prima, alla stima

$$\hat{\pi}_x = \frac{1}{n} \sum_{h=1}^n h(x | Z_h).$$

Analogamente,

$$\frac{\partial \phi}{\partial p_{ix}(s)} = \sum_{h=1}^n \pi_x \frac{\partial g(Z_h | x)}{\partial p_{ix}(s)} / f(Z_h)$$

e

$$\frac{\partial g(Z_h | x)}{\partial p_{ix}(s)} = \frac{\partial}{\partial p_{ix}(s)} \exp \sum_{i=1}^p \sum_{s=1}^{c_i} z_{ih}(s) \log p_{ix}(s) = g(Z_h | x) z_{ih}(s) / p_{ix}(s),$$

da cui si ricava

$$\frac{\partial \phi}{\partial p_{ix}(s)} = \pi_x \sum_{h=1}^n \frac{g(Z_h | x)}{f(Z_h)} \frac{z_{ih}(s)}{p_{ix}(s)} + \beta_{ix} = \sum_{h=1}^n \frac{h(x | Z_h) z_{ih}(s)}{p_{ix}(s)} + \beta_{ix}.$$

Ponendo uguale a zero si ottiene

$$\sum_{h=1}^n h(x | Z_h) z_{ih}(s) + p_{ix}(s) \beta_{ix} = 0$$

e, sommando ambo i lati dell'equazione per s , si ricava

$$\sum_{h=1}^n h(x | Z_h) + \beta_{ix} = 0 \quad \text{ovvero} \quad \beta_{ix} = - \sum_{h=1}^n h(x | Z_h)$$

che sostituita nell'equazione precedente conduce alla stima

$$\hat{p}_{ix}(s) = \sum_{h=1}^n h(x | Z_h) z_{ih}(s) / \sum_{h=1}^n h(x | Z_h) = \sum_{h=1}^n h(x | Z_h) z_{ih}(s) / n \hat{\pi}_x$$

per $i = 1, \dots, p$ e $x = 1, \dots, K$.

Come nel caso di variabili dicotomiche, le due stime possono essere ottenute stimando le equazioni tramite l'algoritmo EM. Come primo passo, si scelgono i

valori iniziali di $\hat{h}(x|Z_h)$ ottenendo così una prima approssimazione di $\hat{\pi}_x$ e $\hat{p}_{ix}(s)$. Queste vengono quindi usate per migliorare le stime di $\hat{h}(x|Z_h)$ e così via.

2.2.10 L'analisi a posteriori

Facendo riferimento all'equazione (2.14) si può mostrare che

$$\frac{h(x|Z_h)}{h(x'|Z_h)} = \frac{\pi_x}{\pi_{x'}} \exp \sum_{i=1}^p \sum_{s=1}^{c_i} z_i(s) \log p_{ix}(s) / p_{ix'}(s).$$

L'allocatione degli oggetti (individui) quindi si basa sul confronto delle funzioni lineari

$$B_x = \sum_{i=1}^p \sum_{s=1}^{c_i} z_i(s) \log p_{ix}(s) \text{ per } x = 1, \dots, K.$$

Perciò, la classe x è preferita alla classe x' se $B_x - B_{x'} > \log(\pi_{x'} / \pi_x)$, il che implica che l'oggetto venga classificato nella classe per cui $B_x + \log \pi_x$ risulta più elevato.

2.2.11 Il modello a classi latenti con variabili manifeste politomiche ordinabili

I modelli a classi latenti con variabili manifeste politomiche e ordinabili si basano sulle probabilità cumulate

$$\Pi_{ix}(s) = \sum_{r=s}^{c_i} p_{ix}(r) \text{ con } s = 2, \dots, c_i.$$

Questa è la probabilità di un oggetto (una risposta) nella categoria s o superiore della variabile i per un membro della classe latente x .

2.2.12 L'identificabilità del modello

Come tutti i modelli costituiti da numerosi parametri, c'è la possibilità che il modello a classi latenti possa risultare non identificabile. In altri termini, può esistere più di un punto nello spazio dei parametri in cui la funzione di verosimiglianza assume lo stesso valore. Si pensi ad un insieme di dati raggruppati in una tabella di contingenza di dimensione $c_1 \times c_2 \times \dots \times c_p$. La probabilità della cella designata da Z è data dalla (2.13). Ci sono solo $\prod_{i=1}^p c_i - 1$ probabilità indipendenti a causa della condizione che sommino a uno. Queste probabilità di cella sono funzioni dei parametri del modello che appaiono nel lato destro dell'equazione (2.13). Nel caso in cui vi siano più parametri che probabilità di cella indipendenti, allora, evidentemente, ci sarebbero diversi vettori dei parametri del modello che porterebbero alle stesse funzioni di $f(Z)$ e dunque alla stessa verosimiglianza. Il numero di parametri nel modello è uguale a

$$NPar = K \sum_{i=1}^p (c_i - 1) + (K - 1)$$

dove la prima sommatoria fa riferimento ai parametri p nel modello (2.13), mentre $K - 1$ è il numero dei parametri π .

Il modello risulterà sicuramente non identificato se

$$\prod_{i=1}^p c_i - 1 < K \sum_{i=1}^p (c_i - 1) + K - 1.$$

Tuttavia, questa condizione non è sufficiente in quanto i parametri sono probabilità e perciò sono soggetti a vincoli sui valori che possono assumere.

Non esistono condizioni generali conosciute per ottenere l'identificabilità ma, grazie a Goodman (1974), si è riusciti a delineare il concetto di identificabilità locale. Questo significa che, in un piccolo intorno, il massimo della verosimiglianza è unico. La condizione di Goodman si basa sul fatto che il massimo è unico se la trasformazione dalle probabilità di cella ai parametri è biunivoca in un piccolo intorno. Questo può essere testato verificando la matrice

Jacobiana della trasformazione: il modello è localmente identificabile se il rango della matrice Jacobiana è uguale al numero di colonne $K\left(\sum_{i=1}^p (c_i - 1) + 1\right)$.

La non identificabilità comporta una perdita di precisione nelle stime e impedisce all'algoritmo EM di convergere. Tuttavia, questo problema può essere facilmente ovviato se ci si restringe all'utilizzo di modelli che possono essere stimati con ragionevole precisione. Di solito, questo significa scegliere modelli con non più di tre o quattro classi latenti. Un'altra soluzione consiste nel ricorrere ai modelli a classi latenti fattoriali (si veda paragrafo **3.1**).

2.3 I modelli a classi latenti: altre impostazioni ed estensioni

I modelli a classi latenti presentati nel paragrafo precedente possono essere specificati anche seguendo due impostazioni alternative. La prima è la classica parametrizzazione con probabilità condizionali introdotta da Lazarsfeld (1950a, 1950b), la seconda, proposta principalmente da Haberman (1979), è la parametrizzazione log-lineare ripresa anche da Hagenaars (1990).

Si supponga di considerare un modello a classi latenti non vincolato con una variabile latente X con indice x e quattro variabili manifeste A , B , C e D (equivalenti a Z_1 , Z_2 , Z_3 e Z_4 nell'impostazione proposta nel paragrafo 2.2) con indici a , b , c e d , che indicano le determinazioni delle variabili manifeste, come rappresentato in figura 2.1.

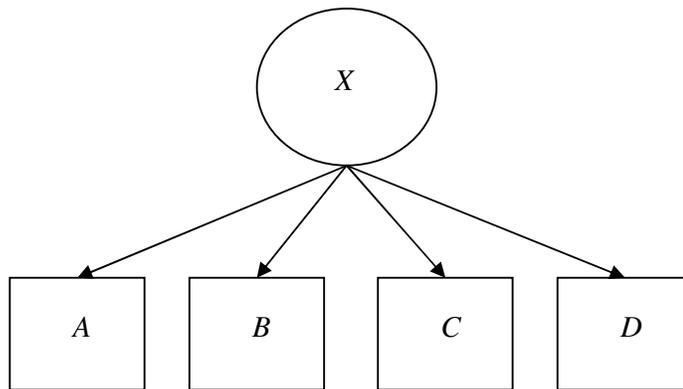


Figura 2.1: il modello a classi latenti

Le equazioni alla base del modello a classi latenti con parametrizzazione a probabilità condizionali sono

$$\pi_{abcd}^{ABCD} = \sum_{x=1}^K \pi_{abcd}^{xABCD} \quad (2.15)$$

in cui K indica il numero di classi latenti e

$$\pi_{abcd}^{XABCD} = \pi_x^X \pi_{abcd|x}^{ABCD|X} = \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X} \pi_{d|x}^{D|X}. \quad (2.16)$$

Il termine π_{abcd}^{XABCD} indica la probabilità della cella (x, a, b, c, d) nella distribuzione congiunta $XABCD$. Come in precedenza, π_x^X denota la probabilità di appartenere alla classe latente x , mentre $\pi_{abcd|x}^{ABCD|X}$ è la probabilità di osservare una particolare sequenza di risposte condizionata a $X = x$. Gli altri parametri π indicano le probabilità condizionali di risposta. Ad esempio, $\pi_{a|x}^{A|X}$ è la probabilità di essere nella categoria a della variabile manifesta A , condizionatamente all'appartenenza alla classe latente x .

Come indicato nell'equazione (2.15), il modello a classi latenti assume che la popolazione sia suddivisa in K classi esaustive e mutualmente esclusive. Quindi, la probabilità congiunta delle variabili osservate può essere ottenuta sommando rispetto alla dimensione latente. Dall'equazione (2.16) si osserva come le quattro variabili manifeste siano postulate essere mutualmente indipendenti dato un particolare valore della variabile latente X : l'assunzione di indipendenza locale.

Haberman dimostra come il modello a classi latenti non vincolato dato in equazione (2.16) sia formalmente identico al modello log-lineare gerarchico $\{XA, XB, XC, XD\}$ scritto come

$$\log m_{abcd} = u + u_x^X + u_a^A + u_b^B + u_c^C + u_d^D + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC} + u_{xd}^{XD} \quad (2.17)$$

nel quale $m_{abcd} = n\pi_{abcd}$. L'equazione (2.17) contiene, oltre alla media generale u e ai termini riferiti ad una variabile (ad esempio, u_a^A che indica il valore del log-*odd* parziale relativo ai casi appartenenti alla categoria a della variabile A ottenuto come scarto dalla categoria di riferimento se ci si riferisce alla codifica *dummy* oppure come scarto dalla media generale per la codifica *effect*), solo i termini di interazione a due variabili tra la variabile latente X e le variabili manifeste che indicano le associazioni parziali (per una trattazione più esaustiva dei modelli log-lineari si veda, ad esempio, Hagenaars, 1993). Siccome non viene inclusa nessuna

interazione tra le variabili manifeste, anche per il modello in equazione (2.17) vale l'assunzione di indipendenza locale per la quale A , B , C , e D vengono assunte essere condizionalmente indipendenti l'una delle altre.

La relazione tra i parametri delle due diverse parametrizzazioni del modello a classi latenti, cioè tra le probabilità condizionali in equazione (2.16) e i parametri log-lineari dell'equazione (2.17), risulta essere

$$\pi_{alx}^{AlX} = \frac{\exp(u_a^A + u_{xa}^{xA})}{\sum_a \exp(u_a^A + u_{xa}^{xA})}. \quad (2.18)$$

La classificazione dei casi nelle K classi latenti stimate dal modello avviene attraverso le probabilità a posteriori usando il teorema di Bayes. In particolare,

$$\hat{\pi}_{x|abcd}^{X|ABCD} = \frac{\hat{\pi}_{abcd}^{XABCD}}{\sum_{x=1}^K \hat{\pi}_{abcd}^{XABCD}}$$

definisce la probabilità a posteriori della sequenza osservata (a, b, c, d) per ogni classe latente $x = 1, \dots, K$. L'oggetto viene assegnato alla classe con la probabilità a posteriori massima (la classe modale).

2.3.1 I modelli a classi latenti vincolati

Se non si ha la necessità di imporre ulteriori restrizioni sui parametri, le due formulazioni del modello a classi latenti sono completamente equivalenti. Tuttavia, se il modello viene vincolato in qualche modo, la scelta della parametrizzazione dipende dal tipo di vincolo che è stato imposto. Alcuni tipi di restrizione possono essere usati in entrambe le specificazioni mentre altri sono specifici di una o dell'altra.

Il modello che combina le due parametrizzazioni descritte sopra è noto come il modello logistico lineare a classi latenti di Formann (1992). L'autore specifica il modello a classi latenti in termini di probabilità latenti e condizionali, dove le probabilità sono parametrizzate come in equazione (2.18). Questo tipo di

formulazione rende possibile combinare vincoli sulle probabilità con restrizioni sui parametri log-lineari.

I vincoli tipicamente imposti nei modelli a classi latenti sono di due tipi: a valore fisso (*fixed-value*) e di uguaglianza sulle probabilità latenti e condizionali (Goodman, 1974). D'altra parte, nel modello a classi latenti log-lineare, si usa imporre vincoli lineari sui parametri u , come, ad esempio, un medesimo effetto della variabile latente sugli indicatori, delle interazioni lineari oppure effetti sulle righe e/o sulle colonne. Oltre a questi, esistono altri tipi di vincoli lineari e non-lineari che possono essere imposti sulle probabilità. Ad esempio, Croon (1990) dimostra come alcuni tipi di vincoli di non uguaglianza sulle probabilità condizionali conducano al modello a classi latenti ordinali.

Di solito, è possibile tradurre i vincoli di uguaglianza sulle probabilità in vincoli di uguaglianza sui parametri log-lineari. Come si può vedere dall'equazione (2.18), per esempio, probabilità condizionali di risposta uguali tra gli indicatori possono anche essere ottenute imponendo l'uguaglianza sia tra i termini delle variabili manifeste che tra i loro termini di interazioni con la variabile latente. Tuttavia, l'assunzione di uguaglianza delle probabilità condizionali risulta essere piuttosto restrittiva. Usando la parametrizzazione log-lineare, è possibile imporre un tipo più debole di vincolo sulle probabilità condizionali, in particolare, imporre la stessa intensità di associazione tra la variabile latente e le varie variabili manifeste. Questa restrizione può essere ottenuta vincolando le interazioni doppie che appaiono in equazione (2.17) ad essere uguali tra loro. Questo tipo di vincolo è simile a quello che Jöreskog (1971) ha definito nel contesto dell'analisi fattoriale: se le probabilità condizionali di risposta sono uguali tra i vari item, questi ultimi possono essere definiti paralleli, mentre vengono chiamati *tau-equivalenti* se sono uguali solo le interazioni doppie con la variabile latente.

Un'altra restrizione piuttosto usata nell'analisi a classi latenti è quella di fissare alcune probabilità condizionali a zero. Questo vincolo è equivalente ad imporre i parametri log-lineari corrispondenti a meno infinito.

I vincoli specifici della parametrizzazione log-lineare sono gli effetti lineari e quelli di riga e/o colonna. Queste restrizioni sono utili nel caso si assuma che la variabile latente, le variabili manifeste od entrambe siano variabili intervallari. Heinen (1996) dimostra che quando una variabile latente viene discretizzata, la maggior parte dei modelli a tratto latente possono essere parametrizzati come modelli a classi latenti con vincoli sui parametri log-lineari.

2.3.2 I modelli con più di una variabile latente

Si può ricorrere a diverse estensioni del modello a classi latenti standard. Alcune di queste sono specifiche della formulazione con probabilità condizionali o della specificazione log-lineare del modello a classi latenti, mentre altre possono essere implementate per entrambe le parametrizzazioni. La prima estensione riguarda l'aggiunta di altre variabili latenti: Goodman (1974) e Haberman (1979) mostrano come specificare dei modelli a classi latenti con più di una variabile latente. Questa specificazione porta ad un modello che è analogo al modello fattoriale analitico con più di un fattore. I modelli a classi latenti con più variabili latenti possono essere ottenuti sia imponendo dei vincoli di uguaglianza sulle probabilità condizionali sia formulando un modello log-lineare.

Si supponga di avere un modello con quattro indicatori e due variabili latenti, nel quale A e B sono le variabili manifeste associate alla variabile latente X e C e D sono indicatori della variabile latente Y . Inoltre, si assume che X ed Y siano associate tra loro. Il modello log-lineare a classi latenti che ne scaturisce è $\{XY, XA, XB, YC, YD\}$ oppure

$$\log m_{xyabcd} = u + u_x^X + u_y^Y + u_a^A + u_b^B + u_c^C + u_d^D + u_{xa}^{XA} + u_{xb}^{XB} + u_{yc}^{YC} + u_{yd}^{YD} .$$

In termini di probabilità condizionali, il modello a classi latenti con due variabili latenti può essere scritto come

$$\pi_{xyabcd}^{XYABCD} = \pi_{xy}^{XY} \pi_{alx}^{AlX} \pi_{blx}^{BlX} \pi_{cly}^{ClY} \pi_{dly}^{DlY} .$$

Il modello appena specificato viene rappresentato in figura 2.2.

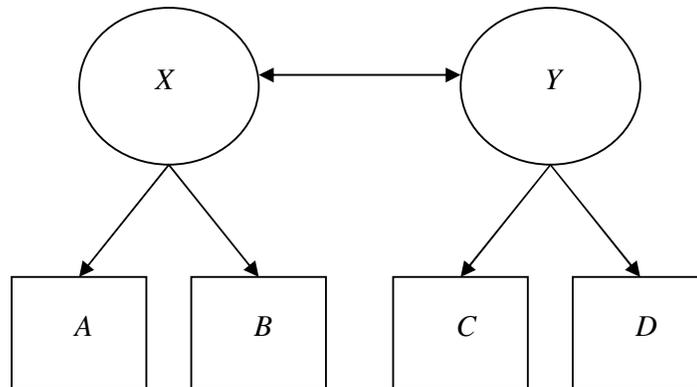


Figura 2.2: il modello a classi latenti con 2 variabili latenti

2.3.3 I modelli con variabili esterne

Un'altra estensione al modello a classi latenti è quella di incorporare delle variabili esterne, dette anche covariate. Siccome le variabili esterne sono variabili manifeste che non sono indicatori, per loro l'assunzione di indipendenza locale non viene rispettata. In altri termini, le variabili latenti spiegano tutte le associazioni esistenti tra gli indicatori ma non le relazioni che intercorrono tra le covariate. Anzi, sono le variabili esterne ad influenzare in maniera causale le variabili latenti. Per questa ragione, queste variabili possono essere definite anche esogene.

Si consideri un modello a classi latenti con due variabili esterne I e J , una variabile latente X e tre indicatori A , B e C . Inoltre, si assuma che le covariate influenzino X e che non vi sia interazione tripla tra I , J e X . Il modello risultante

viene rappresentato in figura 2.3 ed è equivalente al modello a classi latenti log-lineare $\{IJ, IX, JX, XA, XB, XC\}$, ovvero

$$\log m_{xabcij} = u + u_i^I + u_j^J + u_x^X + u_a^A + u_b^B + u_c^C + u_{ij}^{IJ} + u_{ix}^{IX} + u_{jx}^{JX} + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC}.$$

In termini di probabilità condizionali, il modello può essere scritto come

$$\pi_{xabcij}^{XABCIJ} = \pi_{ij}^{IJ} \pi_{x|ij}^{X|IJ} \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}.$$

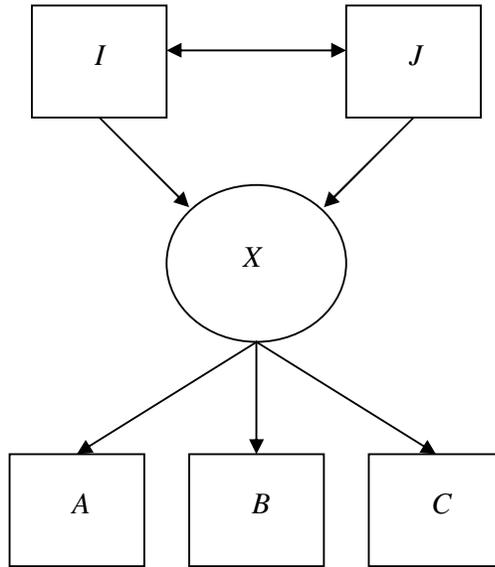


Figura 2.3: il modello a classi latenti con variabili esterne

2.3.4 I modelli per diverse sottopopolazioni

Un'ulteriore estensione riguarda l'introduzione di diverse sottopopolazioni nel modello a classi latenti. Queste possono differire rispetto alla distribuzione latente e alle relazioni tra le variabili latenti e i loro indicatori (Hagenaars, 1990).

Il modello a classi latenti multi-gruppo si ottiene incorporando una variabile di gruppo nel modello standard. La variabile di gruppo può influenzare la distribuzione latente e le probabilità condizionali di risposta. Se si indica con G la

variabile di gruppo (con $g = 1, \dots, Q$) in un modello con una variabile latente X e tre indicatori A, B e C , il modello multi-gruppo non vincolato può essere scritto come

$$\pi_{xabcg}^{XABCG} = \pi_g^G \pi_{x|g}^{X|G} \pi_{a|xg}^{A|XG} \pi_{b|xg}^{B|XG} \pi_{c|xg}^{C|XG},$$

o, in forma log-lineare, con $\{GXA, GXB, GXC\}$. Una specificazione di questo tipo implica che la distribuzione latente, le distribuzioni degli indicatori e le relazioni tra le variabili latenti e gli indicatori siano assunte diverse nelle varie sottopopolazioni.

Tuttavia, spesso si preferisce imporre dei vincoli sui parametri tra i gruppi. Un esempio di modello vincolato è dato dal modello log-lineare $\{GX, XA, XB, XC, GA, GB, GC\}$. In questo caso, si assume che le distribuzioni latenti e manifeste differiscano tra i gruppi ma che le intensità delle relazioni tra la variabile latente e gli indicatori siano le stesse per tutte le sottopopolazioni. Un modello ancora più vincolato è dato da $\{GX, XA, XB, XC\}$. La parte di misurazione di questo modello viene assunta essere uguale per tutti i gruppi. In termini di probabilità condizionali, può essere specificato come

$$\pi_{xabcg}^{XABCG} = \pi_g^G \pi_{x|g}^{X|G} \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}.$$

Questo ultimo modello è equivalente al caso del modello a classi latenti con una variabile esterna (la variabile di gruppo G). L'assunzione di invarianza del costrutto latente che è alla base di questa specificazione può essere utile nel caso si vogliano confrontare le distribuzioni latenti dei diversi gruppi. Queste ultime, infatti, possono essere messe a confronto solo quando la variabile latente assume il medesimo significato per tutte le sottopopolazioni e, perciò, le relazioni tra X e gli indicatori devono essere uguali per ogni gruppo.

2.3.5 I modelli di dipendenza locale

Il modello a classi latenti log-lineare può essere usato anche per specificare dei modelli in cui alcuni indicatori sono collegati tra loro. Hagenaars (1988) dimostra come ottenere questi modelli di dipendenza locale. In figura 2.4 viene mostrato un esempio, in particolare il modello di dipendenza locale $\{XA, XB, XC, XD, CD\}$. In questo modello con una variabile latente e quattro indicatori, esiste una associazione diretta tra le variabili manifeste C e D . In termini di probabilità condizionali, il modello può essere scritto come

$$\pi_{abcd}^{XABCD} = \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{cd|x}^{CD|X}.$$

I modelli di dipendenza locale sono analoghi ai modelli fattoriali con termine di errore correlato (Sörbom, 1975).

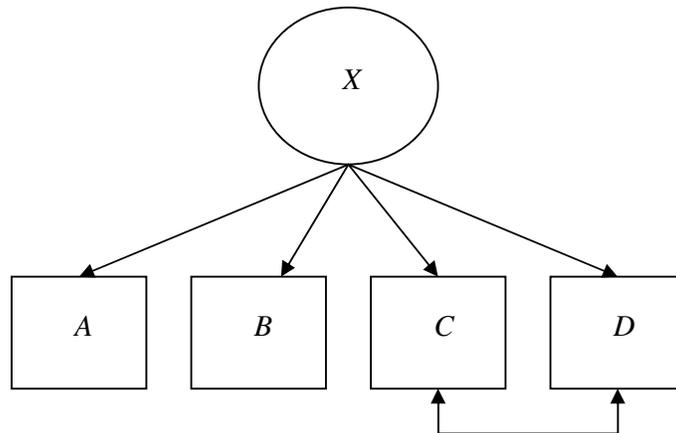


Figura 2.4: il modello di dipendenza locale

2.4 I modelli a profilo latente

I modelli a profilo latente si ottengono nel caso in cui il modello a struttura latente fa riferimento a variabili manifeste continue (si veda tabella 2.1).

L'uso di questi modelli implica la scelta della forma della distribuzione condizionale delle Z_i . Quando si trattano dati dicotomici o politomici, la distribuzione binomiale e quella multinomiale sono una scelta ovvia, ma nel caso di dati continui ci possono essere delle difficoltà nel determinare quale distribuzione assumere. Come si vedrà nel seguito, qualche informazione può essere ricavata dalle distribuzioni marginali, oppure, una volta scelta una distribuzione, l'adattamento del modello risultante e, quindi, il ricorso ad un test sulla bontà, può o meno giustificare la decisione presa a priori.

La distribuzione congiunta delle Z nel modello a profilo latente ha la forma

$$f(Z) = \sum_{x=1}^K \pi_x \prod_{i=1}^p g(z_i | x)$$

per un qualche valore di K , dove $g(z_i | x)$ è la distribuzione condizionale delle Z_i per i membri della classe latente x .

In assenza di informazioni sulla forma di $g(z_i | x)$, qualche consiglio può essere ottenuto dall'ispezione delle distribuzioni marginali date da

$$f(z_i) = \sum_{x=1}^K \pi_x g(z_i | x) \text{ per } i = 1, \dots, p.$$

Una distribuzione bimodale può suggerire una mistura di normali con due classi mentre una distribuzione fortemente asimmetrica potrebbe mettere in guardia contro la scelta di una mistura di gaussiane con qualsiasi numero di componenti. Con questo metodo, si possono ottenere solo informazioni piuttosto grezze ma è comunque possibile evitare assunzioni grossolanamente inappropriate (Bartholomew e Knott, 1999).

2.4.1 La stima di massima verosimiglianza

L'approccio seguito per la stima dei parametri è lo stesso che viene usato per il modello con variabili manifeste categoriche, ma, in questo caso, dipende dalla scelta di $g(z_i | x)$. Le equazioni ottenute ponendo le derivate di LL rispetto a π uguali a zero sono le stesse per qualsiasi distribuzione, comprese quelle usate per dati categorici, e quindi risultano

$$\hat{\pi}_x = \frac{1}{n} \sum_{h=1}^n h(x | Z_h) \text{ per } x = 1, \dots, K. \quad (2.19)$$

Supponendo di considerare la distribuzione condizionale tale che

$$g(z_i | x) = g(z_i | \theta_{ix})$$

allora

$$LL = \sum_{h=1}^n \log \sum_{x=1}^K \pi_x g(z_{ih} | \theta_{ix})$$

e

$$\frac{\partial LL}{\partial \theta_{ix}} = \sum_{h=1}^n \pi_x \frac{\partial g}{\partial \theta_{ix}} / g(z_{ih} | \theta_{ix}).$$

Ponendo queste derivate uguali a zero e risolvendo per θ_{ix} , si hanno equazioni della forma

$$\theta_{ix} = \psi(Z_h, \pi) \text{ con } i = 1, \dots, p \text{ e } x = 1, \dots, K. \quad (2.20)$$

Come prima, si può ricorrere alla procedura EM che alterna le equazioni (2.19) e le (2.20). Inoltre, le derivate seconde possono essere usate per ottenere le stime degli errori standard asintotici.

Le equazioni (2.20) assumono una forma particolare se $g(z_i | \theta_{ix})$ è un membro della famiglia esponenziale e in modo particolare se è normale con media θ_{ix} e varianza unitaria. In questo caso si ottiene

$$\frac{\partial g}{\partial \theta_{ix}} = (z_{ih} - \theta_{ix}) g(z_i | \theta_{ix})$$

e perciò

$$\sum_{h=1}^n h(x | Z_h)(z_{ih} - \theta_{ix}) = 0$$

oppure

$$\hat{\theta}_{ix} = \frac{\sum_{h=1}^n z_{ih} h(x | Z_h)}{\sum_{h=1}^n h(x | Z_h)}. \quad (2.21)$$

Le equazioni (2.21) e (2.19) si prestano senza problemi all'applicazione dell'algoritmo EM.

2.4.2 Ulteriori metodi di stima

La procedura di stima originale proposto per il modello a profilo latente da Lazarsfeld e Henry (1968) consiste nel metodo dei momenti. In questo caso, si considerano sufficienti equazioni di momenti e momenti incrociati in modo da determinare i parametri ignoti. Come metodo di stima, questo approccio è stato surclassato da quello della massima verosimiglianza, ma mostra un importante collegamento con il modello dell'analisi fattoriale (Bartholomew e Knott, 1999): a livello dei momenti secondi, il modello a classi latenti con K classi è indistinguibile dal modello fattoriale con $K-1$ fattori. Questo implica che, una volta stimato un modello fattoriale lineare per una data matrice di covarianze, non c'è garanzia che lo spazio latente sia continuo. Potrebbe essere in egual misura categorico. Inoltre, se una matrice di covarianze è scaturita da un modello fattoriale lineare è sempre possibile trovare un modello a classi latenti con la stessa struttura di covarianza (Molenaar e Von Eye, 1994) e, a causa dell'indeterminatezza associata al modello fattoriale che ammette rotazioni nello spazio fattoriale, questo avviene anche per i modelli a classi latenti. Di conseguenza, l'insieme delle classi identificate dall'analisi può non essere unico.

2.4.3 L'allocatione nelle classi latenti

Se $g(z_i | x)$ è un membro della famiglia esponenziale allora segue dalla teoria generale che la distribuzione a posteriori, e quindi la regola di classificazione, è basata sulla funzione lineare delle Z . In particolare, se z_i data x è normale con media $\mu_i(x)$ e varianza unitaria allora

$$g(Z | x) = (2\pi)^{-\frac{1}{2}p} \exp\left\{-\frac{1}{2}\sum_{i=1}^p (z_i - \mu_i(x))^2\right\}.$$

La classe x è più probabile rispetto alla classe x' se $\pi_x g(Z | x) > \pi_{x'} g(Z | x')$, il che è vero se

$$\sum_{i=1}^p \mu_i(x) z_i - \frac{1}{2} \sum_{i=1}^p \mu_i^2(x) + \log \pi_x > \sum_{i=1}^p \mu_i(x') z_i - \frac{1}{2} \sum_{i=1}^p \mu_i^2(x') + \log \pi_{x'}.$$

2.5 I modelli a tratto latente

I modelli a tratto latente offrono la possibilità di spiegare le associazioni esistenti tra un insieme di variabili manifeste categoriche mediante una variabile latente continua. Oltre alla tipologia di X , un'ulteriore differenza tra il modello a tratto latente e quello a classi latenti risiede nel fatto che nel primo le probabilità condizionali sono collegate ai valori della variabile latente attraverso delle relazioni funzionali esplicite (Masters, 1985). In altre parole, il modello a tratto latente propone una relazione funzionale tra le probabilità $g(Z|x)$ e i valori della variabile X nel continuo. Questo si traduce in una maggiore parsimonia nel numero di parametri rispetto ai modelli a classi latenti. Le forme funzionali più usate in letteratura sono la distribuzione logistica (*logit*) e, qualche volta, la normale ogiva (*probit*). Heinen (1996) mostra come questa differenza tra i due modelli a struttura latente scompaia nel caso si confronti il modello a tratto latente rispetto a modelli a classi latenti con vincoli lineari nella relazione che lega le variabili manifeste con quelle latenti. Inoltre, diversi autori tra cui Bock e Aitkin (1981) hanno dimostrato che l'analisi di uno stesso data set usando modelli a classi latenti e a tratto latente conducano agli stessi risultati.

L'analisi a tratto latente è nota anche come *item response theory* (IRT) ed ha la propria radice nella letteratura psicometrica. La relazione tra il tratto latente e le variabili manifeste viene descritta modellando la probabilità che un individuo con punteggio latente θ_h risponda nella categoria c della variabile (*item*) i . Perciò, in questi modelli, la probabilità condizionata $\pi_c^{i|\theta_h}$ viene espressa usando una certa distribuzione parametrica. Per esempio, quando le variabili manifeste sono dicotomiche e, ad ogni item i , un soggetto può rispondere in maniera corretta ($c=1$) oppure in maniera errata ($c=0$) si è soliti definire un modello per la probabilità che l'individuo, con un certo punteggio latente θ_h , dia la risposta

esatta. Questa probabilità viene espressa con $\pi_1^{i\theta_h}$ e, il suo complemento ad 1 è definito dalla $\pi_0^{i\theta_h}$.

Nel caso più classico della IRT, quando il tratto latente rappresenta l'abilità, la probabilità $\pi_1^{i\theta_h}$ dovrebbe crescere in maniera monotona al crescere dei punteggi latenti. Quando la posizione latente dell'individuo h -esimo è espressa da θ_h mentre la posizione latente dell'item i è indicata da b_i , allora la probabilità $\pi_1^{i\theta_h}$ aumenta in maniera monotona rispetto alla differenza $(\theta_h - b_i)$. Questa relazione può essere ottenuta usando alcune funzioni *link* di distribuzioni cumulate in grado di collegare $\pi_1^{i\theta_h}$ a $(\theta_h - b_i)$: in generale, $\pi_1^{i\theta_h} = \Phi(\theta_h - b_i)$. Le proprietà che la funzione Φ dovrebbe avere sono: $0 \leq \Phi(u) \leq 1$ per $-\infty \leq u \leq +\infty$, se $u_1 \leq u_2$ allora $\Phi(u_1) \leq \Phi(u_2)$ e $\Phi(u)$ deve essere continua in u . È quindi possibile usare la distribuzione logistica ottenendo

$$\pi_1^{i\theta_h} = \frac{1}{1 + \exp[-(\theta_h - b_i)]} = \frac{\exp(\theta_h - b_i)}{1 + \exp(\theta_h - b_i)}. \quad (2.22)$$

L'espressione (2.22) risulta molto simile alla formula specificata per i modelli a classi latenti con la parametrizzazione log-lineare (2.18). L'unica differenza risiede nel fatto che nella (2.22) le quantità θ_h e b_i sono continue mentre nella (2.18) le variabili sono categoriche e, pertanto, non seguono una distribuzione logistica cumulata. In altre parole, si dice che la relazione tra le probabilità di risposta e la variabile latente sia *parametrizzata* nel caso dei modelli a tratto latente e *non-parametrizzata* nel caso dei modelli a classi latenti. Come affermato in precedenza, la distinzione tra i due tipi di modelli dipende dalla differente natura delle variabili latenti.

2.5.1 I modelli a tratto latente per variabili dicotomiche

In letteratura, i modelli a tratto latente per variabili dicotomiche vengono definiti in base al numero dei parametri riferiti agli item che sono utilizzati. Il modello espresso dall'equazione (2.22) denota il modello a tratto latente con un parametro: b_i che esprime la difficoltà dell'item. In figura 2.5 viene rappresentata la funzione di risposta all'item per un modello a tratto latente con quattro parametri. Si nota come il parametro b_i (*Item Calibration* in figura 2.5) marchi un punto nel continuo latente per il quale la probabilità di dare una risposta corretta è esattamente 0,5.

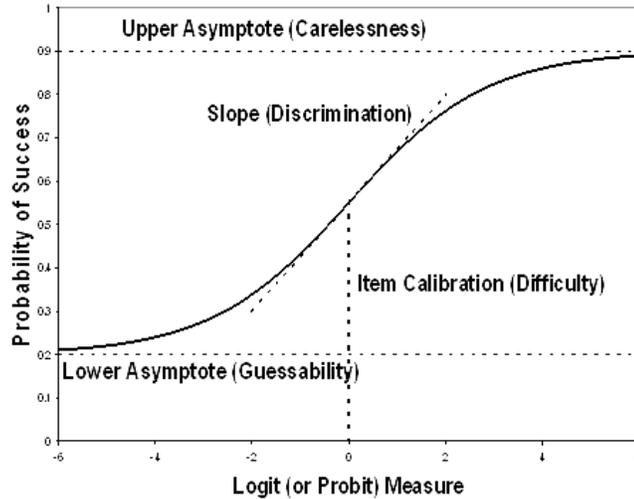


Figura 2.5: la funzione di risposta all'item per un modello a tratto latente logistico con quattro parametri

La probabilità di dare la risposta corretta all'item dicotomico i è uguale a

$$\log \frac{\pi_1^{i\theta_h}}{\pi_0^{i\theta_h}} = \theta_h - b_i.$$

Questa probabilità è una semplice funzione lineare dell'abilità individuale e della difficoltà dell'item.

Il modello logistico con un parametro è identico al modello sviluppato da Rasch (1960) e richiede l'esistenza di statistiche sufficienti sia per il parametro individuale sia per quello relativo alla difficoltà dell'item.

La limitatezza del modello a tratto latente appena mostrato è quella di permettere al solo parametro di difficoltà di variare. Pertanto, questo modello assume che tutti gli item discriminano nello stesso modo individui con abilità diverse. Può essere opportuno, quindi, introdurre un ulteriore parametro in grado di spiegare la variabilità connessa al modo in cui item diversi possono discriminare le abilità degli individui ottenendo

$$\pi_1^{i|\theta_h} = \frac{\exp[a_i(\theta_h - b_i)]}{1 + \exp[a_i(\theta_h - b_i)]}$$

Il parametro di discriminazione a_i rappresenta la pendenza della curva della funzione di risposta (indicata con *Slope* in figura 2.5). Nel modello con un parametro, le curve delle funzioni di risposta per i vari item sono parallele in quanto si ipotizza che il parametro di discriminazione sia lo stesso per ogni item. Dalla figura 2.6 si evince come un valore differente di a_i comporti pendenze differenti delle curve ed, in particolare, come aumentando il valore del parametro aumenti la forza discriminante dell'item.

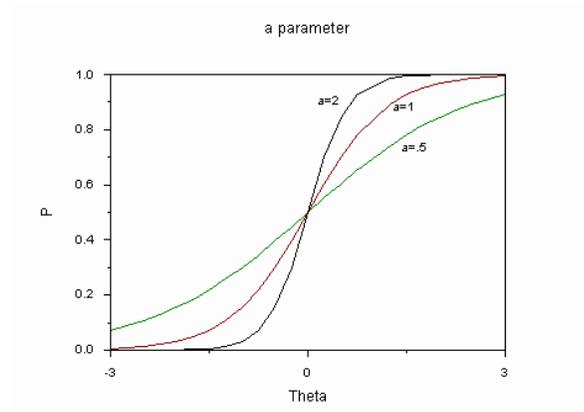


Figura 2.6: la funzione di risposta all'item per un modello a tratto latente logistico a due parametri con diversi valori del parametro di discriminazione

I modelli con tre e quattro parametri permettono di inserire nel modello a tratto latente altre due aspetti: il fatto che un individuo possa rispondere correttamente ad un item tirando ad indovinare e quello che un individuo molto abile possa sbagliare la risposta ad un item per noncuranza. Il modello a tre parametri ammette che l'asintoto di sinistra sia maggiore di zero come mostrato in figura 2.5 e, quindi, rende possibile rilevare l'eventualità che individui con una bassa abilità possano rispondere correttamente indovinando. Il modello a quattro parametri proposto da McDonald (1967) assume che l'asintoto destro sia inferiore ad uno (figura 2.5). L'interpretazione del quarto parametro è che individui con un'abilità alta possano sbagliare la risposta per una qualche ragione (ad esempio, per noncuranza o distrazione).

2.5.2 I modelli a tratto latente per variabili politomiche

Lo scopo di questo paragrafo è di fare un elenco delle principali tipologie di modelli a tratto latente per variabili politomiche, senza scendere nel dettaglio. Per una trattazione più accurata dei vari tipi di modelli a tratto latente per dati politomici si rimanda a Heinen (1996).

Thissen e Steinberg (1986) propongono una prima distinzione tra i modelli a differenze (*difference models*) e i quelli "divisi per totale" (*divide-by-total models*). I modelli appartenenti alla prima categoria non si riferiscono alle probabilità che un individuo h risponda nella categoria c per l'item i , ma si concentrano sulla probabilità di rispondere nella categoria c o superiore. Si nota quindi come le probabilità cumulate (o i complementi delle probabilità cumulate) assumano una posizione centrale nei modelli a differenze. Nel contesto dei modelli a tratto latente, il più comune modello a differenze è il *graded response model* proposto da Samejima (1969). Il termine modelli a differenze si riferisce al fatto che le probabilità di risposta possono essere derivate calcolando il complemento di due probabilità cumulate.

I modelli divide-by-total si basano sulle probabilità condizionate che un individuo risponda nella categoria c , dato il fatto che la risposta sia nella categoria c o $c - 1$, piuttosto che sulle probabilità cumulate e i loro complementi. L'idea di usare le probabilità condizionate come struttura portante dei modelli a tratto latente per variabili politomiche è stata proposta da Masters (1982) e sviluppata nel modello *partial credit*. Un modello più generale di quello definito da Masters viene proposto da Bock (1972) ed è noto nella letteratura psicometrica come il *nominal response model*. Questo modello è strettamente collegato ad un modello a classi latenti con vincoli lineari.

2.5.3 Le principali procedure di stima

Dai paragrafi precedenti è chiaro che vi sono due tipi di parametri nella specificazione di ogni modello a tratto latente: i parametri individuali θ_h e i parametri degli item (la difficoltà, la discriminazione e i due parametri soglia legati agli asintoti della funzione di risposta). La stima dei parametri usati in uno specifico modello a tratto latente può essere effettuata in vari modi.

Un primo approccio consiste nella stima simultanea sia dei parametri individuali sia di quelli degli item. Il metodo è noto come stima di massima verosimiglianza congiunta (*JML*) ma presenta alcuni difetti e, per questo, non è molto usata.

Un'altra possibilità è quella di stimare i parametri degli item, il che richiede l'eliminazione dei parametri individuali dalla funzione di verosimiglianza. Si può ricorrere a due metodi. Il primo si basa sulla conoscenza delle statistiche sufficienti per i parametri individuali. Condizionando rispetto a queste statistiche, si ottiene una funzione di verosimiglianza nella quale compaiono solo i parametri degli item. Questa procedura viene chiamata massima verosimiglianza condizionata (*CML*). Un secondo metodo per sbarazzarsi dei parametri individuali è quello di imporre delle assunzioni circa la distribuzione della variabile latente.

Se viene scelta l'assunzione corretta, i parametri individuali possono essere integrati fuori dalla funzione di verosimiglianza e il metodo che si segue è noto come massima verosimiglianza marginale (*MML*).

La massima verosimiglianza condizionata può essere usata solo se le statistiche sufficienti per i parametri individuali sono note, il che risulta vero per modelli a tratto latente che non usano parametri di discriminazione. Quando gli unici parametri degli item che vengono usati sono parametri soglia (o loro funzioni), le statistiche sufficienti per i parametri individuali sono note. In maniera analoga, la massima verosimiglianza condizionata può essere usata per stimare il modello partial credit e tutti i modelli che derivano da quel modello imponendo vincoli aggiuntivi sui parametri soglia.

Quando sia i parametri soglia che i parametri di discriminazione sono inseriti nel modello, possono essere stimati tramite la *MML*. Per applicare questo metodo, alcune restrizioni devono essere imposte sulla distribuzione della variabili latente. L'assunzione più generale dichiara che la distribuzione di θ è completamente nota. In questo caso si parla di *MML* parametrica. Di solito, vengono considerate assunzioni meno rigorose: la distribuzione di θ può essere approssimata da una distribuzione discreta. In molti casi, inoltre, si assume che i valori della variabile latente discretizzata (cioè, i nodi latenti) siano noti. Questa procedura di stima, piuttosto usata nella pratica, è conosciuta come *MML* semiparametrica e permette di stimare le probabilità latenti insieme ai parametri degli item.

2.6 Il Latent Markov Model

Un modello a classi latenti che sta avendo un grande successo nell'analisi delle serie storiche e dei dati longitudinali è il latent Markov model (LMM). Conosciuto anche come hidden Markov model (Baum et al., 1970), modello a transizione latente o modello a cambiamento di regime (Hamilton, 1989), il LMM descrive la relazione tra due processi stocastici: uno osservabile ed uno sottostante e latente. Rispetto ai modelli a variabili latenti trattati finora, il LMM si basa su un insieme di variabili latenti Y_t che, in funzione del tempo $t = 1, \dots, T$, analizza la dinamica di una serie storica osservabile $\{Z_t\}_{t=1}^T$ ed il passaggio delle osservazioni tra S stati latenti discreti.

A differenza dei modelli di mistura nei quali, dato lo stato latente al tempo t , la distribuzione delle osservazioni in quell'istante temporale risulta completamente specificata, l'impostazione del LMM assume che gli stati latenti non siano tra loro indipendenti ma seguano una struttura di tipo markoviano. In tal modo, il modello risulta più generale e flessibile. Una conseguenza di questa assunzione è data dal fatto che anche i dati osservati Z_t risultano (auto)correlati con una dipendenza che decresce a zero al crescere della loro distanza nel tempo.

Come evidenziato da MacKay (2003), il ricorso a questi modelli permette il raggiungimento di due obiettivi principali: il primo consiste nel fare inferenza su di un processo stocastico non osservabile sulla base di uno che è possibile osservare direttamente; il secondo è quello di spiegare la variazione del processo osservato sulla base della variazione del processo latente postulato. In particolare, la variabilità dei dati osservati può essere analizzata attraverso un LMM assumendo che le osservazioni appartengano ad una di tante distribuzioni condizionali differenti, ognuna associata ad un diverso stato latente.

Il LMM consiste di due variabili:

1. una serie storica osservata z_{ht} , realizzazione di un processo stocastico $\{Z_t\}_{t \geq 0}$, che esprime l'osservazione h -esima al tempo t , con $h = 1, \dots, n$ e $t = 1, \dots, T$. La variabile z_{ht} può assumere qualsiasi tipologia: continua o categorica;
2. una variabile latente categorica Y_t , che varia in funzione del tempo, definita da un insieme di S stati (latenti) ed una matrice di transizione P di dimensione $S \times S$ contenente le probabilità $p_{jk} = \Pr(Y_t = k | Y_{t-1} = j)$ di transitare dallo stato j al tempo $t - 1$ allo stato k al tempo t (per $j, k = 1, \dots, S$).

La variabile latente Y_t consiste, perciò, in una catena di Markov del primo ordine che viene assunta come discreta, omogenea, aperiodica, irriducibile e definita su uno spazio di stato finito³.

In altre parole, viene introdotto il modello (condizionale) per il processo osservato $\{Z_t\}_{t=1}^T$, $f(Z_t | \theta) = \Pr(Z_t | Y_t = j, \theta)$, dove θ indica il corrispondente insieme di parametri.

Il processo osservato deve soddisfare due condizioni:

1. la condizione di indipendenza condizionale (locale): le variabili Z_t per $t = 1, \dots, T$ sono condizionalmente indipendenti date le variabili latenti Y_t ;
2. la condizione di dipendenza contemporanea: la distribuzione di ogni Z_t , date le variabili Y_t , dipende soltanto dallo stato latente occupato al tempo t .

Tenendo in considerazione queste due condizioni, il modello LMM riferito alla serie storica dell'osservazione h -esima è definito come

$$f(Z_h) = \sum_{y_1=1}^S \sum_{y_2=1}^S \dots \sum_{y_T=1}^S f(y_1, \dots, y_T) f(Z_h; y_1, \dots, y_T) \quad (2.23)$$

dove

³ Per un maggiore approfondimento sui dettagli del processo markoviano alla base della variabile latente Y_t si rimanda a Cappé et al. (2005).

$$f(y_1, \dots, y_T) = f(y_1) \prod_{t=2}^T f(y_t | y_{t-1}) \quad (2.24)$$

e

$$f(Z_h; y_1, \dots, y_T) = \prod_{t=1}^T f(z_{ht} | y_t). \quad (2.25)$$

L'equazione (2.23) descrive la funzione di densità di probabilità associata alla serie storica dell'osservazione h . La parte a destra dell'equazione mostra che il modello è una mistura contenente T variabili latenti. Il numero totale di classi latenti (o componenti di mistura) è uguale a S^T . Come per ogni modello di mistura, $f(Z_h)$ si ottiene marginalizzando rispetto alle variabili latenti categoriche, ovvero attraverso la somma di una media pesata di densità di probabilità $f(Z_h; y_1, \dots, y_T)$, dove le probabilità di appartenenza alle classe latenti (le probabilità a priori) $f(y_1, \dots, y_T)$ sono usate come pesi (McLachlan e Peel, 2000). Nelle scienze sociali, questo tipo di modello viene usato per l'analisi di dati longitudinali di natura categorica (Van de Pol e Langeheine, 1990; Vermunt et al., 1999; Bartolucci et al., 2007). Recentemente, altri autori (Rydén et al., 1998; Castellano e Scaccia, 2007; Dias et al., 2008; Frühwirth-Schnatter e Kaufmann, 2008) hanno invece applicato questo modello a serie storiche finanziarie ed esteso l'analisi allo studio di variabili continue.

Dalle equazioni (2.24) e (2.25) si deducono le due assunzioni imposte dal modello LMM che permettono la semplificazione della forma delle densità $f(y_1, \dots, y_T)$ e $f(Z_h; y_1, \dots, y_T)$. In particolare, l'equazione (2.24) mostra che y_t è associata solo a y_{t-1} e y_{t+1} ed è, quindi, indipendente dagli stati latenti occupati in altri istanti temporali. Questa ulteriore imposizione del modello LMM rispetto ai modelli di mistura è la nota assunzione del primo ordine delle catene di Markov. L'equazione (2.25) mostra che, condizionatamente a y_t , l'osservazione al tempo t (z_{ht}) è indipendente dalle osservazioni e dagli stati latenti di altri istanti temporali.

Dalle equazioni (2.24) e (2.25) si evince che il modello di interesse è caratterizzato da tre funzioni di probabilità:

1. $f(y_1)$ è la probabilità dello stato iniziale, ovvero la probabilità di essere in un particolare stato latente al tempo 1;
2. $f(y_t | y_{t-1})$ è la probabilità di transizione latente, ovvero la probabilità di essere in un particolare stato latente al tempo t condizionatamente allo stato latente occupato al tempo $t - 1$: assumendo un processo di transizione omogeneo rispetto al tempo, si ottiene la matrice di transizione P ;
3. $f(z_{ht} | y_t)$ è la funzione di densità delle osservazioni, che esprime la densità di probabilità di avere una determinata osservazione h al tempo t condizionatamente allo stato latente occupato al tempo t . Il vettore dei parametri definito con θ dipende dalla natura della variabile osservata. Ad esempio, se la variabile Z è continua allora la distribuzione di riferimento è la gaussiana caratterizzata da un vettore di parametri $\theta_j = (\mu_j, \sigma_j^2)$ contenente le medie e le varianze del generico stato latente j .

2.6.1 La stima dei parametri

La stima dei parametri del modello LMM avviene tramite la massima verosimiglianza e implica la massimizzazione della funzione di log-verosimiglianza $LL = \sum_{h=1}^n \log f(Z_h)$. Questo problema può essere risolto ricorrendo all'algoritmo EM dove, al passo E, si calcola $f(y_1, \dots, y_T | Z_h) = f(y_1, \dots, y_T, Z_h) / f(Z_h)$, ovvero la distribuzione congiunta delle T variabili latenti condizionata ai dati (probabilità a posteriori) e le stime provvisorie dei parametri del modello. Al passo M, si ricorre ai metodi di massima verosimiglianza per dati completi per aggiornare i parametri incogniti del modello prendendo in considerazione una matrice estesa dei dati con

$f(y_1, \dots, y_T | Z_h)$ come pesi. Si nota come l'algoritmo EM standard, ad ogni iterazione, necessita il calcolo di S^T valori di $f(y_1, \dots, y_T | Z_h)$ per ogni oggetto osservato o per ogni unica sequenza di dati. Questo implica che il tempo computazionale e la memoria necessaria al computer aumentano esponenzialmente in ragione del numero degli istanti temporali, rendendo questo algoritmo impraticabile o addirittura impossibile da applicare anche per serie storiche piuttosto ridotte.

L'assunzione di indipendenza condizionale imposta dal modello LMM viene sfruttata per definire una speciale variante dell'algoritmo EM che, usualmente, viene chiamata algoritmo *forward-backward* o di Baum-Welch (Baum et al., 1970; Hamilton, 1989). Questo algoritmo circuisce il calcolo della distribuzione a posteriori congiunta sfruttando l'indipendenza condizionale delle Z_t date le Y_t .

Il passo M dell'algoritmo EM necessita soltanto delle distribuzioni marginali $f(y_t | Z_h)$ e $f(y_{t-1}, y_t | Z_h)$. L'algoritmo Baum-Welch ottiene queste quantità direttamente invece di calcolare prima $f(y_1, \dots, y_T | Z_h)$ e quindi collassare sulle rimanenti dimensioni come avviene nell'algoritmo EM.

I due componenti chiave dell'algoritmo Baum-Welch sono le probabilità *forward* $\alpha_{ht}(j)$ e le probabilità *backward* $\beta_{ht}(j)$. Queste quantità sono definite come

$$\begin{aligned}\alpha_{ht}(j) &= f(y_t = j, z_{h1}, \dots, z_{ht}) \\ \beta_{ht}(j) &= f(z_{ht+1}, \dots, z_{hT} | y_t = j)\end{aligned}$$

quindi, $\alpha_{ht}(j)$ si riferisce all'informazione dell'unità h -esima fino al tempo t mentre $\beta_{ht}(j)$ considera le osservazioni di h dopo l'istante temporale t . In particolare, la probabilità forward $\alpha_{ht}(j)$ denota la probabilità congiunta di osservare un determinato insieme di valori dell'unità h -esima fino al tempo t e di appartenere allo stato latente j ; mentre $\beta_{ht}(j)$ è la probabilità di avere un

determinato insieme di osservazioni dal tempo $t + 1$ al tempo T data l'appartenenza allo stato j .

Usando $\alpha_{ht}(j)$ e $\beta_{ht}(j)$ si ottengono le distribuzioni posteriori marginali

$$f(y_t = j | Z_h) = \frac{\alpha_{ht}(j)\beta_{ht}(j)}{f(Z_h)}$$

$$f(y_{t-1} = k, y_t = j | Z_h) = \frac{\alpha_{ht-1}(k)f(y_t = j | y_{t-1} = k)f(z_{ht} | y_t = j)\beta_{ht}(j)}{f(Z_h)}$$

L'elemento chiave dell'algorithm forward-backward consiste nel fatto che i termini $\alpha_{ht}(j)$ e $\beta_{ht}(j)$ vengono calcolati usando schemi ricorsivi. Lo schema ricorsivo forward per $\alpha_{ht}(j)$ comincia dal primo istante temporale e continua fino al raggiungimento di T :

$$\alpha_{h1}(j) = f(y_1 = j)f(z_{h1} | y_1 = j)$$

$$\alpha_{ht}(j) = \left[\sum_{k=1}^S \alpha_{ht-1}(k)f(y_t = j | y_{t-1} = k) \right] f(z_{ht} | y_t = j)$$

per $t = 2, \dots, T$. Lo schema ricorsivo backward per $\beta_{ht}(j)$ comincia con l'ultimo istante temporale ($t = T$) e procede a ritroso fino al raggiungimento di $t = 1$:

$$\beta_{hT}(j) = 1$$

$$\beta_{ht}(j) = \sum_{k=1}^S \beta_{ht+1}(k)f(y_{t+1} = k | y_t = j)f(z_{ht+1} | y_{t+1} = k)$$

per $t = T - 1, \dots, 1$.

Successivamente, si calcolano $f(y_t = j | Z_h)$ e $f(y_{t-1} = k, y_t = j | Z_h)$ usando le precedenti equazioni. Al passo M , queste quantità sono necessarie per ottenere le nuove stime dei parametri del modello LMM che compongono le equazioni (2.24) e (2.25), attraverso il ricorso a metodi standard quali, ad esempio, il Newton-Raphson. La procedura continua fino al raggiungimento della convergenza.

2.6.2 Le estensioni del LMM

Esistono due ulteriori estensioni fondamentali ai modelli latent Markov ottenibili apportando delle modifiche all'equazione riferita alla catena markoviana (2.24) o, in maniera analoga, alla (3.4) nel paragrafo 3.2.

La prima consiste nel rendere il modello eterogeneo nel tempo, in altre parole, permettere alla matrice di transizione P di variare rispetto a t . In questo contesto, si elimina l'assunzione di invarianza nel tempo delle probabilità di transizione p_{jk} ammettendo che queste ultime possano cambiare anche ad ogni istante temporale. Il modello LMM eterogeneo rispetto al tempo risulta più appropriato nei casi in cui esistano fattori endogeni che influiscono in maniera significativa su determinati istanti temporali e non su altri, rendendo perciò le probabilità di transizione dallo stato k allo stato j diverse nel tempo. La differenza nel numero dei parametri da stimare tra il modello eterogeneo e quello omogeneo rispetto al tempo è pari a $S(S-1)(T-2)$. Perciò, quando il numero di osservazioni temporali T è elevato il LMM eterogeneo nel tempo risulta spesso avere un adattamento ai dati peggiore rispetto al LMM omogeneo a causa dell'alto numero di parametri da stimare. Inoltre, è possibile modellare le probabilità di transizione attraverso regressioni logistiche per mostrare come alcune covariate influenzino i cambiamenti tra gli stati latenti, ovvero come tali variabili esterne determinino l'eterogeneità temporale delle probabilità p_{jk} (Muenz e Rubinstein, 1985).

La seconda estensione al modello LMM consiste nell'adottare una catena di Markov del secondo ordine invece che del primo. In questo modo, la matrice di transizione $P = \{p_{jkw}\}$ diventa da bi- a tri-dimensionale e la probabilità di transitare nello stato j al tempo t , $p_{jkw} = \Pr(y_t = j \mid y_{t-1} = k, y_{t-2} = w)$ con $j, k, w = 1, \dots, S$, dipende non solo dallo stato occupato al tempo precedente ma anche dallo stato di due osservazioni fa. La struttura markoviana con approssimazione al secondo ordine può essere in taluni casi più appropriata, ad

esempio, Thede e Harper (1999) mostrano come tale modello conduca a risultati migliori se applicato a dati riferiti all'analisi grammaticale.

3. Metodi statistici per le variabili latenti: gli ultimi sviluppi

In questo paragrafo si mostrano alcuni sviluppi recenti nel campo dei metodi a struttura latente. Nel paragrafo **3.1** si descrive un'estensione al modello a classi latenti particolarmente indicato per la riduzione del numero di classi: il *modello a classi latenti fattoriali*. Nei paragrafi **3.2** e **3.3**, si valutano due estensioni del LMM: il *latent Markov di mistura* e il *latent Markov misto*.

3.1 Il modello a classi latenti fattoriali

Nel paragrafo **2.3.2** si è mostrato come specificare un modello a classi latenti con più di una variabile latente. Un modo alternativo per ottenere un modello di questo tipo consiste nel ricorrere al modello a classi latenti fattoriali (*LC factor model*). I modelli a classi latenti fattoriali sono stati proposti come un'alternativa generale all'analisi a classi latenti di tipo esplorativa tradizionale da Madigson e Vermunt (2001). Il punto di forza di questa tipologia di modelli consiste nell'ottenere una parametrizzazione più parsimoniosa e quindi, di solito, risultati più facilmente interpretabili a fronte di una minima perdita di adattamento.

I modelli a classi latenti fattoriali sono stati originariamente introdotti da Goodman (1974) nel contesto dell'analisi a classi latenti confermativa. Certi modelli a classi latenti tradizionali contenenti quattro o più classi possono essere interpretati in termini di due o più componenti di variabili latenti, dove queste componenti vengono trattate come una variabile congiunta (si veda, ad esempio, Hagenaars, 1990). Per esempio, una variabile latente X con $K = 4$ classi può essere espressa come due variabili latenti dicotomiche $V = \{1, 2\}$ e $W = \{1, 2\}$ usando la seguente corrispondenza:

	$W = 1$	$W = 2$
$V = 1$	$X = 1$	$X = 2$
$V = 2$	$X = 3$	$X = 4$

Quindi, $X = 1$ corrisponde a $V = 1$ e $W = 1$, $X = 2$ a $V = 1$ e $W = 2$, e così via.

Formalmente, per quattro variabili manifeste, il modello con quattro classi latenti può essere ri-parametrizzato come un modello fattoriale con due variabili latenti dicotomiche:

$$\pi_{vwabcd}^{VWABCD} = \pi_{vw}^{VW} \pi_{abcd|vw}^{ABCD|vw} = \pi_{vw}^{VW} \pi_{avw}^{AVW} \pi_{blvw}^{B|vw} \pi_{clvw}^{C|vw} \pi_{dlvw}^{D|vw} . \quad (3.1)$$

Il modello a classi latenti fattoriali non vincolato può essere scritto anche in forma log-lineare:

$$\begin{aligned} \log m_{vwabcd} = & u + u_v^V + u_w^W + u_{vw}^{VW} + u_a^A + u_b^B + u_c^C + u_d^D + u_{av}^{AV} + u_{bv}^{BV} + u_{cv}^{CV} + u_{dv}^{DV} + \\ & + u_{aw}^{AW} + u_{bw}^{BW} + u_{cw}^{CW} + u_{dw}^{DW} + u_{avw}^{AVW} + u_{bvw}^{BVW} + u_{cvw}^{CVW} + u_{dvw}^{D|vw} \end{aligned} \quad (3.2)$$

dove

$$\pi_{vw}^{VW} = \frac{\exp(u_v^V + u_w^W + u_{vw}^{VW})}{\sum_v \sum_w \exp(u_v^V + u_w^W + u_{vw}^{VW})}, \quad \pi_{avw}^{AVW} = \frac{\exp(u_a^A + u_{av}^{AV} + u_{aw}^{AW} + u_{avw}^{AVW})}{\sum_a \exp(u_a^A + u_{av}^{AV} + u_{aw}^{AW} + u_{avw}^{AVW})},$$

così via.

La corrispondenza tra questa rappresentazione e il modello a classi latenti in equazione (2.17) si ha scrivendo il termine log-lineare riguardante la variabile X come $u_{2(v-1)+w}^X = u_v^V + u_w^W + u_{vw}^{VW}$ ed i termini doppi riferiti alla X come $u_{a,2(v-1)+w}^{AV} = u_{av}^{AV} + u_{aw}^{AW} + u_{avw}^{AVW}$, $u_{b,2(v-1)+w}^{BV} = u_{bv}^{BV} + u_{bw}^{BW} + u_{bvw}^{BVW}$, eccetera. È facile da verificare come questa ri-parametrizzazione non modifica il numero dei parametri nel modello.

Madigson e Vermunt (2001) definiscono il modello a classi latenti con R fattori come un modello fattoriale vincolato contenente R variabili latenti dicotomiche mutualmente indipendenti con parametri (intesi come i pesi fattoriali)

che misurano l'associazione di ogni variabile latente su ogni indicatore. Nello specifico, il modello *base R*-fattoriale è definito imponendo due insiemi di restrizioni sul modello a classi latenti fattoriale non vincolato. Il modello a due fattori risultante è una forma vincolata del modello a classi latenti con quattro classi. Senza alcuna restrizione, il modello con due fattori sarebbe equivalente al modello con quattro classi latenti.

Il primo insieme di vincoli pone uguale a zero ogni termine di interazione di ordine superiore a due. Per il modello base con due fattori, si ha $u_{av}^{AVW} = u_{bv}^{BVW} = u_{cv}^{CVW} = u_{dv}^{DVW} = 0$. Dopo l'imposizione di queste restrizioni, i termini log-lineari del secondo ordine nel modello base con due fattori diventano $u_{a,2(v-1)+w}^{AV} = u_{av}^{AV} + u_{aw}^{AW}$, $u_{b,2(v-1)+w}^{BV} = u_{bv}^{BV} + u_{bw}^{BW}$, eccetera. Per la variabile manifesta A , u_{av}^{AV} rappresenta il peso di A sul fattore V , mentre u_{aw}^{AW} denota il peso di A sul fattore W , e così via. Fissando il termine u di interazione tra le tre variabili uguale a zero, si ottiene un modello che è concettualmente simile all'analisi fattoriale esplorativa classica: ognuno dei fattori può avere un effetto su ogni indicatore ma non vi sono termini di interazione di ordine superiore. I vincoli finora specificati sono necessari per permettere che le quattro classi latenti siano espresse come tabella incrociata delle due variabili latenti e quindi sono essenziali per distinguere il modello a classi latenti fattoriale da quello a classi latenti.

Il secondo insieme di restrizioni pone le variabili latenti come indipendenti tra loro. Per il modello a due fattori, questo vincolo impone l'indipendenza nella tavola a doppia entrata $[VW]$. Questa restrizione rende il modello simile all'analisi fattoriale esplorativa. Come viene mostrato più avanti nel paragrafo, si può anche evitare di imporre questa assunzione per ottenere i modelli a classi latenti fattoriali per analisi confermativa.

Nel modello base a due fattori, le probabilità condizionali in equazione (3.1) sono vincolate dai modelli logit

$$\pi_{vw}^{vW} = \frac{\exp(u_v^V + u_w^W)}{\sum_v \sum_w \exp(u_v^V + u_w^W)}, \quad \pi_{avw}^{AVW} = \frac{\exp(u_a^A + u_{av}^{AV} + u_{aw}^{AW})}{\sum_a \exp(u_a^A + u_{av}^{AV} + u_{aw}^{AW})}, \text{ ecc.},$$

i quali escludono l'associazione marginale tra le variabili latenti, oltre ai termini di interazione di ordine superiore.

Sebbene il modello base R -fattoriale sia un caso speciale di un modello a classi latenti con 2^R classi, in realtà, a causa dei vincoli imposti, esso è comparabile ad modello con solo $K = R + 1$ classi latenti in termini di parsimonia.

Il numero di parametri nel modello base R -fattoriale è $(1 + R) \sum_{i=1}^p (c_i - 1) + R$ dove

c_i sono le categorie dell' i -esima variabile manifesta per $i = 1, \dots, p$, mentre il

numero dei parametri del modello a classi latenti⁴ è $K \sum_{i=1}^p (c_i - 1) + (K - 1)$.

Quindi, si nota come il grado di parsimonia del modello base R -fattoriale sia lo stesso del modello a classi latenti con $K = R + 1$ classi.

Siccome il modello a classi latenti R -fattoriale non vincolato è equivalente ad un modello con $K = 2^R$ classi latenti, il numero dei vincoli imposti in un modello fattoriale base può essere calcolato come differenza tra il numero dei parametri nei due modelli:

$$\left[2^R \sum_{i=1}^p (c_i - 1) + (2^R - 1) \right] - \left[(1 + R) \sum_{i=1}^p (c_i - 1) + R \right] = (2^R - R - 1) \left(\sum_{i=1}^p (c_i - 1) + 1 \right).$$

Questa notevole riduzione nel numero dei parametri è sufficiente a rendere il modello identificabile in molte situazioni. In altre termini, frequentemente si è nel caso in cui il modello base R -fattoriale è identificabile mentre il modello con 2^R classi latenti non lo è.

La tabella 3.1 verifica l'equivalenza nel numero dei parametri e dei relativi gradi di libertà tra vari modelli a classi latenti standard e i corrispondenti fattoriali

⁴ Si veda il paragrafo 2.2.12.

base nel caso di cinque variabili manifeste dicotomiche. Dalla tabella è possibile inoltre calcolare, ad esempio, che il modello base con due fattori richiede $23 - 17 = 6$ parametri in meno rispetto al modello con quattro classi latenti. Questa riduzione corrisponde alle cinque restrizioni $u_{avw}^{AVW} = u_{bvw}^{BVW} = u_{cvw}^{CVW} = u_{dvw}^{DVW} = u_{evw}^{EVW} = 0$, più il vincolo che attesta l'indipendenza tra V e W .

Infine, si nota un'importante differenza tra il modello a classi latente fattoriale ed i modelli con molte classi latenti proposte in letteratura (ad esempio, Goodman, 1974; Haberman, 1979; Hagenaaars, 1990). Il modello a classi latenti fattoriale base appena descritto include tutti i pesi fattoriali tra le variabili latenti e gli indicatori. Questo significa che nessuna assunzione viene richiesta sulle relazioni tra le variabili manifeste e quelle latenti, rendendo il modello fattoriale più adatto per l'analisi esplorativa dei dati rispetto ai modelli a classi latenti con molte classi che, invece, necessitano di restrizioni nelle associazioni tra indicatori e variabili latenti.

Modelli a classi latenti standard			Modelli a classi latenti fattoriali base		
# classi latenti	# parametri	# gradi di libertà	# fattori	# parametri	# gradi di libertà
1	5	26	0	5	26
2	11	20	1	11	20
3	17	14	2	17	14
4	23	8	3	23	8
5	29	2	4	29	2

Tabella 3.1: relazione di equivalenza tra i modelli a classi latenti standard e fattoriali (esempio con cinque variabili manifeste dicotomiche)

3.1.1 L'interpretazione dei fattori

Il confronto tra l'approccio dell'analisi a classi latenti di tipo esplorativo e quello fattoriale effettuato sui risultati ottenuti con diversi data set mostra come

quest'ultimo conduca ad un modello più parsimonioso e più facile da interpretare nella quasi totalità dei casi (Magidson e Vermunt, 2001).

Nei modelli a classi latenti fattoriali, i fattori sono di solito interpretabili, come nell'analisi fattoriale tradizionale, come un segmento orientato che raggruppa le variabili che condividono una sorgente di variazione comune. Oltre alla possibilità di considerare variabili di diversa tipologia (categoriche o continue), la differenza principale di questo approccio rispetto all'analisi fattoriale tradizionale consiste nel fatto che non servono rotazioni affinché le soluzioni risultino unicamente identificate e interpretabili.

Quando si applica un modello a classi latenti non è noto il numero delle dimensioni che la soluzione riesce a catturare: ad esempio, un modello con tre classi può descrivere una o due dimensioni, mentre un modello con quattro classi può descriverne una, due oppure tre. Quando un modello a tre classi latenti descrive una dimensione, è molto probabile che un modello con un fattore con 3 o più livelli rappresenti i dati quasi con lo stesso adattamento ma, come si è mostrato, con un numero di parametri inferiore.

Nella maggior parte dei casi nei quali la stima dei modelli a classi latenti suggerisce la necessità di almeno due dimensioni per ottenere un adattamento adeguato ai dati, è ragionevole attendersi che vi siano due fattori sottostanti in grado di spiegare i livelli "basso" e "alto" di ogni dimensione. Nel caso di variabili manifeste dicotomiche, se quattro classi latenti fossero interpretabili come una soluzione bi-dimensionale dove le classi corrispondono alle combinazioni (basso, basso), (alto, basso), (basso, alto) e (alto, alto), questo spiegherebbe il perché il modello a classi latenti fattoriale base, solitamente, dia un migliore adattamento ai dati. Ad esempio, quando un modello a tre classi latenti descrive due dimensioni, c'è lo svantaggio di non riuscire a catturare tutte le quattro combinazioni base delle due dimensioni. In questo caso, il modello 2-fattoriale risulta approssimare meglio le relazioni esistenti nei dati.

In altri termini, è possibile interpretare ogni fattore attraverso alcune delle variabili manifeste considerate che, graficamente, si dispongono lungo il segmento che raffigura il fattore come mostrato in figura 3.1. In questo esempio, i segmenti che identificano i due fattori del modello descrivono le associazioni tra le quattro variabili manifeste dicotomiche *A*, *B*, *C* e *D* e possono essere caratterizzati nel seguente modo: il fattore 1, posizionato in orizzontale, viene definito dalle variabili *A* e *D*, mentre il fattore 2, orientato in verticale in figura 3.1, spiega la relazione delle altre due variabili manifeste (*B* e *C*).

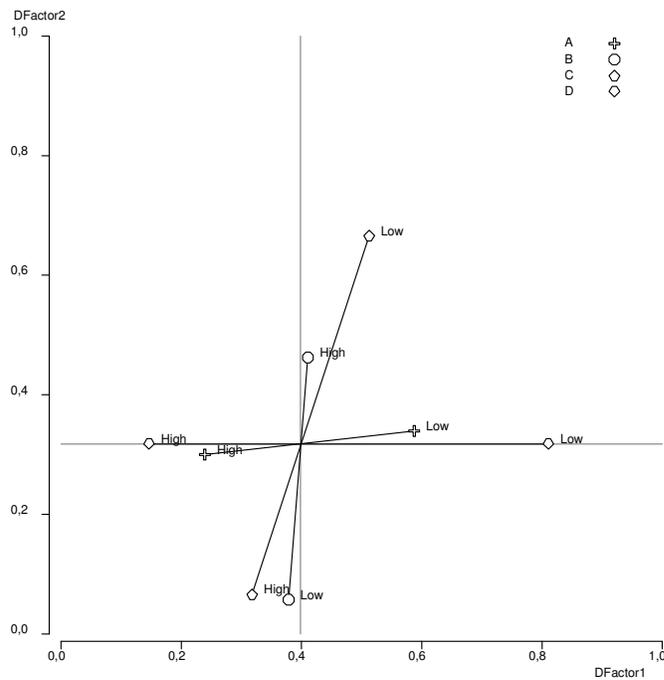


Figura 3.1: il modello a classi latenti fattoriale con 2 fattori e 4 indicatori

3.1.2 Le estensioni al modello a classi latenti fattoriale base

In alcune situazioni può essere utile apportare alcune modifiche ed estensioni al modello a classi latenti fattoriale base.

Primo, ove applicabile, è possibile trattare le variabili manifeste categoriche come ordinabili in diversi modi. Nel caso di una variabile A tricotomica, l'approccio più semplice è quello di assumere la categoria centrale come equidistante dalle altre e modificare il modello log-lineare descritto nell'equazione (3.2) usando la variabile uniforme $v_a^A = \{0 \text{ se } a = 1; 0,5 \text{ se } a = 2; 1 \text{ se } a = 3\}$ per le categorie della variabile A : il parametro del modello diventa perciò $u_{av}^{AV} = u_{av}^{AV} v_a^A$.

Secondo, analogamente all'analisi fattoriale confermatrice, si può assumere che i due fattori V e W siano correlati e vincolare alcune variabili manifeste (ad esempio, A e B) al primo fattore ed altre (C e D) al secondo. La rappresentazione log-lineare per il modello confermatrice di questo tipo si ottiene ponendo $u_{vw}^{VW} \neq 0$ e $u_{av}^{AW} = u_{bw}^{BW} = u_{cv}^{CV} = u_{dv}^{DV} = 0$. Ovviamente, un buon adattamento del modello vincolato ai dati corrisponde ad una conferma delle assunzioni a priori che sono state imposte.

Terzo, finora si sono considerati solo fattori dicotomici. È possibile estendere il modello fattoriale in modo che ogni fattore possa contenere più di due livelli ordinati, assegnando dei pesi numerici tra zero ed uno equidistanti tra loro ai livelli del fattore. Heinen (1996) usa questa strategia per definire dei modelli a classi latenti simili a certi modelli a tratto latente. L'uso di pesi fissi per i livelli dei fattori nei termini di interazione garantisce che ogni fattore catturi una singola dimensione. Per i fattori con più di due livelli, si considerano le medie condizionate $\hat{E}_a^{VIA} = \sum_{v=1}^R \hat{\pi}_{va}^{VIA} v_v^V$ invece che le probabilità condizionate $\hat{\pi}_{va}^{VIA}$. Si noti che se si assegna il valore zero al primo livello ed uno per l'ultimo, per fattori dicotomici la media condizionale è uguale alla probabilità condizionale di essere nel livello 2.

Infine, come per il modello a classi latenti, l'inclusione di covariate e variabili esterne è un'estensione piuttosto semplice ed immediata.

3.2 Il Latent Markov Model di mistura

Il latent Markov model di mistura (MLMM, Vermunt et al., 2007) è un'estensione del LMM descritto nel paragrafo 2.6 nel quale, oltre alla variabile latente dinamica Y_t viene inclusa anche una variabile latente statica X . In maniera analoga, il MLMM può essere visto anche come un'estensione del modello a classi latenti che, oltre alla variabile latente X , presenta la variabile latente Y_t , funzione del tempo $t = 1, \dots, T$, in grado di spiegare la dinamica temporale ed il passaggio di un'osservazione tra S stati latenti.

In particolare, il modello MLMM consiste in tre tipi di variabili:

1. una serie storica z_{ht} che esprime l'osservazione h al tempo t , con $h = 1, \dots, n$ e $t = 1, \dots, T$;
2. una variabile latente categorica X costante nel tempo e suddivisa in K classi;
3. una variabile latente categorica Y_t che varia in funzione del tempo ed è definita da S stati latenti.

Nel resto del paragrafo, per una maggiore chiarezza, con x si intende la classe latente (per $x = 1, \dots, K$) e con y_t lo stato latente (per $y_t = 1, \dots, S$ e $t = 1, \dots, T$). La variabile X rappresenta, come nel modello a classi latenti, la suddivisione delle osservazioni in gruppi tra loro eterogenei ma che presentano una forte omogeneità al loro interno. La variabile latente funzione del tempo con struttura di transizione markoviana Y_t aggiunge flessibilità al modello, considerando la distribuzione delle osservazioni temporali e catturando i cambiamenti che avvengono in punti temporali adiacenti. In figura 3.1 viene mostrato il modello MLMM.

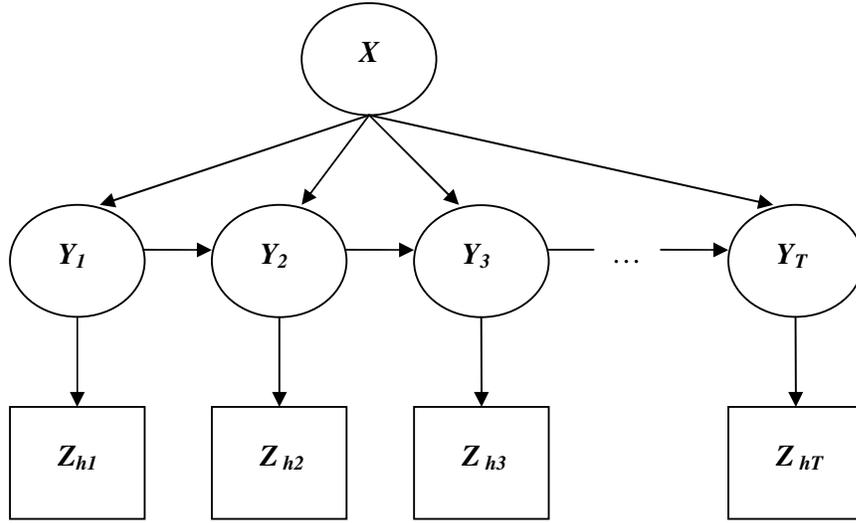


Figura 3.1: il modello LMM di mistura (MLMM)

Analogamente alle equazioni (2.23)-(2.25), il modello MLMM è definito come

$$f(Z_h) = \sum_{x=1}^K \sum_{y_1=1}^S \sum_{y_2=1}^S \dots \sum_{y_T=1}^S f(x, y_1, \dots, y_T) f(Z_h; x, y_1, \dots, y_T) \quad (3.3)$$

dove

$$f(x, y_1, \dots, y_T) = f(x) f(y_1 | x) \prod_{t=2}^T f(y_t | y_{t-1}, x) \quad (3.4)$$

e

$$f(Z_h; x, y_1, \dots, y_T) = \prod_{t=1}^T f(z_{ht} | y_t). \quad (3.5)$$

Come per la (2.23), l'equazione (3.3) mostra che il modello è una mistura contenente una variabile latente statica e T variabili latenti che variano nel tempo. Il numero totale di classi latenti (o componenti di mistura) è perciò uguale a KS^T , che è il prodotto del numero di categorie di x e y_t per $t = 1, \dots, T$.

Le equazioni (3.4), (3.5) e la figura 3.1 mostrano l'assunzione di indipendenza condizionata imposta dal modello LMM. In particolare, l'equazione (3.4) evidenzia il fatto che all'interno della classe latente x , y_t viene assunta come una catena di Markov del primo ordine, mentre l'equazione (3.5) mostra che condizionatamente a y_t , l'osservazione al tempo t risulta indipendente dalle osservazioni, classi latenti e stati latenti di altri istanti temporali (assunzione di indipendenza condizionale).

Il modello è caratterizzato da quattro funzioni di probabilità:

1. $f(x)$ è la probabilità di appartenere alla classe latente x , la probabilità a priori $\pi_x = P(X = x)$;
2. $f(y_1 | x)$ è la probabilità dello stato iniziale, ovvero la probabilità di essere in un particolare stato latente al tempo 1 condizionatamente all'appartenenza alla classe latente x : $\lambda_{jx} = P(y_1 = j | X = x)$;
3. $f(y_t | y_{t-1}, x)$ è la probabilità di transizione latente, ovvero la probabilità di essere in un particolare stato latente al tempo t condizionatamente allo stato latente occupato al tempo $t - 1$ e all'appartenenza alla classe latente: assumendo un processo di transizione omogeneo rispetto al tempo, si ha $p_{jkx} = P(y_t = j | y_{t-1} = k, X = x)$ contenute nelle K matrici di transizione P_x ;
4. $f(z_{ht} | y_t)$ è la funzione di densità delle osservazioni, che esprime la densità di probabilità di avere un determinata osservazione h al tempo t condizionatamente allo stato latente occupato al tempo t . Se la distribuzione di riferimento è di tipo gaussiano, questa è caratterizzata da un vettore di parametri $\theta_j = (\mu_j, \sigma_j^2)$ contenente le medie e le varianze del generico stato latente j , invarianti rispetto alle classi latenti.

Nel caso di distribuzione normale delle osservazioni, il numero dei parametri del modello MLMM è pari a $S(KS + 2) - 1$ e consiste in $K - 1$ classi latenti, $K(S - 1)$ stati iniziali, $KS(S - 1)$ probabilità di transizione e $2K$ medie e varianze condizionate delle variabili osservate.

3.2.1 I casi speciali del modello MLMM vincolato

Vari casi speciali del modello MLMM definito dalle equazioni (3.3)-(3.5) possono essere ottenuti eliminando uno o più dei tre elementi principali: la struttura di autocorrelazione di tipo markoviano, le variabili latenti varianti nel tempo oppure la variabile latente costante nel tempo.

Per esempio, come detto, il modello latent Markov può essere visto come caso particolare del modello MLMM ottenuto eliminando la variabile latente costante nel tempo x , ovvero assumendo che non vi sia eterogeneità non osservabile nei dati. Questo modello si ottiene ponendo $K = 1$, il che implica che tutte le osservazioni appartengano alla stessa classe latente.

Se si assume che non vi sia autocorrelazione tra le variabili latenti y_t , così che $P(y_t | y_{t-1}, x) = P(y_t | x)$, si ottiene il modello di mistura multilevel o gerarchico (Vermunt, 2003, 2007). Questo mostra come il modello MLMM possa essere visto come un modello di mistura che viene esteso con una struttura di autocorrelazione.

Infine, nel caso la variabile osservata sia continua, il modello di mistura gaussiano può essere visto come una variante del modello MLMM vincolato ottenuto rimuovendo le variabili latenti y_t : $f(Z_h) = \sum_{x=1}^K f(x) \prod_{t=1}^T f(z_{ht} | x)$. Questo modello è equivalente al modello a classi latenti per T variabili manifeste con l'assunzione che le osservazioni temporali siano condizionatamente indipendenti all'interno della classe latente x .

3.2.2 La stima dei parametri

Come per il modello LMM (si veda il paragrafo **2.6.1**), anche i parametri del MLMM vengono stimati massimizzando la funzione di log-verosimiglianza e ricorrendo alla variante dell'EM nota come algoritmo di Baum-Welch.

Oltre alle probabilità forward α_{ht} e backward β_{ht} , è necessario aggiungere anche la quantità $\gamma_h(x)$ riferita all'eterogeneità latente presente nei dati. Queste tre quantità sono definite come

$$\begin{aligned}\alpha_{ht}(x, j) &= f(y_t = j, z_{h1}, \dots, z_{ht} \mid X = x) \\ \beta_{ht}(x, j) &= f(z_{ht+1}, \dots, z_{hT} \mid y_t = j, X = x) \\ \gamma_h(x) &= f(X = x, Z_h)\end{aligned}$$

che conducono alle distribuzioni posteriori marginali

$$\begin{aligned}f(X = x \mid Z_h) &= \frac{\gamma_h(x)}{f(Z_h)} \\ f(X = x, y_t = j \mid Z_h) &= \frac{\alpha_{ht}(x, j)\beta_{ht}(x, j)}{f(Z_h)} \\ f(X = x, y_{t-1} = k, y_t = j \mid Z_h) &= \\ &= \frac{\gamma_h(x)\alpha_{ht-1}(x, k)f(y_t = j \mid y_{t-1} = k)f(z_{ht} \mid y_t = j, X = x)\beta_{ht}(x, j)}{f(Z_h)}\end{aligned}$$

Gli schemi ricorsivi per le probabilità forward α_{ht} e backward β_{ht} sono i medesimi di quelli descritti nel paragrafo **2.6.1**, mentre la quantità $\gamma_h(x)$ viene ottenuta come

$$\gamma_h(x) = \sum_{j=1}^S f(X = x)\alpha_{ht}(j)\beta_{ht}(j)$$

per ogni t . Quindi, prima si ottengono le probabilità $\alpha_{ht}(j)$ e $\beta_{ht}(j)$ per ogni istante temporale e conseguentemente si ricava $\gamma_h(x)$. Successivamente, si calcolano $f(X = x \mid Z_h)$, $f(X = x, y_t = j \mid Z_h)$ e $f(X = x, y_{t-1} = k, y_t = j \mid Z_h)$

usando le equazioni precedenti e, al passo M , queste quantità sono usate per ottenere le nuove stime dei parametri del modello MLMM.

3.2.3 La scelta del numero di classi latenti e stati latenti

Un altro importante argomento è quello legato alla scelta del numero delle classi e degli stati latenti necessari o, in altre parole, su come stimare e/o scegliere K e S . L'approccio tradizionale per la selezione del modello quando si ricorre alla stima di massima verosimiglianza è quello di scegliere il miglior modello considerando i test del rapporto di verosimiglianza tra modelli annidati. Purtroppo, nel contesto dei modelli a classi latenti questo approccio è problematico in quanto l'ipotesi nulla da testare è definita sul confine dello spazio parametrico e, conseguentemente, la condizione di regolarità di Cramer sulle proprietà asintotiche della stima di massima verosimiglianza non sono più valide sotto l'ipotesi nulla. Come alternativa, si può determinare la bontà del modello attraverso alcuni criteri di informazione. Il principio alla base di queste statistiche è, come visto in precedenza, quello della parsimonia e permette di valutare sia l'adattamento che la complessità del modello. Le statistiche di informazione più usate sono il criterio bayesiano (BIC) e quello di Akaike (AIC). Esistono studi di simulazione che mostrano come l'AIC tenda a sovrastimare il numero di parametri del modello (Dias e Vermunt, 2007), perciò è stata proposta una variante, il Consistent Akaike Information Criterion (CAIC), che penalizza la complessità del modello in maniera più marcata. Per cui, K e S possono essere scelti in modo da minimizzare il valore delle statistiche BIC e/o CAIC.

3.3 Il Latent Markov Model misto

Il latent Markov model misto (MHMM⁵) permette un approccio multivariato in cui si considerano processi multipli ed estende la classe dei LMM ammettendo l'incorporazione di covariate e di effetti casuali sia nella parte condizionata del modello, sia in quella latente.

MacKay (2003) individua due vantaggi fondamentali. In primo luogo, modellare processi multipli simultaneamente permette la stima degli effetti a livello di popolazione, oltre ad una stima più efficiente dei parametri che sono comuni a tutti i processi ed agli effetti a livello individuale, riconoscendo le differenze tra i diversi processi. In secondo luogo, il MHMM permette una maggiore flessibilità nel modellare la struttura di correlazione, in quanto viene rilassata l'assunzione che le osservazioni siano indipendenti dati gli stati latenti.

3.3.1 Il MHMM con effetti casuali per il processo osservato

Si denota l'osservazione z_{ht} riferita all'unità h -esima, con $h = 1, \dots, n$, al tempo t , per $t = 1, \dots, T$, ed il relativo stato latente Y_{ht} che assume valori da un insieme finito $\{1, 2, \dots, S\}$, dove S è noto a priori. Inoltre, si assume che, condizionatamente agli effetti casuali, $\{Y_{ht}\}_{t=1}^T$ sia una catena di Markov e, nel caso sia condizionatamente stazionaria con distribuzioni stazionarie uniche, è possibile usare tali distribuzioni come probabilità iniziali. In altri termini, si ottengono le distribuzioni iniziali sulle basi delle probabilità di transizione, in modo tale che queste probabilità possano variare tra le unità. Altrimenti, si assume che le probabilità iniziali siano dei parametri fissi uguali per tutte le unità. Infine, si

⁵ Per differenziare l'abbreviazione da quella del latent Markov model di mistura presentato nel paragrafo 3.2, si riprende il nome usato in letteratura, *Mixed Hidden Markov Model*, e la relativa sigla MHMM (MacKay, 2003, 2007).

assume che, condizionatamente agli effetti casuali, $\{Z_{ht}\}_{t=1}^T$ sia un LMM e che tali modelli LMM siano indipendenti gli uni dagli altri.

Il MHMM con effetti casuali nella parte condizionata del modello per i dati osservati assume che gli effetti casuali non appaiano nel modello per i processi latenti. In particolare, si assume che i processi latenti siano omogenei con probabilità di transizione e probabilità iniziali comuni a tutte le unità.

Ricorrendo alla specificazione dei modelli lineari generalizzati misti (si veda, ad esempio, McCulloch e Searle, 2001), si assume che, condizionatamente agli effetti casuali u e agli stati latenti Y , $\{Z_{ht}\}_{t=1}^T$ siano indipendenti con distribuzione nella famiglia esponenziale

$$f(z_{ht} | Y_{ht} = j, u, \theta) = \exp\{(z_{ht} \eta_{htj} - k(\eta_{htj})) / a(\phi) + d(z_{ht}, \phi)\}$$

dove θ è il vettore dei parametri del modello che comprende anche il parametro \square comune a tutti gli stati latenti.

In questa specificazione,

$$\eta_{htj} = \tau_j + c'_{ht} \beta_j + w'_{htj} u$$

dove τ_s denota l'effetto fisso quando $Y_{ht} = j$, c'_{ht} sono le covariate per l'unità h al tempo t e w'_{htj} è la riga della matrice degli effetti casuali per l'unità h -esima al tempo t ed appartenente allo stato j .

Indicando con $f(u, \theta)$ la distribuzione degli effetti casuali e assumendoli indipendenti dagli stati latenti, la funzione di verosimiglianza del modello proposto risulta

$$\begin{aligned}
L(\theta) &= f(Z; \theta) = \int_u \sum_y f(Z | Y, u, \theta) f(Y | u, \theta) f(u; \theta) du \\
&= \int_u \sum_y f(Z | Y, u, \theta) f(Y; \theta) f(u; \theta) du \\
&= \int_u \sum_y \left\{ \prod_{h=1}^n \prod_{t=1}^T f(z_{ht} | y_{ht}, u, \theta) \right\} \left\{ \prod_{h=1}^n f(y_{h1}) \prod_{t=2}^T f(y_{ht} | y_{ht-1}) \right\} f(u; \theta) du \\
&= \int_u \sum_y \left\{ \prod_{h=1}^n f(y_{h1}) f(z_{h1} | y_{h1}, u, \theta) \prod_{t=2}^T f(y_{ht} | y_{ht-1}) f(z_{ht} | y_{ht}, u, \theta) \right\} f(u; \theta) du
\end{aligned}$$

e, per semplificare l'espressione scrivendo la sommatoria come un prodotto di matrici, si valutano, dato un valore di u , il vettore A^{h1} con elementi $A_j^{h1} = f(y_{ht} = j) f(z_{h1} | y_{h1} = j, u)$, la matrice A^{ht} dove gli elementi sono $A_{jk}^{ht} = f(y_{ht} = j | y_{ht-1} = k) f(z_{ht} | y_{ht} = j, u)$ e il vettore S -dimensionale di 1, $\mathbf{1}$, ottenendo

$$L(\theta) = \int_u \prod_{h=1}^n \left\{ (A^{h1})^i \left(\prod_{t=2}^T A^{ht} \right) \mathbf{1} \right\} f(u; \theta) du .$$

Perciò, l'unico impatto dell'inclusione degli effetti casuali sulla funzione di verosimiglianza consiste nell'introduzione di un'integrazione sulla distribuzione degli u . In altri termini, a differenza della verosimiglianza del modello LMM classico, si aggiunge la sola complessità dell'integrale in u . MacKay (2007) dimostra che, in molte applicazioni, $L(\theta)$ si riduce a forme molto semplici.

L'algoritmo EM può sembrare una scelta naturale per la stima dei parametri, valutando gli effetti casuali come variabili mancanti al pari dei valori di $\{Y_{ht}\}$. Tuttavia, si riscontrano i problemi di efficienza già espressi in precedenza. Nei suoi lavori, MacKay suggerisce dei metodi Monte Carlo (MC) utili per aggirare la valutazione dell'integrale e, perciò, utilizzabili sia per massimizzare direttamente la funzione di verosimiglianza, sia per implementare l'algoritmo EM. Ad esempio, si può ricorrere agli algoritmi nelle versioni MC di Newton-Raphson o EM presentati da McCulloch (1997) nel contesto dei modelli lineari generalizzati misti.

Data la complessità del MHMM, è importante essere in grado di interpretare sia gli effetti fissi che quelli casuali. Un modo per capire il loro impatto sul modello è quello di esaminare i momenti marginali del processo osservato risultanti. Usando le proprietà delle famiglie esponenziali, $E(Z_{ht} | Y_{ht} = j, u) = k'(\eta_{hj})$ e $Var(Z_{ht} | Y_{ht} = j, u) = k''(\eta_{hj})a(\phi)$, oltre alla assunzione che $Cov(Z_{ht}, Z_{h't'} | Y_{ht} = j, Y_{h't'} = k, u) = 0$, allora

$$\begin{aligned} E(Z_{ht}) &= E(k'(\eta_{hj})) \\ Var(Z_{ht}) &= E(k''(\eta_{hj}))a(\phi) + Var(k'(\eta_{hj})) \\ Cov(Z_{ht}, Z_{h't'}) &= Cov(k'(\eta_{hj}), k'(\eta_{h't'k})) \text{ per } t' < t \end{aligned}$$

Generalmente, questi momenti non hanno una forma chiusa. Tuttavia, per alcune distribuzioni comuni di $f(Z_{ht} | Y_{ht}, u)$, ad esempio la normale o la Poisson, e degli effetti casuali, ad esempio la normale multivariata, esistono delle forme chiuse (si veda l'esempio di MacKay (2003), pag. 59).

Includendo gli effetti casuali nella parte condizionata del modello per i dati osservati implica il rilassamento dell'assunzione che le osservazioni siano condizionatamente indipendenti dati gli stati latenti.

3.3.2 Il MHMM con effetti casuali per il processo latente

Può essere talvolta desiderabile ammettere che sia la catena di Markov latente a variare tra le unità osservate. In altri termini, le osservazioni possono trascorrere proporzioni di tempo differenti nei vari stati latenti. Tuttavia, includere gli effetti casuali nel processo latente del LMM è un problema impegnativo, a prescindere che vi siano effetti casuali anche nel modello condizionato per i dati osservati.

In questa specificazione, si assume che $\{Y_{ht} | u\}_{t=1}^T$ sia una catena di Markov e che $Y_{ht} | u$ sia indipendente da $Y_{h't'}$ per $h \neq h'$. Affinché le probabilità di transizione giacciono tra 0 ed 1 e che le righe della matrice di transizione sommino ad 1, si modellano le p_{jk} come

$$P(y_{ht} = j \mid y_{ht-1} = k, u, \theta) = \frac{\exp\{\tau_{jk} + c'_{ht} \beta_{jk} + w'_{htjk} u\}}{\sum_{j=1}^S \exp\{\tau_{jk} + c'_{ht} \beta_{jk} + w'_{htjk} u\}}.$$

La funzione di verosimiglianza del modello può essere scritta come

$$L(\theta) = \int \prod_{h=1}^n \left\{ (A^{h1})' \left(\prod_{t=2}^T A^{ht} \right) \mathbf{1} \right\} f(u; \theta) du$$

dove, in questo caso, $A_j^{h1} = f(y_{ht} = j) f(z_{h1} \mid y_{h1} = j, u, \theta)$ e $A_{jk}^{ht} = f(y_{ht} = j \mid y_{ht-1} = k, u, \theta) f(z_{ht} \mid y_{ht} = j, u, \theta)$. Questa $L(\theta)$ può sembrare molto simile alla precedente, tuttavia l'integrale risulta piuttosto complicato, anche in situazioni semplici e questo rende la stima dei parametri un problema difficile da risolvere.

Un altro problema nell'aggiungere gli effetti casuali nel modello per il processo latente consiste nella difficoltà della valutazione del loro impatto sul modello in generale e sui momenti marginali in particolare. Le espressioni dei momenti risultano le stesse di quelle proposte nel paragrafo **3.3.1**, tuttavia l'integrazione risulta molto più difficile in questo contesto.

Tuttavia, si può interpretare l'impatto degli effetti casuali sulla covarianza asintotica. Considerando il caso in cui $\{Y_{ht} \mid u\}$ sia omogenea e stazionaria con distribuzione stazionaria unica e dove $c_{ht} \equiv c_h$ e $w_{htj} \equiv w_h$ sono indipendenti da t e j , si dimostra che $Cov(Z_{ht}, Z_{ht'}) \rightarrow \kappa$ se $|t-t'| \rightarrow \infty$, dove κ è una costante positiva. Si ricordi che $Cov(Z_{ht}, Z_{ht'}) \rightarrow 0$ se non vi sono effetti casuali nel modello, ovvero quando si assume lo stesso modello per ogni unità osservata. Perciò, gli effetti casuali permettono una dipendenza positiva di lungo termine per ognuna delle osservazioni di una unità.

In definitiva, il ruolo degli effetti casuali nella parte del modello per il processo latente permette di rilassare l'assunzione che il processo sia markoviano.

Tuttavia, sia la procedura di stima che l'interpretazione dei risultati sono problemi ancora irrisolti e necessitano di ulteriori contributi.

4. Soluzioni metodologiche per l'analisi dei fenomeni finanziari: il caso del mercato azionario italiano

Il ricorso ai metodi statistici a variabili latenti nell'analisi dei fenomeni finanziari permette di sottolineare l'importanza che la dimensione metodologica ricopre in questi studi e che, tuttavia, viene frequentemente trascurata. Inoltre, la stima dei modelli a struttura latente permette di proporre delle soluzioni innovative ad alcuni problemi rilevanti in campo finanziario.

In questo capitolo, l'efficacia delle proposte metodologiche presentate nel capitolo **1** viene valutata facendo riferimento ai modelli statistici per lo studio delle variabili latenti descritti nei capitoli **2** e **3** al fine di ottenere una analisi del mercato azionario italiano.

Il paragrafo **4.1** è dedicato alla descrizione delle principali caratteristiche del mercato azionario italiano, mentre nel paragrafo **4.2** viene introdotto l'insieme dei dati oggetto di studio e vengono proposte alcune analisi preliminari.

Il primo contributo metodologico proposto nel paragrafo **4.3** valuta i modelli a classi latenti per variabili manifeste dicotomiche introdotti nel paragrafo **2.2** e l'analisi a profilo latente descritta nel paragrafo **2.3** per assegnare le singole attività finanziarie in classi diverse al fine di ottenere una nuova classificazione azionaria. La nuova classificazione viene, quindi, valutata rispetto al profilo di rischio – rendimento atteso che permette di discriminare le caratteristiche finanziarie fondamentali delle diverse classi latenti (paragrafo **4.4**). La definizione del profilo rischio-rendimento di ogni gruppo consente di definire due ulteriori proposte metodologiche innovative nelle analisi finanziarie: (i) un miglioramento dei processi di diversificazione del rischio e (ii) un metodo decisionale avanzato e corretto per la scelta di un portafoglio efficiente di attività rischiose.

Inoltre, in un ambito più aggregato, la nuova classificazione azionaria viene usata per analizzare i problemi riferiti alla costruzione e alla rappresentatività dei

numeri indici dei prezzi finanziari. In particolare, lo studio proposto nel paragrafo **4.5** valuta due aspetti metodologici fondamentali: i processi di definizione del paniere di titoli sottostante e gli effetti che una diversa struttura di ponderazione implica sulla performance degli indici finanziari.

Infine, l'analisi proposta nel paragrafo **4.6** illustra come il latent Markov model presentato nel paragrafo **2.6** possa essere impiegato per studiare l'andamento dinamico del mercato azionario italiano. In particolare, includendo esplicitamente nell'analisi a struttura latente la dimensione temporale è possibile esaminare la serie storica dei rendimenti del principale indice del mercato italiano per riconoscere i diversi regimi che la caratterizzano. L'identificazione e l'interpretazione rispetto al profilo di rischio-rendimento dei regimi e della loro alternanza nel tempo permettono alcune valutazioni finanziarie di forte interesse: (i) l'individuazione dei periodi di crisi, (ii) la definizione della loro fine ed il passaggio ad un regime di stabilità e (iii) la possibilità di prevedere gli sviluppi futuri del mercato.

4.1 Il mercato azionario italiano

Il mercato azionario italiano è gestito da Borsa Italiana, società per azioni istituita nel 1998 a seguito della privatizzazione dei mercati borsistici. Secondo la World Federation of Exchanges (<http://www.world-exchanges.org>), nel 2008 la Borsa Valori di Milano si posizionava al quindicesimo posto nella graduatoria dei principali mercati azionari mondiali in termini di capitalizzazione ed al settimo in Europa dietro al gruppo Euronext che comprende le Borse di Parigi, Amsterdam, Bruxelles e Lisbona, la Borsa Valori di Londra, il mercato azionario tedesco, spagnolo, svizzero ed il gruppo OMX Nordic che comprende le Borse di Svezia, Finlandia e Danimarca.

Dal 1 aprile 2001, le azioni del Mercato Telematico Azionario (MTA) sono ripartite da Borsa Italiana in segmenti di mercato omogenei in funzione della loro capitalizzazione⁶. La segmentazione adottata tenta di rappresentare al meglio la struttura dell'economia italiana, caratterizzata da società di grandi dimensioni e da un numero elevato di piccole e medie imprese, e di rispondere alle esigenze degli investitori, aiutandoli a specializzare le proprie decisioni di investimento e a diversificare i propri portafogli finanziari. Tale ripartizione è effettuata all'atto dell'ammissione delle azioni sul listino e monitorata con periodicità semestrale. Sostanzialmente, le azioni sono suddivise in base alla capitalizzazione di mercato in due gruppi: il primo è il segmento delle cosiddette *Blue-Chip* ed è formato dalle società con alta capitalizzazione (superiore a 1.000 milioni di euro); il secondo, diviso a sua volta nei segmenti Standard e STAR, è formato dai titoli di società con media o bassa capitalizzazione (inferiore a 1.000 milioni di euro e superiore a 40 milioni di euro). A sua volta, il segmento Standard è diviso in due classi sulla base della frequenza degli scambi e del controvalore medio giornaliero negoziato

⁶ Sul MTA possono essere quotati e negoziati anche diritti di opzione, warrant e obbligazioni convertibili. In questo lavoro si fa riferimento solo ai titoli azionari.

nel semestre precedente. La differenza tra lo Standard e lo STAR (Segmento Titoli con Alti Requisiti) deriva dal fatto che in quest'ultimo sono negoziate le azioni di piccole e medie imprese con buone prospettive, che devono soddisfare requisiti aggiuntivi rispetto a quelli richiesti per l'ammissione allo Standard e rispettare impegni particolari in termini di liquidità, trasparenza e corporate governance.

La riclassificazione delle azioni avviene in occasione della revisione ordinaria dell'indice S&P-MIB (a marzo ed a settembre): Borsa Italiana individua le società non comprese nel segmento Blue-Chip la cui capitalizzazione sia divenuta superiore al miliardo di euro e le trasferisce a tale segmento; viceversa, se la capitalizzazione delle azioni di società del segmento Blue-Chip è scesa al di sotto della soglia del miliardo di euro, esse saranno declassate se entro sei mesi non viene ripristinato tale livello. Le revisioni riguardano anche la ripartizione delle società a media e bassa capitalizzazione: nel caso in cui vengano meno le condizioni che hanno permesso l'inserimento nel segmento STAR, le azioni sono trasferite al segmento Standard. Borsa Italiana con l'ausilio di società di *index provider* quali Standard and Poor's (S&P) e, dal 30 marzo 2009, FTSE Group elabora, calcola ed aggiorna una serie di indici di borsa che rappresentano il valore di un determinato mercato o segmento.

Uno dei principali indici di Borsa Italiana è il MIBTEL. Questo indice dei prezzi "riassume" il valore di tutte le azioni quotate sul MTA. Il MIBTEL viene aggiornato ogni minuto durante la fase di negoziazione continua sulla base dei prezzi degli ultimi contratti conclusi su ciascuna azione componente il paniere. La base dell'indice risale al 3 gennaio 1994 ed è pari a 10000 punti. Al 31 dicembre 2008 il MIBTEL valeva 15096 punti.

L'indice S&P-MIB costituisce il principale benchmark del mercato borsistico italiano e comprende le azioni delle 40 maggiori società italiane ed estere quotate sul MTA rappresentando circa l'80% della capitalizzazione del mercato azionario

italiano. Lo S&P-MIB è stato introdotto il 2 giugno 2003 e calcolato partendo dal 31 dicembre 1997 (base = 10000). Al 31 dicembre 2008 il suo valore corrispondeva a 19460 punti. I criteri di selezione dei titoli appartenenti all'indice si basano sulla classificazione settoriale, sulla capitalizzazione del flottante e sulla liquidità. La scelta dei titoli appartenenti al paniere dell'indice, le revisioni ordinarie effettuate a marzo e a settembre di ogni anno e quelle straordinarie a seguito di operazioni sul capitale (incremento del numero di azioni superiore al 5%), di variazioni rilevanti del flottante (superiore al 5%), di scorpori di rami d'azienda (*spin off*), fusioni o cancellazione dalle contrattazioni (*delisting*) sono compiti della commissione congiunta composta da rappresentanti di S&P e di Borsa Italiana (*Index Committee*) con lo scopo di mantenere più che mai aggiornato l'indice borsistico alla realtà economica italiana. La metodologia di calcolo dello S&P-MIB avviene ponderando il prezzo delle azioni facenti parte il paniere alla capitalizzazione del flottante. Il termine flottante indica la quantità di quote azionarie emesse da una società che gli investitori possono liberamente scambiare sul mercato azionario. Di questo ammontare di quote non fanno parte la quantità di capitale che costituisce la partecipazione di controllo, le quote vincolate da patti parasociali e di quelle soggette ai vincoli alla trasferibilità (come clausole di *lock-up*) di durata superiore ai 6 mesi. Inoltre, nel calcolo del flottante non si tiene conto delle partecipazioni azionarie superiori ad una determinata soglia (per S&P questa soglia è fissata al 5%). Al contrario, rientrano nel computo le azioni possedute da organismi di investimento collettivo del risparmio, da fondi pensione e da enti previdenziali. Il valore del flottante non corrisponde alla capitalizzazione di una società quotata. La capitalizzazione risulta essere infatti il valore, al prezzo di mercato, di tutte le azioni che l'azienda ha emesso. Perciò, il valore del flottante risulta essere una parte della capitalizzazione: quando il flottante coincide con la capitalizzazione significa che tutte le azioni emesse sono "in mano" al mercato e quindi oggetto di contrattazione e non detenute da soggetti

dominanti. Il problema principale alla base del flottante consiste nel fatto che non esiste una procedura comune per la definizione di questa quantità: index provider diversi calcolano il flottante applicando regole differenti.

A seguito della fusione di Borsa Italiana con il London Stock Exchange (LSE), dal 1 giugno 2009 l'indice S&P-MIB è stato sostituito dal FTSE-MIB mantenendo però le medesime caratteristiche del suo predecessore.

Un altro indice relativo ai titoli quotati nel segmento Blue-Chip e non inclusi nel paniere dello S&P-MIB è il MIDEX. Anche questo indice viene calcolato sulla base del flottante e viene aggiornato ogni minuto sulla base dei prezzi degli ultimi contratti conclusi su ciascuna azione componente. Il MIDEX è stato creato il 31 dicembre 1994 con valore pari a 10000 e al 31 dicembre 2008 il suo valore era pari a 17178 punti. A seguito della fusione con il LSE, questo indice è stato rinominato FTSE Italia Mid Cap.

Per quanto riguarda il segmento STAR, l'indice più rappresentativo è l'ALL-STARS introdotto il 31 dicembre 2001 (base = 10000) e sostituito dal FTSE Italia Star. Il 31 dicembre 2008 l'indice ALL-STARS valeva 8554 punti.

Infine, prima della fusione tra Borsa Italiana e il LSE, il segmento Standard non era rappresentato da alcun indice borsistico. Attualmente, l'indice di riferimento per questo segmento di mercato è il FTSE Italia Small Cap.

La classificazione settoriale GICS (Global Industry Classification Standard), adottata come uno dei principali criteri per la formazione del paniere dell'indice S&P-MIB, è stata sviluppata da Standard and Poor's e Morgan Stanley Capital International ed è suddivisa in 10 settori: Beni di Consumo Voluttuari, Beni di Consumo di Prima Necessità, Energia, Finanza, Sanità, Industriali, Informatica, Materiali, Servizi di Telecomunicazione e Pubblica Utilità. L'intento di questa classificazione è quello di "venire incontro alle esigenze degli investitori in merito a un sistema di classificazione che rifletta la performance finanziaria di una società" (Standard and Poor's, 2008).

4.2 I dati

Per analizzare il mercato azionario italiano, si considera la serie dei rendimenti mensili in percentuale di 208 titoli quotati nei principali segmenti del mercato MTA presentati nel paragrafo 4.1 per il periodo 2000-2008. Oltre ai titoli appartenenti al segmento Blue-Chip e classificati negli indici S&P-MIB e MIDEX, si valutano anche le azioni di società a media e bassa capitalizzazione classificate nell'indice ALL-STARS e nel segmento Standard. Per mantenere un certo grado di coerenza, le azioni con un numero di osservazioni inferiore a 24, che, nel caso di dati mensili, corrispondono alle società con quotazione successiva al primo febbraio 2007, vengono eliminate dalla base dei dati. In particolare, vengono escluse dall'analisi 11 azioni: una appartenente all'indice S&P-MIB, due al MIDEX e otto all'ALL-STARS. Per il segmento Standard vengono presi in considerazione un totale di 72 titoli. Nell'**appendice A** si fornisce la lista dei 208 titoli analizzati ed alcune statistiche di sintesi.

In tabella 4.1 vengono mostrate le statistiche descrittive per i tre indici azionari. In particolare, oltre al numero di titoli, la tabella 4.1 mostra la media, la deviazione standard, la mediana, il primo ed il quinto percentile della distribuzione dei rendimenti da gennaio 2000 a dicembre 2008 per gli indici azionari considerati.

Indice	Nr. Azioni Totali (Analizzate)	Media	Dev. Std.	Mediana	I Perc.	V Perc.
S&P-MIB	40 (39)	-0,423	5,699	0,620	-15,78	-11,92
MIDEX	31 (29)	-0,252	6,338	0,515	-16,54	-11,88
ALL-STARS	76 (68)	-0,069	4,837	0,691	-11,96	-9,60

Tabella 4.1: statistiche descrittive degli indici S&P-MIB, MIDEX e ALL-STARS

Dalla tabella 4.1 si nota come i rendimenti medi dei tre indici azionari nel periodo 2000-2008 siano negativi. Questo risultato, in antitesi con la teoria economica che assume un premio per il rischio positivo, è dovuto alla presenza di due importanti crisi finanziarie avvenute nell'orizzonte temporale analizzato. Come mostrato in figura 4.1, l'andamento dei rendimenti mensili dei tre indici azionari è sensibilmente diverso in alcuni periodi rispetto ad altri. In particolare, tra il 2001 ed il 2003 e dopo la seconda metà del 2007 si riscontra una maggiore instabilità nelle serie dei rendimenti degli indici. Si può notare infatti come, nei periodi citati, vi sia un aumento significativo della variabilità o, in termini finanziari, della volatilità, ovvero un sostanziale aumento nell'intensità delle fluttuazioni attorno all'asse delle ascisse con picchi positivi che si alternano a picchi fortemente negativi. Inoltre, a partire da metà 2007, i rendimenti dei tre indici azionari risultano negativi anche per diversi mesi consecutivamente. È noto infatti che nei periodi 2001-2003 e dalla fine del 2007 e per tutto il 2008, il mercato azionario italiano, come la maggior parte dei mercati mondiali, sia stato vittima di due acute crisi finanziarie.

La tabella 4.2 mostra la media, la deviazione standard ed il primo percentile della serie dei rendimenti mensili dell'indice S&P-MIB nei semestri dei nove anni analizzati. Dalla tabella 4.2, si nota come il mercato azionario italiano abbia subito una contrazione in corrispondenza degli anni 2001 e 2002 seguita da una serie di semestri con media positiva fino al secondo semestre 2007, in cui comincia una fase di forte declino. Oltre al rendimento medio, è possibile osservare anche le differenze accentuate nei valori delle deviazioni standard. Si nota infatti come in corrispondenza dei semestri nei quali avvengono le crisi finanziarie, i valori della deviazione standard risultano maggiori, ad indicare un incremento della volatilità.

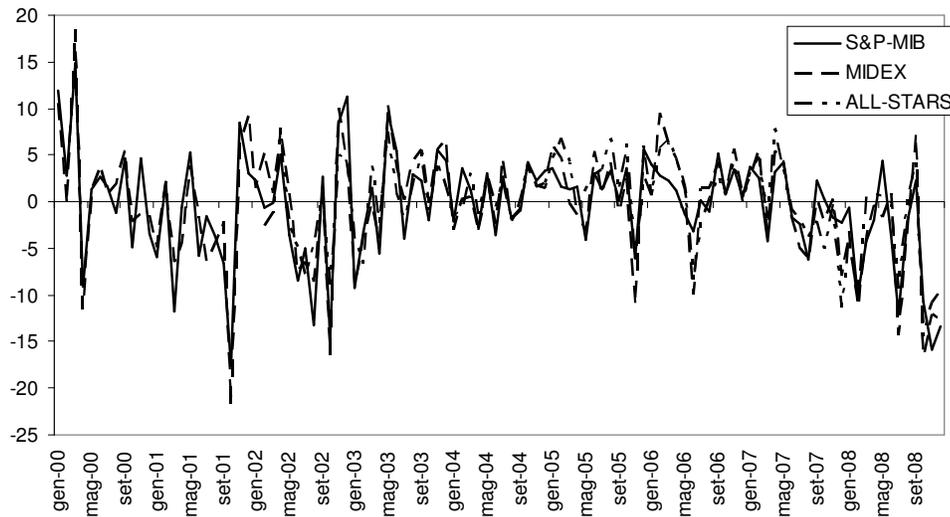


Figura 4.1: l'andamento dei rendimenti mensili degli indici S&P-MIB, MIBEX e ALL-STARS per il periodo 2000-2008

Semestre	Media	Dev. Std.	I Perc.
I-2000	4,068	7,775	-8,635
II-2000	0,222	3,689	-4,850
I-2001	-2,802	5,662	-11,467
II-2001	-2,990	8,161	-17,098
I-2002	-0,911	4,277	-8,189
II-2002	-1,766	10,017	-14,607
I-2003	-0,245	6,444	-9,025
II-2003	1,513	3,402	-3,872
I-2004	-0,110	2,837	-3,508
II-2004	1,605	2,215	-1,785
I-2005	1,219	2,517	-3,824
II-2005	1,316	3,742	-5,798
I-2006	0,975	2,457	-3,034
II-2006	1,495	2,103	-0,959
I-2007	1,287	3,118	-4,056
II-2007	-1,659	2,590	-5,970
I-2008	-2,711	4,377	-9,914
II-2008	-8,651	6,450	-15,735

Tabella 4.2: media, deviazione standard e primo percentile dei rendimenti dell'indice S&P-MIB nei semestri 2000-2008

Settori GICS	Nr. Azioni	Media	Dev. Std.
Beni Voluttuari	59	-0,840	11,310
Beni Prima Necessità	7	-0,548	8,016
Energia	8	0,590	9,667
Finanza	52	-0,006	10,315
Sanità	3	-0,304	10,252
Industriali	32	0,151	10,622
Informatica	18	-0,874	13,327
Materiali	12	-0,348	9,072
Telecomunicazione	2	-0,910	12,592
Pubblica Utilità	15	0,122	9,680

Tabella 4.3: statistiche descrittive della classificazione settoriale GICS

In tabella 4.3 vengono mostrate le statistiche descrittive dei settori della classificazione GICS ed il relativo numero di azioni appartenenti a ciascuno dei 10 settori. Inoltre, nella figura 4.2 vengono raffigurati i settori GICS nella classica rappresentazione deviazione standard - rendimento medio. Come si può osservare dalla tabella 4.3 e dalla figura 4.2, considerando la classificazione settoriale GICS può essere difficile differenziare in modo netto il profilo rischio-rendimento: i valori delle medie e delle deviazioni standard di alcuni settori risultano, infatti, piuttosto simili tra loro. Ad esempio, i punti riferiti a Beni Voluttuari, Informatica e Telecomunicazione giacciono tutti nella parte inferiore destra del grafico di figura 4.2 evidenziando le pessime caratteristiche finanziarie di questi tre settori: basso rendimento medio e alta volatilità; oppure i valori di media e deviazione standard in tabella 4.3 per Finanza, Industriali e Pubblica Utilità che risultano molto simili tra loro.

Uno degli scopi di questo lavoro è quello di proporre una alternativa alla classificazione settoriale GICS che risulti consistente sotto il profilo rischio-rendimento dei titoli azionari. Questa questione viene analizzata nei paragrafi **4.3**

e 4.4 ricorrendo ai metodi statistici per le variabili latenti che permettono di ottenere una corretta classificazione dei titoli finanziari descritti in questo paragrafo in gruppi omogenei ed una accurata misura e valutazione del loro profilo di rischio-rendimento atteso, consentendo, di conseguenza, un miglioramento delle procedure decisionali di investimento. Inoltre, sempre nel contesto della classificazione, ma in un ambito più aggregato, nel paragrafo 4.5, si analizzano i problemi di costruzione e rappresentatività dei numeri indici finanziari. Questa analisi permette di sviluppare alcune proposte innovative sulla costruzione dei numeri indici dei prezzi e consente di testare la robustezza della metodologia proposta. Infine, il paragrafo 4.6 si studia l'andamento dinamico dell'indice azionario S&P-MIB per valutare l'alternanza dei periodi di crisi e di stabilità finanziaria. I metodi a variabili latenti permettono infatti di includere esplicitamente nell'analisi la dimensione temporale consentendo il passaggio da una specificazione statica ad una dinamica.

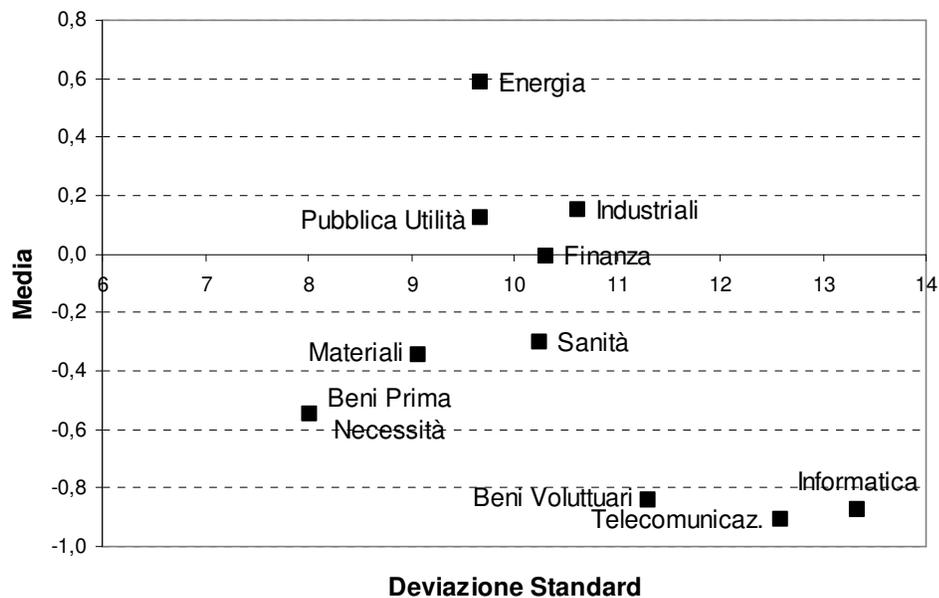


Figura 4.2: media e deviazione standard dei 10 settori della classificazione GICS

4.3 La classificazione dei titoli finanziari

La prima soluzione metodologica si basa sulla proposta descritta nel paragrafo **1.2** e permette la definizione di una nuova classificazione dei titoli finanziari. In particolare, viene sfruttata la capacità che i modelli a struttura latente hanno nell'allocare le unità statistiche in gruppi composti da elementi con caratteristiche comuni, seguendo un processo di classificazione non supervisionato. Nell'approccio esplorativo, il numero dei gruppi, non noto a priori, viene determinato sulla base di alcuni indicatori statistici. Nell'analisi a classi latenti, la procedura di validazione del processo di classificazione coincide con la verifica del buon adattamento del modello ai dati, ovvero con il controllo della validità dell'assunzione di indipendenza locale.

4.3.1 I modelli a struttura latente per la classificazione

Nei paragrafi **1.2** e **1.5** si è evidenziato come i modelli a struttura latente offrano una perfetta impostazione metodologica per ottenere una corretta classificazione degli oggetti. L'approccio probabilistico su cui si basa l'analisi a classi latenti permette di definire la nuova classificazione seguendo un processo metodologico rigoroso: l'allocatione delle osservazioni campionarie nelle K classi latenti avviene in base alle stime di massima verosimiglianza delle probabilità a posteriori.

Come presentato nel paragrafo **1**, la teoria classica del portafoglio valuta il profilo rischio-rendimento sulla base di due variabili latenti: il rischio V e il rendimento atteso $E(R)$. Queste due variabili possono essere approssimate per mezzo di tre variabili (continue) ottenute sulla base dei rendimenti passati osservabili dei titoli azionari: la media (\bar{R}), la deviazione standard (σ) ed il primo percentile (τ_1). Nel paragrafo **1.1**, si è argomentato come la caratteristica di tali variabili di essere semplici approssimazioni, probabilmente diverse rispetto ai

veri valori non osservabili del rendimento atteso e del rischio, viene spesso trascurata. Per valorizzare l'importanza di questo punto e con l'intento di ottenere una maggiore flessibilità nella classificazione dei titoli azionari, si è proposto di esprimere \bar{R} , σ e τ_1 come variabili categoriche ordinabili (M , S e P). In particolare, in questa analisi, si è scelto di riclassificare i valori puntuali delle tre variabili nella formulazione più semplice: ogni indicatore continuo viene trasformato in uno categorico con due categorie, dove le classi indicano i livelli basso e alto e vengono definiti sulla base del valore mediano.

Per ottenere la classificazione dei titoli azionari riferendosi alle variabili M , S e P (con indici m , s e p , rispettivamente), si ricorre al modello a classi latenti per variabili manifeste dicotomiche descritto nel paragrafo 2.2.2. Inoltre, si specifica questo modello includendo la classificazione settoriale GICS denotata con C (con indice c) come covariata. Come descritto nel paragrafo 4.1, la variabile categorica C esprime uno dei metodi di classificazione più usati nei mercati finanziari ma, come si può notare dall'analisi preliminare riportata in tabella 4.3 e dalla figura 4.2, il raggruppamento dei titoli azionari rispetto al settore economico GICS non permette una discriminazione soddisfacente del profilo di rischio-rendimento dei gruppi.

Riprendendo la specificazione introdotta nel paragrafo 2.3, il modello a classi latenti per le variabili dicotomiche oggetto di analisi è

$$\pi_{mspc}^{MSPC} = \sum_{x=1}^K \pi_{xmspc}^{XMSPC} \quad (4.1)$$

e

$$\pi_{xmspc}^{XMSPC} = \pi_c^C \pi_{xlc}^{X1C} \pi_{msplx}^{MSP1X} = \pi_c^C \pi_{xlc}^{X1C} \pi_{mlx}^{M1X} \pi_{slx}^{S1X} \pi_{plx}^{P1X} . \quad (4.2)$$

In questo modello, la variabile (di intervento) latente X permette di determinare la classificazione dei titoli finanziari ricercata.

In figura 4.3 viene rappresentato il modello a classi latenti specificato nelle equazioni (4.1) e (4.2). Dalla figura si può notare come la classificazione settoriale

tradizionale GICS influenzi direttamente la nuova classificazione X ma non abbia nessuna relazione con i tre indicatori.

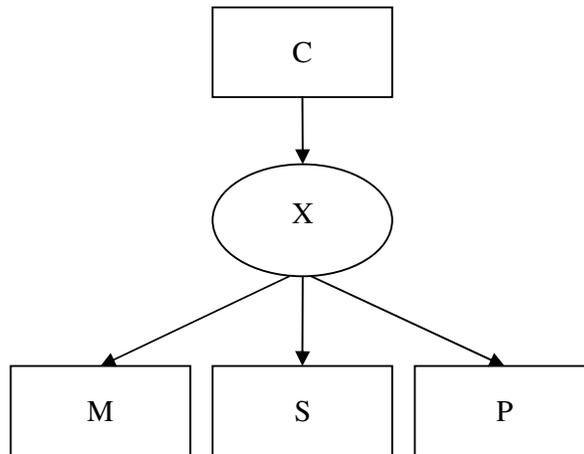


Figura 4.3: il modello a classi latenti per la classificazione dei titoli azionari

Una seconda proposta per la classificazione dei titoli azionari che valuta direttamente i valori puntuali delle variabili manifeste continue, si basa sull'analisi a profilo latente descritta nel paragrafo **2.4**. A differenza del modello a classi latenti proposto nelle equazioni (4.1) e (4.2), l'analisi a profilo latente viene eseguita includendo nel modello quattro variabili manifeste continue e due covariate categoriche. Oltre al rendimento medio \bar{R} usato come approssimazione del rendimento atteso, allo scarto quadratico medio σ ed al primo percentile τ_1 della distribuzione dei rendimenti dei titoli azionari, un ulteriore indicatore riferito al rischio è dato dalla deviazione standard del titolo durante i periodi di crisi σ_C . L'inclusione di questa variabile manifesta nel modello permette la valutazione degli effetti che i periodi di turbolenza finanziaria hanno sul comportamento delle azioni ed, in particolare, sulla loro volatilità. Oltre alle quattro variabili manifeste, si includono nel modello anche le due covariate, settore economico GICS (C) e indice di mercato (I).

Come nel modello a classi latenti, la variabile latente X definita da K classi è interpretabile come la nuova classificazione dei titoli azionari in gruppi omogenei rispetto alle caratteristiche che definiscono il loro profilo di rischio-rendimento⁷.

Analogamente a quanto presentato nel paragrafo 2.4, la specificazione del modello a profilo latente risulta essere

$$f(Z) = \sum_{x=1}^K \pi_{x,c,i} g(\bar{R} | x) g(\sigma | x) g(\tau_1 | x) g(\sigma_c | x) \quad (4.3)$$

dove le probabilità condizionate $g(.|x)$ sono assunte come normalmente distribuite.

In figura 4.4 si fornisce la rappresentazione grafica del modello a profilo latente specificato.

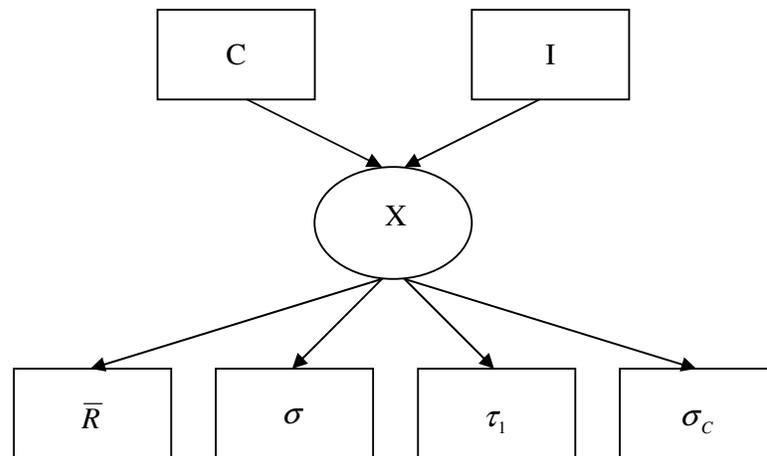


Figura 4.4: il modello a profilo latente per la classificazione dei titoli azionari

In entrambi i modelli a struttura latente considerati in questo paragrafo, la determinazione del numero di classi latenti K è un passo fondamentale

⁷ L'analisi del profilo rischio-rendimento dei gruppi ottenuti dalle analisi a classi e a profilo latente viene affrontata nel paragrafo 4.4.

nell'analisi, in quanto rappresenta il numero dei gruppi che costituiscono la nuova classificazione. Il ricorso ad una o più statistiche test per la scelta di K implica l'introduzione di un rigoroso procedimento metodologico nei processi di classificazione azionaria. Nella realtà dei mercati finanziari, i processi di classificazione si basano infatti su decisioni piuttosto arbitrarie, non sempre facilmente valutabili e senza il coinvolgimento di un approccio di tipo scientifico.

Dopo aver ottenuto la stima dei parametri dei modelli, l'ultima fase delle analisi consiste nell'assegnare le unità osservate alla classe latente più appropriata. La classificazione dei titoli nei K gruppi avviene riferendosi alle probabilità a posteriori in equazione (2.6) per il modello a classi latenti o, analogamente,

$$\hat{\pi}_{xmspc} = \frac{\hat{\pi}_{xmspc}}{\sum_{x=1}^K \hat{\pi}_{xmspc}} \quad (4.4)$$

e, per il modello a profilo latente, all'analisi a posteriori descritta nel paragrafo **2.4.3**.

4.3.2 Le implicazioni dell'analisi: il confronto tra la nuova classificazione e quella settoriale tradizionale

La classificazione settoriale ricopre un ruolo fondamentale nei mercati finanziari in quanto rappresenta uno dei riferimenti principali nei processi di diversificazione del portafoglio.

La classificazione ottenuta sfruttando le potenzialità dei metodi a variabili latenti permette di valutare i K gruppi di titoli individuati rispetto alle loro caratteristiche del profilo rischio-rendimento. L'ipotesi alla base di questo approccio definisce le azioni appartenenti allo stesso settore come omogenee o, quantomeno, presuppone che i settori caratterizzano ed influenzano l'andamento di un titolo in modo rilevante. In questo contesto, settori diversi dovrebbero essere caratterizzati da livelli di rischio e rendimento diversi tra loro. Inoltre, i settori dovrebbero essere influenzati dal ciclo economico, distinguendo tra settori pro-

ciclici e anti-ciclici. Per ottenere questi obiettivi è essenziale che l'assegnazione di una singola azione ad uno specifico settore avvenga seguendo un processo metodologico corretto e rigoroso.

La classificazione settoriale tradizionale mostra alcuni difetti ai quali è necessario porre rimedio proponendo delle soluzioni efficaci. In primo luogo, la classificazione settoriale risulta fortemente statica, in quanto viene raramente aggiornata dal momento in cui una società viene quotata sul mercato azionario. In secondo luogo, le società azionarie, di solito, operano in diversi settori mentre la classificazione tradizionale considera solo l'attività principale. Infine, il settore basato sulla categoria merceologica può non rappresentare il criterio classificatorio più appropriato per discriminare correttamente il profilo di rischio-rendimento.

In questa analisi, la classificazione settoriale tradizionale viene confrontata con la nuova classificazione basata sul profilo rischio-rendimento non osservabile e, per questo motivo, ottenuta nel campo delle classi latenti.

4.3.3 I dati per le analisi

Per l'analisi a classi latenti, si considerano 5 settori della classificazione GICS: Beni Voluttuari, Energia, Finanza, Materiali e Pubblica Utilità. Oltre alla loro importanza in termini di capitalizzazione sul mercato azionario italiano, i settori considerati risultano essere piuttosto eterogenei rispetto al loro profilo rischio-rendimento: nella figura 4.2, i punti nello spazio deviazione standard-rendimento medio riferiti a questi settori appaiono ben distanziati gli uni dagli altri. Tuttavia, dalla rappresentazione delle 146 azioni oggetto di analisi in figura 4.5, si nota come, al loro interno, i settori tradizionali presentino una certa eterogeneità se valutati sulla base del profilo rischio-rendimento. In particolare, si può osservare come praticamente tutti i 5 settori GICS siano caratterizzati da una notevole dispersione lungo almeno uno degli assi cartesiani di figura 4.5.

In tabella 4.4 sono riportate le statistiche descrittive relative ai 5 settori GICS. Anche i valori sintetici in tabella 4.4 confermano la scarsa omogeneità dei titoli azionari appartenenti ai 5 settori GICS analizzati.

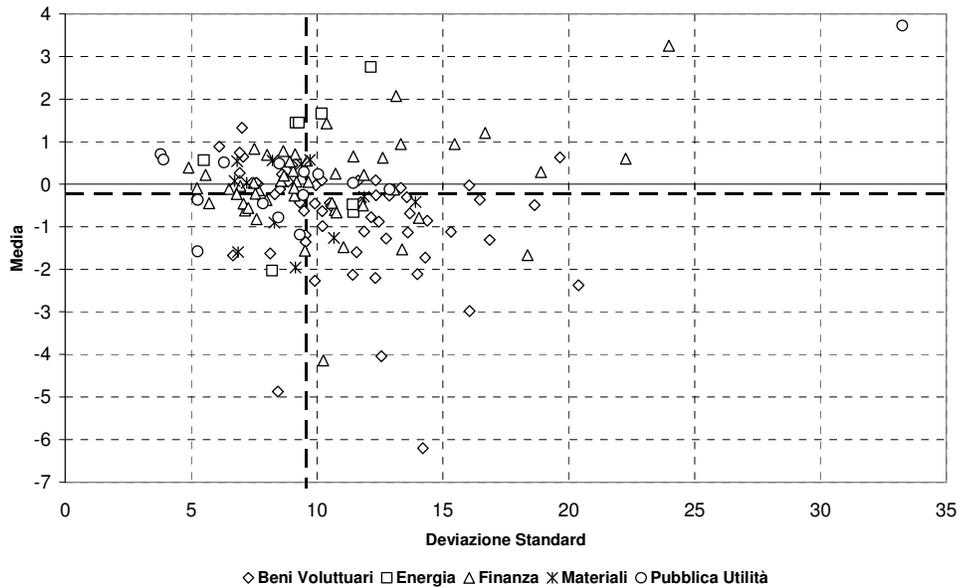


Figura 4.5 : i 146 titoli azionari appartenenti ai 5 settori GICS

Settore GICS	Nr.	Rendimento	Deviazione	Min / Max	Min / Max
	Azioni	Medio	Standard	Rendimento	Dev. Std.
Beni Voluttuari	59	-0,840	11,310	-6,2 / 1,32	6,13 / 20,39
Energia	8	0,590	9,667	-2,03 / 2,75	5,50 / 12,14
Finanza	52	-0,006	10,315	-4,14 / 3,26	4,90 / 23,99
Materiali	12	-0,348	9,072	-1,96 / 0,57	6,73 / 13,92
Pubblica Utilità	15	0,122	9,680	-1,58 / 3,73	3,79 / 33,26

Tabella 4.4: statistiche descrittive dei 5 settori GICS considerati nell'analisi

Seguendo la proposta del paragrafo 1.1, i valori puntuali delle variabili dei 146 titoli azionari vengono trasformati in variabili categoriche dicotomiche, valutando come valore soglia la mediana che identifica le due categorie:

- Indicatore media (M): basso = 1, alto = 2 (mediana = -0,232);
- Indicatore deviazione standard (S): alto = 1, basso = 2 (mediana = 9,575);
- Indicatore primo percentile (P): basso = 1, alto = 2 (mediana = -22,08).

In figura 4.5, le linee tratteggiate indicano i valori soglia per le variabili dicotomiche M ed S . La figura 4.6 mostra le frequenze assolute delle categorie per le tre variabili manifeste riferite ai cinque settori GICS. Anche dagli istogrammi in figura 4.6 emergono dei profili di rischio-rendimento dei settori tradizionali non ben delineati, ad eccezione dei Beni Voluttuari e della Pubblica Utilità che presentano una proporzione maggiore di valori bassi e alti, rispettivamente. Il test chi-quadro per l'indipendenza risulta significativo per un livello pari a $\alpha = 0,05$ ma non per $\alpha = 0,01$ ($p\text{-value} = 0,029$).

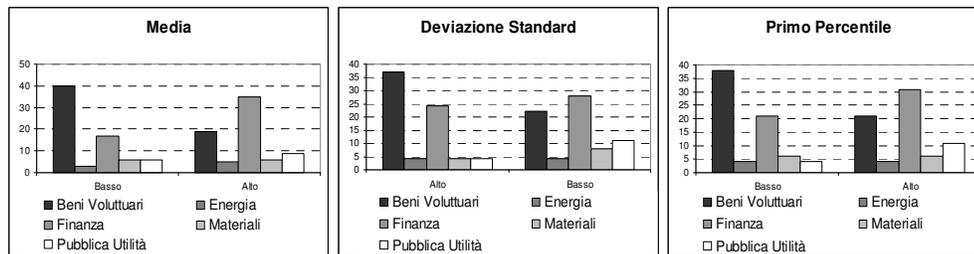


Figura 4.6: le categorie dei cinque settori GICS per le variabili manifeste

Per l'analisi a profilo latente, si valutano i rendimenti mensili di tutti i 208 titoli azionari presentati nel paragrafo 4.2, per i quali si sono rilevate le seguenti variabili manifeste continue come indicatori:

- la media (\bar{R});
- la deviazione standard (σ);
- il primo percentile (τ_1);
- la deviazione standard nei periodi di crisi (σ_c);

e due variabili categoriche come covariate:

- il settore economico GICS (C);
- l'indice di mercato (I).

La variabile σ_C indica la deviazione standard dei rendimenti giornalieri dei titoli azionari nei periodi associati a ben note crisi finanziarie e/o a forti cali del mercato azionario. I periodi considerati sono: settembre-ottobre 2001, luglio-ottobre 2002 e settembre-dicembre 2008. L'inclusione delle informazioni fornite dalle due covariate permette di aumentare l'attendibilità delle stime e, come viene mostrato nel seguito, la loro utilità sia a livello strategico che operativo. In particolare, si fa riferimento ai due metodi di classificazione più comuni nei mercati azionari. Da un lato, la covariata C consiste nei dieci settori della classificazione GICS e, come descritto nel paragrafo 4.3.2, rappresenta uno dei maggiori riferimenti nei processi di diversificazione del portafoglio. Dall'altro, la covariata I che comprende gli indici S&P-Mib, Midex, All-Stars e Standard denota la principale segmentazione del mercato azionario italiano ed è quindi collegata alla rilevanza del titolo in termini di capitalizzazione⁸.

4.3.4 La stima dei modelli

La stima del modello a classi latenti specificato nelle equazioni (4.1) e (4.2) e di quello a profilo latente mostrato nell'equazione (4.3) viene effettuata per diversi valori di K in modo da determinare il minor numero di classi latenti in grado di spiegare le relazioni osservate tra le variabili manifeste.

Solitamente, l'analisi comincia con la stima del modello base con una sola classe latente che corrisponde all'assumere l'indipendenza tra le variabili osservate. Se il modello base fornisce un adattamento adeguato ai dati allora non

⁸ Come segnalato nel paragrafo 4.1, recentemente gli indici azionari del mercato italiano elencati in questa analisi sono stati sostituiti da altri indici realizzati seguendo gli standard del FTSE Group di Londra. Tuttavia, i nuovi indici risultano molto simili e, il più delle volte identici, a quelli considerati soprattutto per quanto riguarda la metodologia di calcolo ed il numero delle azioni appartenenti ad ogni indice.

c'è bisogno di effettuare alcuna analisi, in quanto non esistono relazioni tra le variabili manifeste che possono essere spiegate attraverso la variabile latente e tutte le osservazioni appartengono alla stessa popolazione e, quindi, alla medesima classe.

I risultati del modello a classi latenti per diversi valori di K sono riportati nella tabella 4.5. Il valore significativamente diverso da zero della statistica L^2 indica che il modello con una classe latente debba essere rifiutato e, per questo motivo, le associazioni esistenti nei dati non possono essere spiegate senza il ricorso ad una variabile latente con almeno 2 classi.

Modello	LL	$NPar$	L^2	gdl	$p-value$	$AIC(LL)$
1 classe	-303,585	3	140,61	32	1,4E-15	613,17
2 classi	-252,244	11	37,93	24	0,035	526,49
3 classi	-241,309	19	16,06	16	0,45	520,62
4 classi	-238,982	27	11,41	8	0,18	531,96

Tabella 4.5: risultati dei modelli a classi latenti con un diverso numero di classi

Il modello con 2 classi latenti fornisce una forte riduzione del valore di L^2 (73% rispetto al modello base), tuttavia, questa statistica risulta ancora troppo elevata ($L^2 = 37,93$; $gdl = 24$; $p-value < 0,05$). Aggiungendo una terza classe al modello si ottiene una ulteriore riduzione di L^2 (89% rispetto al modello base) ed un adattamento ai dati adeguato. La tabella 4.5 mostra che, in base alla statistica AIC che prende in considerazione anche la parsimonia, il modello con 3 classi latenti è preferibile rispetto al modello con 4 classi.

Un altro metodo di diagnostica utile per valutare l'adattamento ai dati del modello è definito sulla base dei residui bivariati (BVR; Magidson e Vermunt, 2001). In questo test basato sul chi-quadrato di Pearson, la relazione tra due variabili viene analizzata confrontando le frequenze osservate con quelle teoriche

stimate dal modello. Un valore del BVR sostanzialmente maggiore di uno suggerisce che il modello a classi latenti non è in grado di spiegare adeguatamente l'associazione esistente tra quelle due variabili.

In tabella 4.6 vengono riportati i valori del BVR per ogni coppia di variabili riferiti ai vari modelli a classi latenti stimati. La tabella 4.6 mostra come i modelli che presentano tutti i valori di BVR inferiori ad 1 siano quelli con 3 e 4 classi latenti, confermando il fatto che $K = 3$ può considerarsi come la miglior soluzione in grado di spiegare adeguatamente le associazioni esistenti tra le variabili manifeste. La scelta del numero di classi sulla base di indicatori statistici rappresenta un passo fondamentale dell'analisi, in quanto permette di definire il numero dei gruppi in cui vengono classificati i titoli azionari.

Coppie di variabili	1 classe	2 classi	3 classi	4 classi
[MS]	6,967	0,249	0,002	0,000
[MP]	10,886	0,013	0,007	0,031
[SP]	70,757	0,033	0,011	0,002
[MC]	3,671	1,920	0,403	0,274
[SC]	2,167	0,287	0,053	0,059
[PC]	2,506	0,090	0,156	0,047

Tabella 4.6: valori di residui bivariati (BVR) relativi ai vari modelli a classi latenti

La tabella 4.7 illustra i risultati della stima di massima verosimiglianza per il modello con tre classi latenti. Le stime dei parametri mostrano che due classi latenti hanno probabilità (a priori) piuttosto simili mentre la terza ha una dimensione inferiore: quasi il 45% delle azioni sono stimate essere nella Classe 1 ($\pi_{x=1} = 0,448$), il 40% nella Classe 2 ($\pi_{x=2} = 0,402$) ed il restante 15% nella Classe 3 ($\pi_{x=3} = 0,150$).

Variabile manifesta		Classe 1	Classe 2	Classe 3
	Categorie	0,4478	0,4022	0,1500
<i>M</i>	$\pi_{m=bassolx}$	0,3605	0,8177	0,0211
	$\pi_{m=altolx}$	0,6395	0,1823	0,9789
	\bar{z}_{mlx}	1,6395	1,1823	1,9789
<i>S</i>	$\pi_{s=altolx}$	0,0061	0,9063	0,8875
	$\pi_{s=bassolx}$	0,9939	0,0937	0,1125
	\bar{z}_{slx}	1,9939	1,0937	1,1125
<i>P</i>	$\pi_{p=bassolx}$	0,0732	0,9200	0,6503
	$\pi_{p=altolx}$	0,9268	0,0800	0,3497
	\bar{z}_{plx}	1,9268	1,0800	1,3497

Tabella 4.7: risultati del modello a 3 classi latenti, probabilità condizionate e medie delle variabili manifeste dicotomiche

La stima del modello a profilo latente per diversi valori di K permette la definizione del numero di classi più adatta a spiegare le relazioni esistenti tra le variabili manifeste continue. In tabella 4.8 vengono illustrati i valori della log-verosimiglianza, il relativo numero di parametri, la statistica AIC ed il test del rapporto di verosimiglianza ottenuto tramite il metodo *bootstrap* e condotto confrontando le log-verosimiglianze dei modelli con K e $K - 1$ classi. Quando il test risulta significativo, l'aggiunta di una ulteriore classe latente fornisce un miglior adattamento del modello ai dati.

Dalla tabella 4.8 è possibile determinare il modello che presenta il migliore adattamento. In primo luogo, si può notare come la funzione di log-verosimiglianza raggiunga il massimo in corrispondenza del modello con 9 classi latenti. Come secondo riferimento, anche il criterio AIC suggerisce il modello a profilo latente con 9 classi come soluzione più adeguata. Infine, una ulteriore conferma emerge dal test del rapporto di verosimiglianza ottenuto tramite il

metodo bootstrap (*Bootstrap -2LL Diff.*). I valori del test, infatti, risultano tutti significativamente diversi da zero, tranne per il confronto tra il modello con 9 e 10 classi latenti, evidenziando, perciò, la presenza di 9 sottopopolazioni diverse in cui i 208 titoli azionari oggetto di analisi sono suddivisi.

Modello	<i>LL</i>	<i>NPar</i>	<i>AIC(LL)</i>	<i>Bootstrap -2LL Diff.</i>	<i>p-value</i>
1 classe	-1895,49	8	3806,98	-	-
2 classi	-1712,91	29	3483,83	365,16	0,000
3 classi	-1659,09	50	3418,17	107,65	0,000
4 classi	-1629,33	71	3400,66	59,51	0,004
5 classi	-1601,11	92	3386,22	56,44	0,002
6 classi	-1570,73	113	3367,46	60,76	0,000
7 classi	-1553,58	134	3375,17	45,89	0,038
8 classi	-1528,40	155	3366,80	50,37	0,004
9 classi	-1503,00	176	3358,00	50,79	0,000
10 classi	-1493,99	197	3381,99	23,36	0,119

Tabella 4.8: risultati dei modelli a profilo latente con un diverso numero di classi

La tabella 4.9 mostra i risultati della stima del modello a profilo latente, illustrando le probabilità a priori e le medie condizionate dei 4 indicatori. Le nove classi latenti sono numerate rispetto alla loro dimensione, ovvero rispetto alle probabilità a priori $\pi_{x|c,i}$ illustrate nella prima riga di tabella 4.9. La Classe 1 rappresenta il gruppo modale e comprende il 18,8% delle azioni, mentre la Classe 9 è la più piccola e raggruppa solo il 3,8% dei titoli azionari. Dalle probabilità a priori si può osservare la presenza di alcuni gruppi piuttosto piccoli, ad esempio le Classi 7, 8 e 9, ed altri alquanto importanti come le Classi 1, 2 e 3 che, se cumulate, raggruppano il 50% dei titoli azionari.

In tabella 4.9, le nove classi latenti sono state ordinate rispetto al valor medio condizionato dell'indicatore \bar{R} ($\hat{\mu}(\bar{R} | x)$).

	Cl 7	Cl 4	Cl 3	Cl 2	Cl 1	Cl 8	Cl 5	Cl 6	Cl 9
$\pi_{x c,i}$	0,065	0,138	0,146	0,166	0,188	0,043	0,112	0,105	0,038
$\hat{\mu}(\bar{R} x)$	-3,17	-0,96	-0,30	-0,24	-0,13	0,01	0,28	0,34	1,71
$\hat{\mu}(\sigma x)$	12,70	15,17	11,44	7,49	9,17	14,92	10,54	5,74	19,37
$\hat{\mu}(\tau_1 x)$	-30,7	-32,5	-25,9	-18,1	-21,5	-35,6	-21,9	-13,4	-26,1
$\hat{\mu}(\sigma_c x)$	3,66	4,69	3,49	2,83	3,18	3,76	3,50	2,67	3,82

Tabella 4.9: risultati del modello a profilo latente con 9 classi, probabilità a priori e medie delle variabili manifeste

4.3.5 La nuova classificazione dei titoli azionari

Il passo successivo alla stima dei modelli a struttura latente consiste nel classificare ogni titolo nella classe più appropriata.

La stima del modello a classi latenti permette di assegnare ogni azione ad una delle tre classi in base alle probabilità a posteriori definite nell'equazione (4.4), ottenendo così la nuova classificazione. I gruppi sono costituiti rispettivamente da 67, 57 e 22 titoli azionari. In **appendice B** vengono riportate le probabilità a posteriori per le tre classi latenti (colonne "CL1", "CL2" e "CL3"), mentre la colonna "CL#" indica la classe modale in cui viene classificato il titolo azionario. Osservando le probabilità a posteriori riportate in **appendice B** si nota come i titoli che l'analisi a classi latenti non è in grado di assegnare ad un gruppo con una probabilità maggiore di 0,75 siano solo 5; inoltre, l'errore di classificazione relativo a questo metodo risulta essere piuttosto basso (pari a 0,051).

La tabella 4.10 mostra il peso che ogni settore tradizionale ha nella composizione delle classi latenti e, quindi, nella nuova classificazione. Dalla

tabella emergono alcune indicazioni interessanti. In primo luogo, vi sono alcuni settori che vengono classificati in buona parte in una classe latente: ad esempio, il 69,4% delle azioni dei Beni Voluttuari viene allocato nella Classe 2 ed il 72,2% dei titoli della Pubblica Utilità sono assegnati alla Classe 1. Inoltre, vi sono casi in cui il settore tradizionale non contribuisce affatto alla definizione della classe latente: praticamente nessun titolo appartenente ai settori dei Beni Voluttuari e dei Materiali viene classificato nella Classe 3 ed alcun titolo della Pubblica Utilità è assegnato alla seconda classe latente. In tutti gli altri casi, i settori tradizionali contribuiscono in modo abbastanza simile alla definizione di tutti i nuovi gruppi.

Settore Tradizionale	Classe 1	Classe 2	Classe 3
Beni Voluttuari	0,3023	0,6946	0,0031
Energia	0,4165	0,2920	0,2914
Finanza	0,5056	0,2072	0,2871
Materiali	0,5900	0,3833	0,0266
Pubblica Utilità	0,7219	0,0036	0,2745

Tabella 4.10: allocazione dei settori tradizionali nelle classi latenti

	Cl 7	Cl 4	Cl 3	Cl 2	Cl 1	Cl 8	Cl 5	Cl 6	Cl 9
$\pi_{xlc,i}$	0,065	0,138	0,146	0,166	0,188	0,043	0,112	0,105	0,038
Azioni	13	29	29	33	41	9	23	22	8

Tabella 4.11: la classificazione dei titoli azionari ottenuta dall'analisi a profilo latente

La classificazione dei titoli azionari nei nove gruppi che scaturiscono dall'analisi a profilo latente viene riportata in tabella 4.11 e, nel dettaglio, in **appendice C**.

4.4 La misura e la valutazione del profilo rischio-rendimento dei gruppi

Una volta stimati i modelli a struttura latente, è possibile estrapolare le caratteristiche di ogni classe osservando i valori delle probabilità condizionate ed il valore medio degli indicatori per ogni classe latente come proposto in equazione (1.4).

Inoltre, allo scopo di confrontare la metodologia a classi latenti con approcci di clustering più tradizionali, ci si riferisce alla tecnica delle k-medie che, come mostrato nel paragrafo 1.2, può essere ottenuta come caso particolare del modello a classi latenti (Magidson e Vermunt, 2002).

4.4.1 Il profilo rischio-rendimento nell'analisi a classi latenti

Nell'analisi a classi latenti stimata nel paragrafo 4.3.4, il profilo di rischio-rendimento delle tre classi può essere determinato sulla base delle probabilità condizionate $\hat{\pi}_{i|x}$ e delle medie delle variabili manifeste condizionate alla variabile latente di appartenenza $\bar{z}_{i|x}$ (per $i = m, s, p$) riportate in tabella 4.7. Nello specifico, la caratteristica principale della prima classe latente è il basso rischio: la Classe 1 ha i valori più alti per le variabili dicotomiche S e P (la loro media è $\bar{z}_{s|x=1} = 1,99$ e $\bar{z}_{p|x=1} = 1,93$) ed una media di M pari a 1,64. La seconda classe è caratterizzata da un basso rischio e da un basso rendimento: la Classe 2 mostra il più basso valor medio di S e P ($\bar{z}_{s|x=2} = 1,09$ e $\bar{z}_{p|x=2} = 1,04$) e la media più bassa per la variabile M ($\bar{z}_{m|x=2} = 1,18$). Infine, la terza classe latente presenta la media della variabile M più alta ($\bar{z}_{m|x=3} = 1,98$) e valori abbastanza bassi per le medie degli indicatori riferiti al rischio: $\bar{z}_{s|x=3} = 1,11$ e $\bar{z}_{p|x=3} = 1,35$.

Le probabilità condizionate $\pi_{m|x}$, $\pi_{s|x}$ e $\pi_{p|x}$ in tabella 4.7 evidenziano che le caratteristiche delle tre classi latenti sono ben definite in riferimento al profilo

rischio-rendimento azionario: la Classe 1 raggruppa i titoli che permettono le migliori opportunità di investimento coniugando un buon rendimento atteso ad un rischio piuttosto contenuto, la Classe 3 è caratterizzata da un alto rendimento atteso ma anche da un rischio elevato, infine, la Classe 2 è definita da un pessimo profilo di rischio-rendimento, in quanto risulta essere la classe con il livello di rischio più elevato ed il rendimento atteso inferiore. In questo gruppo vengono classificati i titoli azionari che andrebbero evitati.

4.4.2 Il confronto tra l'analisi a classi latenti e la tecnica delle k-medie

Per approfondire i vantaggi che il ricorso all'analisi a classi latenti comporta rispetto alla classificazione delle unità statistiche in gruppi omogenei rispetto al profilo di rischio-rendimento, si valuta un metodo di clustering più "tradizionale" come l'algoritmo delle k-medie (MacQueen, 1967), che rappresenta uno dei riferimenti più diffusi per definire la suddivisione di gruppi di oggetti in K partizioni sulla base dei loro attributi.

Il metodo delle k-medie viene eseguito usando i valori continui dei tre indicatori Z_i e consiste nel minimizzare iterativamente una certa misura di dissimilarità, in questo caso, la distanza euclidea quadratica, così da pervenire alla definizione di 3 gruppi. Il numero dei gruppi viene scelto in maniera esogena per permettere un confronto diretto con la classificazione ottenuta dalla stima del modello a classi latenti con 3 classi.

In tabella 4.12 vengono riportati i risultati del metodo delle k-medie applicato ai tre indicatori continui \bar{R} , σ e τ_1 delle 146 azioni analizzate. I titoli azionari vengono classificati in tre cluster abbastanza eterogenei: il Cluster 1 include 88 azioni ed è caratterizzato da una media uguale a -0,07 e dai valori inferiori di deviazione standard e primo percentile (8,42 e -18,68, rispettivamente), il Cluster 2 è composto da 54 titoli e assume il valore più basso per la media (-0,88), il valore intermedio di deviazione standard (12,92) e il primo percentile più basso (-

30,17), mentre il Cluster 3, contenente 4 azioni soltanto, è caratterizzato dalla media più alta (1,57), dal valore maggiore di deviazione standard (24,09) e da un primo percentile di -24,42.

Cluster	\bar{R}	σ	τ_1	Azioni	Classe1	Classe2	Classe3
1	-0,073	8,421	-18,679	88	65	14	9
2	-0,877	12,920	-30,169	54	2	42	10
3	1,568	24,093	-24,416	4	0	1	3

Tabella 4.12: risultati del metodo delle k-medie

La classificazione dei titoli azionari ottenuta tramite il metodo delle k-medie mostra alcune differenze rispetto alla allocazione risultante dalla stima del modello con 3 classi latenti.

Le ultime tre colonne di tabella 4.12 mostrano come le azioni assegnate ad ogni cluster vengano classificate nelle tre classi latenti. I Cluster 1 e 2 sono composti da titoli azionari allocati in tutte e tre le classi. In particolare, il Cluster 1 è costituito da un gran numero di azioni classificate nella Classe 1 mentre nel secondo cluster sono allocati in gran parte i titoli della Classe 2. Infine, il Cluster 3 contiene tre azioni originariamente assegnate alla terza classe latente ed uno proveniente dalla Classe 2.

Analizzando le intersezioni e le differenze tra i cluster delle k-medie rappresentati in figura 4.7 e le classi latenti in figura 4.8 emerge come il modello a classi latenti sia in grado di definire dei gruppi più omogenei rispetto al profilo di rischio-rendimento. Questa superiorità della metodologia a classi latenti, già evidenziata da Magidson e Vermunt (2002a, 2002b) e ripresa nel paragrafo **1.2**, si riscontra anche osservando la maggiore sovrapposizione dei punti dei tre cluster in figura 4.7 rispetto a quelli rappresentati in figura 4.8 che fanno, invece, riferimento alle tre classi latenti.

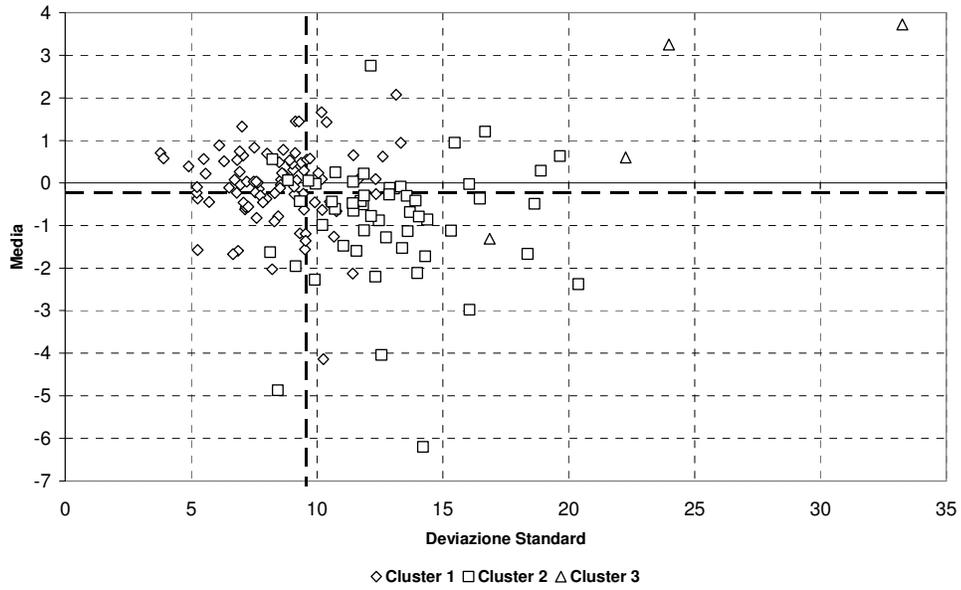


Figura 4.7: risultati del metodo delle k-medie: media e deviazione standard dei rendimenti azionari nei tre cluster

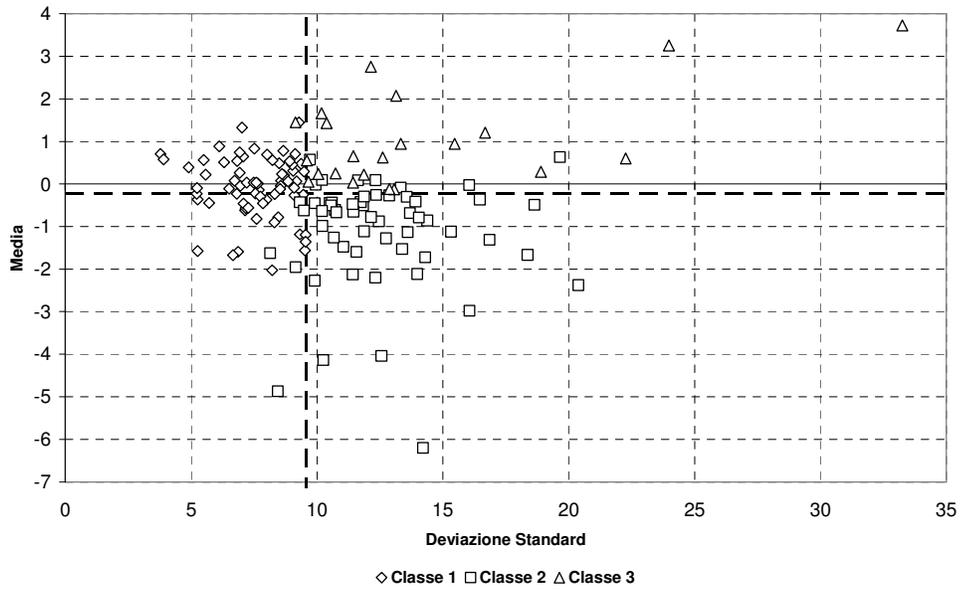


Figura 4.8: risultati del modello a classi latenti: media e deviazione standard dei rendimenti azionari nelle tre classi

Le indicazioni dell'analisi grafica possono, inoltre, trovare una conferma nella valutazione e nel confronto della variabilità relativa alle 3 classi latenti e ai 3 cluster. In particolare, per discriminare le capacità classificatorie dei due metodi, è possibile considerare le varianze entro i gruppi dei tre indicatori Z_i sia per il modello a classi latenti, sia per la tecnica delle k-medie:

$$\text{Var}_e(Z_i | x) = \frac{\sum_{h=1}^{n_x} (z_{ih} - \bar{z}_{i|x})^2}{n_x}$$

dove n_x denota il numero dei titoli classificati nel gruppo (classe latente o cluster) x .

Per il calcolo delle varianze entro si fa riferimento ai valori originali (continui) delle tre variabili manifeste M , S e P , sulla base dei quali sono ottenuti i valori medi per le tre classi latenti, riportati in tabella 4.13.

Classe	\bar{R}	σ	τ_1	Azioni
1	-0,077	7,585	-18,036	67
2	-1,141	12,537	-28,617	57
3	1,031	14,193	-24,136	22

Tabella 4.13: medie condizionate per le tre classi latenti riferite ai valori originali delle tre variabili manifeste

Le caratteristiche delle tre classi latenti, illustrate precedentemente in tabella 4.7 in riferimento alle variabili manifeste trasformate, vengono, così, riprese in tabella 4.13 sulla base dei valori continui di M , S e P . La tabella 4.13 viene introdotta allo scopo di consentire un più immediato e corretto confronto con le caratteristiche dei tre cluster, illustrate in tabella 4.12.

Dall'analisi delle tabella 4.12 e 4.13 è possibile evidenziare alcuni risultati interessanti. In particolare si sottolinea come, il Cluster 1 e la Classe 1, che concentrano il maggior numero di titoli, presentino un valore medio

dell'indicatore \bar{R} quasi identico, mentre il livello di rischio e, in particolare, il valore di $\bar{z}_{\sigma_{lx=1}}$, risulta inferiore con l'analisi a classi latenti. Emerge, così, uno dei punti di forza dell'analisi a classi latenti nel campo di variabili finanziarie: si tratta della capacità di definire un gruppo di azioni caratterizzate da un livello di rischio estremamente limitato. Tale aspetto viene ripreso ed illustrato nel dettaglio nel prossimo paragrafo **4.5**.

Il confronto tra i valori medi introduce quello tra le varianze entro, che sono riportate, per i tre indicatori \bar{R} , σ e τ_1 , rispettivamente nelle tabelle 4.14, 4.15 e 4.16.

x	Modello LC	K-medie
1	0,515	0,829
2	1,624	2,009
3	1,130	4,183

Tabella 4.14: varianze entro riferite all'indicatore \bar{R} per il modello a classi latenti e la tecnica delle k-medie

L'analisi della variabilità dentro ai gruppi indica, per l'indicatore \bar{R} , come le classi latenti consentano di ottenere, per tutti i casi considerati, una dispersione minore.

x	Modello LC	K-medie
1	2,013	3,982
2	7,362	8,298
3	32,426	34,899

Tabella 4.15: varianze entro riferite all'indicatore σ per il modello a classi latenti e la tecnica delle k-medie

Anche facendo riferimento all'indicatore σ si ottiene una situazione analoga: dalla tabella 4.15 è ancora possibile notare come la variabilità all'interno delle tre classi latenti sia sempre inferiore rispetto a quella presente nei tre cluster.

La maggiore omogeneità dei gruppi definiti dall'analisi a classi latenti rispetto agli indicatori \bar{R} e σ conferma il risultato dell'analisi grafica ottenuta precedentemente confrontando le nuvole dei punti illustrate nelle figure 4.7 e 4.8. Il confronto tra i due metodi prosegue, infine, con l'analisi delle varianze entro riferite all'indicatore τ_1 e riportate in tabella 4.16.

x	Modello LC	K-medie
1	12,699	12,732
2	28,510	18,370
3	27,133	2,969

Tabella 4.16: varianze entro riferite all'indicatore τ_1 per il modello a classi latenti e la tecnica delle k-medie

Per il primo cluster e la prima classe si conferma lo schema delineato per gli indicatori \bar{R} e σ e caratterizzato da una minore variabilità per il caso delle variabili latenti. Per il secondo e il terzo cluster, invece, è possibile osservare, a differenza dei casi precedenti, una dispersione più contenuta rispetto alle corrispondenti classi latenti. Per la seconda classe è possibile interpretare questo risultato in un quadro che vede ancora l'analisi a classi latenti in posizione di vantaggio rispetto al metodo delle k-medie. La seconda classe, infatti, come peraltro il secondo cluster, raccoglie i titoli peggiori, cioè quelli che non è opportuno inserire in un portafoglio. Data questa caratteristica, diventa irrilevante il livello della variabilità all'interno del gruppo, in quanto i titoli da non includere nel portafoglio possono essere indifferentemente omogenei o eterogenei tra di loro. Per quanto riguarda, inoltre, la terza classe e il terzo cluster, è opportuno

sottolineare come si tratti del gruppo di titoli a maggior rendimento. Rendimenti elevati vanno necessariamente accompagnati da una elevata rischiosità e, quindi, da una forte variabilità: in questa ottica il risultato ottenuto è pienamente coerente con il quadro interpretativo dei fenomeni analizzati.

Nel complesso, il confronto dei risultati ottenuti con l'analisi per classi latenti e con il metodo delle k-medie consente di confermare la validità del percorso proposto per lo studio delle variabili finanziarie. L'analisi a classe latente, in particolare, dimostra come i propri punti di forza, sottolineati dalla letteratura statistica, trovino un puntuale riscontro nelle strategie di investimento.

4.4.3 Il profilo rischio-rendimento nell'analisi a profilo latente

Il passo successivo alla stima del modello a profilo latente consiste nel fornire una accurata interpretazione dei nove gruppi ottenuti. Le probabilità condizionali $g(.|x)$, di cui si riportano le medie $\hat{\mu}(.|x)$ in tabella 4.9, permettono di definire le caratteristiche di ogni classe latente rispetto al profilo di rischio-rendimento atteso. Ad esempio, la Classe 9 contiene un numero limitato di titoli ($\pi_{x=9} = 0,038$, per un totale di 8 azioni) ma presenta la media più alta per l'indicatore \bar{R} . Tuttavia, la valutazione delle tre variabili manifeste riferite al rischio porta alla conclusione che, nonostante l'alto rendimento atteso ($\hat{\mu}(\bar{R} | x = 9) = 1,71$), questa classe è caratterizzata da una elevata volatilità ($\hat{\mu}(\sigma | x = 9) = 19,37$) e risulta fortemente influenzata dalle crisi finanziarie ($\hat{\mu}(\sigma_c | x = 9) = 3,82$). Nella prima colonna della tabella 4.9, la Classe 7 mostra un rendimento medio fortemente negativo ($\hat{\mu}(\bar{R} | x = 7) = -3,17$) ed un rischio piuttosto elevato. Le Classi 4, 3, 2 e 1 sono caratterizzate da rendimenti medi negativi ma presentano livelli di rischio sensibilmente differenti. In particolare, i titoli classificati nella Classe 4 risultano particolarmente volatili nei periodi di crisi ($\hat{\mu}(\sigma_c | x = 4) = 4,69$) e sono soggetti a forti crolli nel prezzo

($\hat{\mu}(\tau_1 | x = 4) = -32,5$). Per contro, la Classe 2, nonostante il rendimento negativo ($\hat{\mu}(\bar{R} | x = 2) = -0,24$) è caratterizzata da un rischio piuttosto moderato come dimostrano i valori medi dei tre indicatori σ , τ_1 e σ_c . Le Classi 5 e 6 sono entrambe caratterizzate da rendimenti attesi positivi e, soprattutto nel caso della Classe 6, da livelli di rischio molto bassi. Infine, una attenzione particolare va riservata alla Classe 8: questo gruppo di titoli azionari, malgrado un rendimento atteso leggermente positivo ($\hat{\mu}(\bar{R} | x = 8) = 0,01$), è fortemente influenzato da forti crolli nei prezzi come suggerisce il più alto valore medio dell'indicatore τ_1 .

La caratterizzazione dei profili dei nove gruppi di azioni facilita una corretta valutazione finanziaria: un buon investimento dovrebbe evitare le Classi 7, 4 e, probabilmente, anche la Classe 8. Per contro, un portafoglio appetibile dovrebbe includere i titoli classificati nelle Classi 5, 6 e, per un maggiore livello di rischio, anche quelli appartenenti alla Classe 9.

In figura 4.9, coerentemente con l'analisi classica del portafoglio, vengono rappresentate le nove classi latenti nel piano rendimento medio – deviazione standard. Dalla figura 4.9, si può notare come l'analisi a profilo latente riesca correttamente a rilevare le posizioni peggiori posizionate nella parte destra del grafico che corrispondono ai titoli che andrebbero evitati. Inoltre, il modello è in grado di discriminare le unità appartenenti alle Classi 6, 2, 1, 5 e 3 che, come suggerisce la sovrapposizione delle bolle in figura 4.9, risultano piuttosto difficili da analizzare.

Come valutazione aggiuntiva rispetto all'analisi classica del portafoglio, il modello a profilo latente permette di analizzare in maniera specifica anche gli eventi estremi ed i periodi di crisi. Nella figura 4.10, le bolle che rappresentano le classi vengono inserite nello spazio primo percentile – deviazione standard durante i periodi di turbolenza finanziaria. Questa rappresentazione facilita la valutazione finanziaria e ribadisce due caratteristiche peculiari: la Classe 8 risulta

fortemente influenzata da forti ribassi nei prezzi, mentre, per quanto riguarda i periodi di crisi, la Classe 4 è caratterizzata da una volatilità molto elevata.

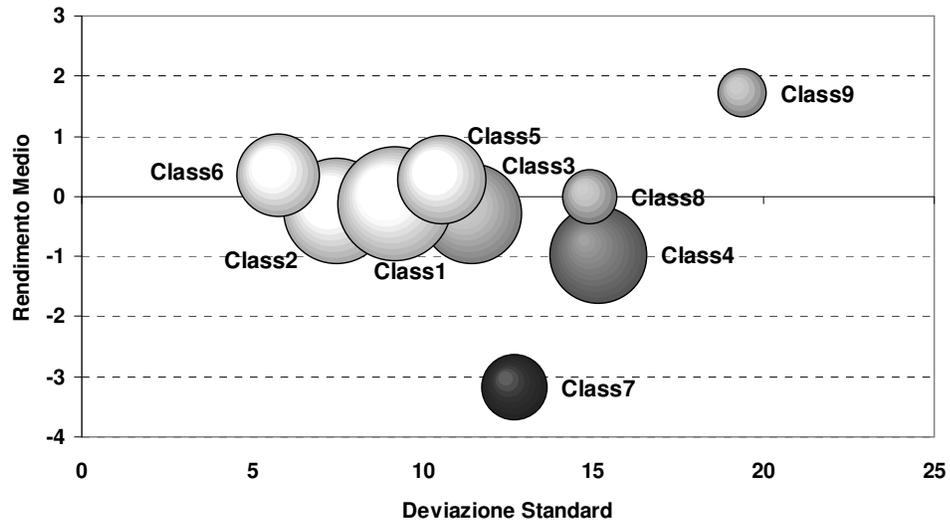


Figura 4.9: le nove classi latenti nello spazio rendimento medio - deviazione standard

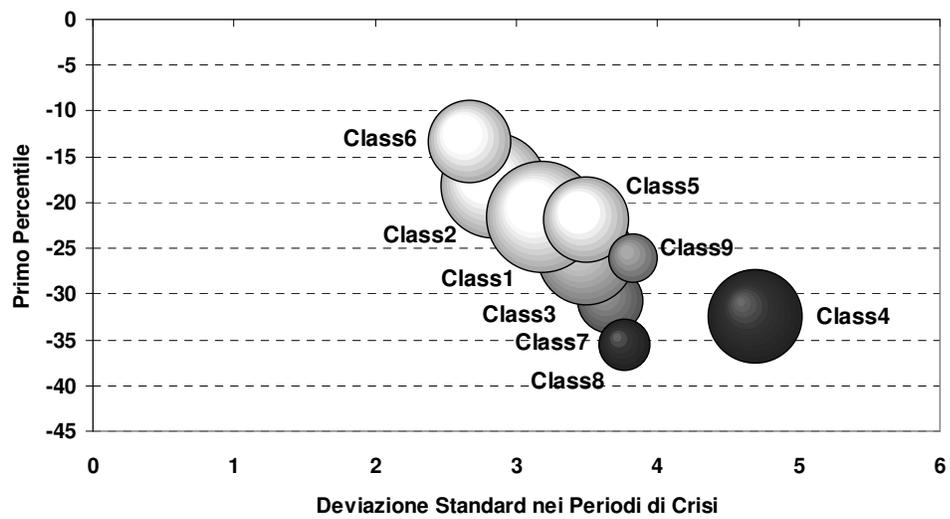


Figura 4.10: le nove classi latenti nello spazio primo percentile – deviazione standard nei periodi di crisi

Infine, con riferimento alla covariata I , dalla figura 4.11 si osserva come le classi latenti con il profilo rischio-rendimento peggiore (le Classi 7, 4 e 8) e la classe con la deviazione standard maggiore (la Classe 9) siano composte totalmente o in buona parte dai titoli appartenenti agli indici meno importanti, ovvero da società a capitalizzazione medio-bassa.

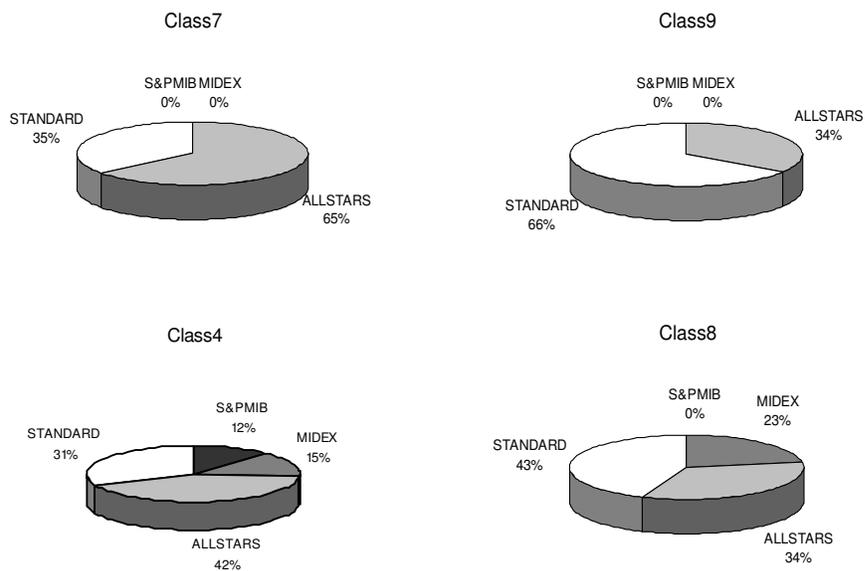


Figura 4.11: la composizione delle Classi 7, 9, 4 e 8 rispetto all'indice azionario (covariata I)

4.4.4 Le implicazioni per l'analisi del portafoglio: la diversificazione settoriale

La stima dei modelli a struttura latente, la classificazione dei titoli nelle classi latenti e la loro interpretazione sotto il profilo rischio-rendimento permette delle interessanti implicazioni per l'analisi del portafoglio.

Riprendendo il confronto tra la classificazione settoriale tradizionale e la nuova ottenuta dal modello a classi latenti introdotta nel paragrafo 4.3.2, è possibile ricorrere alla classificazione dei titoli azionari nelle tre classi per

ottenere un miglioramento nelle procedure di diversificazione degli investimenti. L'attenzione è rivolta alla valutazione dei vantaggi che la nuova classificazione offre rispetto alla definizione settoriale tradizionale nei processi di diversificazione e nella riduzione del rischio del portafoglio finanziario. La precedente tabella 4.10 mostra come i settori GICS contribuiscano in maniera piuttosto simile alla definizione delle tre nuove classi. Tale risultato è interpretabile come prova del fatto che i settori tradizionali non siano consistenti rispetto al profilo rischio-rendimento dei titoli finanziari.

Per poter valutare e confrontare le performance delle due classificazioni, si calcola un numero indice semplice dei prezzi per ognuno dei cinque settori tradizionali e per le tre classi latenti⁹. La tabella 4.17 mostra media, deviazione standard, primo e quinto percentile ed il rapporto di Sharpe di questi indici.

Osservando i rapporti di Sharpe (Sharpe, 1994), che misurano il rendimento in eccesso (rispetto al BOT a 3 mesi) per unità di rischio, la Classe 3 mostra una performance migliore rispetto a tutti i settori tradizionali analizzati, ad eccezione dell'Energia. Al contrario, la Classe 2 risulta avere la performance peggiore.

	Classe1	Classe2	Classe3	Beni Volutt.	Energia	Finanza	Mate- riali	Pubbl. Utilità
\bar{R}	0,016	-0,809	1,048	-0,553	1,110	0,148	-0,321	0,225
σ	4,410	7,404	7,926	6,452	6,896	5,868	5,986	6,265
τ_1	-11,59	-18,97	-17,34	-16,11	-17,14	-14,40	-16,16	-12,38
τ_5	-10,03	-13,06	-10,04	-12,39	-12,55	-10,31	-12,43	-9,87
Sharpe	-0,057	-0,145	0,099	-0,127	0,122	-0,020	-0,098	-0,007

Tabella 4.17: media, deviazione standard, percentili e rapporto di Sharpe degli indici dei prezzi settoriali

⁹ Il metodo di calcolo del numero indice semplice dei prezzi è riportato nella seguente equazione (4.5).

Inoltre, vale la pena notare che la Classe 1 assume i valori più bassi per gli indicatori relativi al rischio, evidenziando la capacità dell'analisi a classi latenti nel definire un gruppo caratterizzato da un livello di rischio molto più contenuto rispetto alla classificazione settoriale tradizionale.

Nel paragrafo **4.5** si analizzano più approfonditamente gli aspetti metodologici riguardanti i numeri indici dei prezzi, valutando come la nuova classificazione conduca ad alcuni importanti miglioramenti nei processi di composizione e del calcolo dei numeri indici dei prezzi finanziari e della loro rappresentatività a livello settoriale.

4.4.5 Le implicazioni per l'analisi del portafoglio: la scelta di un portafoglio efficiente

La determinazione e l'interpretazione su basi finanziarie delle nove classi nell'analisi a profilo latente permette di esaminare uno dei più comuni e diffusi casi di valutazione finanziaria: la scelta di un portafoglio caratterizzato da un profilo rischio-rendimento ottimale.

L'inclusione di un titolo in un portafoglio finanziario dovrebbe essere valutata attentamente in modo da creare la migliore opportunità di investimento possibile. In particolare, oltre al rischio ed al rendimento atteso, è cruciale valutare anche le interrelazioni esistenti tra i titoli che partecipano al portafoglio. In genere, un portafoglio efficiente si ottiene quando si massimizza il rendimento atteso per un dato livello di rischio oppure si minimizza il rischio per un dato livello di rendimento. Sulle frontiere efficienti giacciono i portafogli che presentano la minima deviazione standard per un dato livello di rendimento medio. Nella figura 4.12 vengono rappresentate le frontiere efficienti per ognuna delle nove classi latenti. La figura 4.12 mostra come i gruppi definiti dall'analisi a profilo latente risultino tra loro piuttosto eterogenei e, per tale ragione, siano particolarmente utili per definire delle efficaci strategie di investimento. In particolare, come

mostra la figura 4.13, usando congiuntamente le Classi 6, 5 e 9 si possono proporre un insieme di opportunità di investimento particolarmente attraenti.

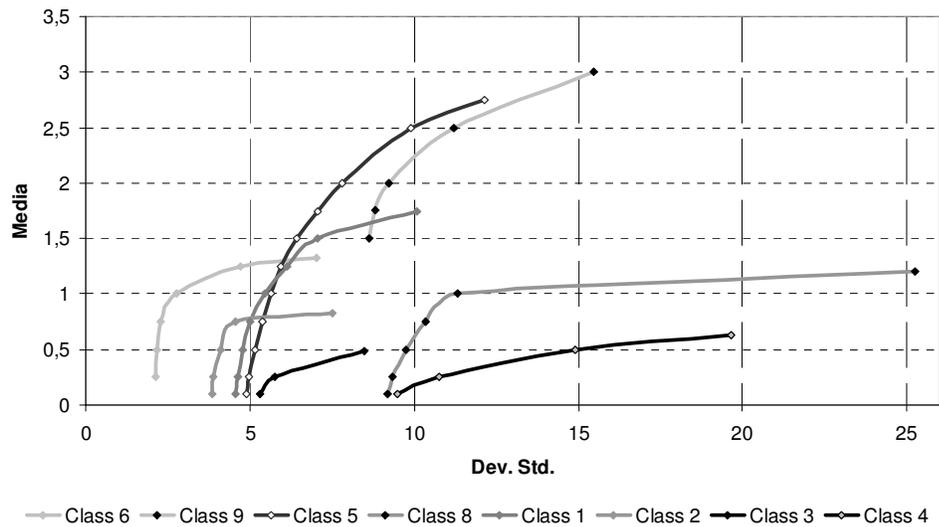


Figura 4.12: le frontiere efficienti per le nove classi latente ottenute dal modello a profilo latente (la Classe 7 non ammette soluzioni positive e quindi non è inserita nel grafico)

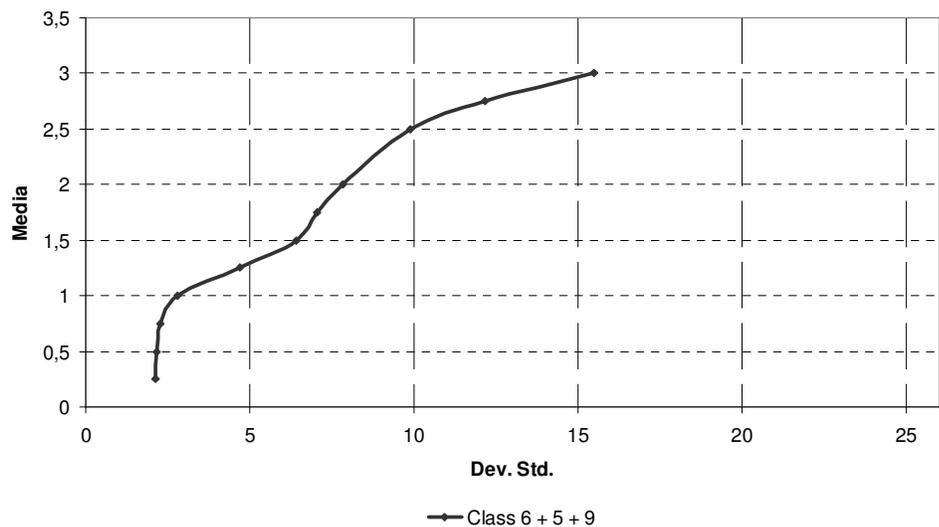


Figura 4.13: la frontiera efficiente ottenuta come combinazione delle Classi 6, 5 e 9

4.5 L'analisi e lo sviluppo dei numeri indici finanziari

La stima del modello a classi latenti e la conseguente classificazione dei titoli azionari in gruppi omogenei rispetto al profilo del rischio – rendimento atteso permettono una ulteriore analisi in un campo in cui il contributo della metodologia statistica risulta fondamentale ma poco sfruttato: lo sviluppo dei numeri indici dei prezzi delle attività finanziarie.

I numeri indici dei prezzi giocano un ruolo di rilievo nei mercati finanziari sia come misure sintetiche delle variazioni, sia come referenza per la diversificazione del rischio. Nonostante l'importanza di questi strumenti, alcune problematiche metodologiche con un rilevante impatto empirico sono state solo marginalmente analizzate. Il presente studio ha lo scopo di contribuire all'analisi di due aspetti.

In primo luogo, riprendendo la classificazione dei titoli azionari ottenuta nel paragrafo **4.3**, si propone un metodo in grado di definire la composizione del paniere sottostante i numeri indici dei prezzi settoriali. I nuovi gruppi ottenuti ricorrendo all'analisi a classi latenti permettono un miglioramento rilevante in termini di copertura e rappresentatività del profilo rischio-rendimento come mostrato nel paragrafo **4.4.4**.

Il secondo aspetto fa riferimento agli effetti che la struttura di ponderazione ha sulla volatilità e sulla correlazione tra gli indici. Questo è un punto chiave in quanto volatilità e correlazione determinano direttamente i processi di diversificazione del rischio. Si confrontano diverse proposte di numeri indici, dalla versione più semplice fino alla metodologia attualmente usata nei mercati finanziari e basata sul concetto di flottante. Inoltre, i diversi numeri indici vengono calcolati considerando la nuova classificazione e quella settoriale tradizionale per evidenziare le interrelazioni tra i processi aggregatori e i metodi di classificazione. L'ultimo passo dello studio fa riferimento alle implicazioni che

la composizione del paniere e la struttura di ponderazione hanno sull'analisi del portafoglio finanziario.

4.5.1 Le questioni metodologiche affrontate

La nuova classificazione settoriale ottenuta ricorrendo alla capacità dei modelli a classi latenti nell'allocare le società per azioni in gruppi omogenei sotto il profilo rischio-rendimento (si veda il paragrafo **4.3.5** e l'**appendice B**) fornisce un metodo semplice ed efficiente per definire la composizione del paniere sottostante gli indici di prezzo settoriali. Inoltre, si contribuisce al dibattito sulla classificazione delle serie storiche finanziarie che sta ricevendo una crescente attenzione nella letteratura statistica (Otranto, 2008; Lisi e Otranto, 2008).

Oltre alla composizione del paniere, anche la struttura di ponderazione dei numeri indici dei prezzi settoriali rappresenta un argomento di grande interesse. In maniera analoga ai numeri indici dei prezzi delle più importanti variabili economiche, vi è stata la tendenza ad adottare le quantità totali dei diversi titoli finanziari come struttura di ponderazione. Rispetto a questa scelta, è stato argomentato che formulazioni alternative, sviluppate sulle basi delle quantità scambiate o potenzialmente scambiabili, possono condurre ad una misura più corretta dei cambiamenti dei prezzi. Diverse strutture di ponderazione possono implicare risultati molto diversi per quanto riguarda la volatilità e le misure di rischio dei numeri indici dei prezzi finanziari (Lisi e Mortandello, 2004). Lo scopo di questa analisi è quello di valutare gli effetti che proposte alternative comportano sui processi di diversificazione del portafoglio.

4.5.2 I numeri indici dei prezzi settoriali

Allo scopo di valutare gli effetti sulle caratteristiche dell'indice, si suggeriscono alcune alternative per la struttura di ponderazione. Inoltre, si confrontano gli indici ottenuti con riferimento alla nuova classificazione ottenuta

con l'ausilio del modello a classi latenti e quella settoriale tradizionale, sviluppando un indicatore sintetico per ogni settore.

L'opzione più diretta ed immediata è il calcolo della media aritmetica degli indici semplici p_{it} / p_{i0} :

$$I1 = \frac{1}{n} \sum_{i=1}^n \frac{p_{it}}{p_{i0}} \quad (4.5)$$

dove p_{it} e p_{i0} sono rispettivamente i prezzi dell' i -esima azione al tempo t e al tempo 0, mentre n è il numero delle azioni incluse nel paniere. Malgrado l'estrema semplicità, l'indice $I1$ può fornire informazioni utili e, dal 1882, è stato adottato per il calcolo degli indici Dow Jones.

La seconda proposta segue le tradizionali regole usate nei mercati finanziari che consistono nel duplicare la metodologia di tipo Laspeyeres sviluppata per i numeri indici dei prezzi relativi alle principali variabili economiche. In questo contesto, la capitalizzazione al tempo base viene usata come struttura di ponderazione:

$$I2 = \frac{\sum_{i=1}^n \frac{p_{it}}{p_{i0}} p_{i0} q_{i0}}{\sum_{i=1}^n p_{i0} q_{i0}} \quad (4.6)$$

dove q_{i0} è il numero di quote dell' i -esima società al tempo 0. La maggioranza dei numeri indici dei prezzi finanziari sono di solito ottenuti seguendo l'espressione (4.6).

Il numero di quote azionarie rappresenta l'elemento fondamentale nella prossima proposta, dove i pesi sono dati da q_{it} :

$$I3 = \frac{\sum_{i=1}^n \frac{p_{it}}{p_{i0}} q_{it}}{\sum_{i=1}^n q_{it}}$$

È inoltre possibile ricorrere ai volumi V_{it} scambiati al tempo t come struttura di ponderazione:

$$I4 = \frac{\sum_{i=1}^n \frac{P_{it}}{P_{i0}} V_{it}}{\sum_{i=1}^n V_{it}}.$$

L'ultima proposta si riferisce alla metodologia attualmente usata nell'indice FTSE-MIB¹⁰ ed è stata formulata nel tentativo di dare una risposta ad alcune osservazioni critiche mosse all'uso dei numeri indici dei prezzi $I2$. L'obiezione principale si riferisce alla rappresentatività di q_i , il numero totale di quote. Quando gli azionisti di maggioranza o di controllo possiedono un numero rilevante di quote, può essere appropriato escludere questa quantità da q_i e dal calcolo dell'indice, in quanto non partecipa alla regolare attività di scambio.

Un riferimento che può assicurare una maggiore rappresentatività è definito sulla base della frazione del flottante f_i . La misura del flottante è tuttora considerata un problema aperto dalla comunità finanziaria poiché non esiste una procedura comune per il calcolo delle quantità f_i . Nel seguito si fa riferimento alla metodologia definita da Standard and Poor's ed adottata anche dal FTSE Group e si suggeriscono sia la ponderazione basata sul numero delle quote pesate per la frazione del flottante:

$$I5 = \frac{\sum_{i=1}^n \frac{P_{it}}{P_{i0}} q_{it} f_{it}}{\sum_{i=1}^n q_{it} f_{it}},$$

sia quella riferita alla capitalizzazione del flottante:

¹⁰ Prima del passaggio agli indici FTSE, la medesima metodologia era applicata all'indice S&P-MIB.

$$I6 = \frac{\sum_{i=1}^n \frac{P_{it}}{P_{i0}} q_{it} f_{it} P_{i0}}{\sum_{i=1}^n q_{it} f_{it} P_{i0}}.$$

Se la scelta del periodo base rappresenta tradizionalmente un punto cruciale nella teoria degli indici dei prezzi, in quelli finanziari gioca un ruolo ancora più rilevante.

Per prima cosa, le revisioni dell'indice modificano frequentemente il paniere originale introducendo o eliminando alcune società. Inoltre, al tempo t , alcune azioni inizialmente incluse nell'indice potrebbero non essere più presenti, mentre alcune altre potrebbero essere state aggiunte al paniere e, di conseguenza, un confronto diretto con il tempo base non sarebbe appropriato. Secondo, le operazioni sul capitale modificano sia i prezzi azionari che il numero delle quote introducendo un gap rispetto al tempo base. Infine, anche pagamenti straordinari o ordinari (ad esempio, i dividendi) modificano il valore dell'azione e, come le operazioni sul capitale, rendono inappropriato il confronto rispetto a p_0 .

Inoltre, nei numeri indici dei mercati finanziari, è necessario aggiustare i valori dei prezzi p_0 e le quantità q_0 relativi al tempo base. Le espressioni degli indici $I1 - I6$ rappresentano un riferimento generale ma la loro computazione empirica richiede la trasformazione di p_0 e q_0 ricorrendo al fattore di aggiustamento k_t che include gli effetti delle modifiche del paniere, delle operazioni sul capitale e dei pagamenti verificatisi nel periodo intercorso tra il tempo base e t .

Siccome non è possibile fare direttamente un confronto corretto tra il tempo t ed il tempo 0, nel seguito si calcolano gli indici precedentemente elencati con riferimento al tempo $t-1$, ma includendo il fattore di aggiustamento k_t :

$$I = \frac{\sum_{i=1}^n \frac{P_{it}}{P_{i0}} k_t W_{it}}{\sum_{i=1}^n W_{it}}$$

dove W_{it} rappresenta la generica struttura di ponderazione.

4.5.3 I risultati

I dati oggetto di indagine empirica, introdotti e descritti nel paragrafo **4.3.3**, consistono nelle serie dei rendimenti mensili di 146 azioni quotate sul mercato italiano e appartenenti a cinque settori della classificazione GICS.

Il primo risultato dell'analisi si riferisce alla composizione del paniere sottostante i numeri indici dei prezzi settoriali, ottenuta attraverso il modello a classi latenti illustrato nel paragrafo **4.3.2**.

Inoltre, si offre un contributo al dibattito sulla struttura di ponderazione dei numeri indici dei prezzi finanziari calcolando gli indici elencati nel paragrafo **4.5.2** considerando come composizioni del paniere la classificazione settoriale tradizionale e la nuova. Lo scopo dell'analisi consiste nell'analizzare gli effetti che la struttura di ponderazione e la composizione del paniere hanno sulla performance degli indici in ottica di diversificazione del portafoglio finanziario.

La stima del modello con l'identificazione di tre classi latenti, permette, come specificato nel paragrafo **4.3.4**, la classificare delle 146 azioni nelle tre classi interpretabili, in questo contesto, come la nuova definizione settoriale. Dalla tabella 4.7 è possibile ricavare le caratteristiche dei profili di rischio-rendimento dei tre nuovi settori.

Per analizzare gli effetti delle diverse strutture di ponderazione, si confrontano le medie e le deviazioni standard dei numeri indici sintetici dei prezzi con panieri relativi alle azioni appartenenti alle tre classi latenti e ai cinque settori GICS e riportate in tabella 4.18.

	<i>I1</i>		<i>I2</i>		<i>I3</i>		<i>I4</i>		<i>I5</i>		<i>I6</i>	
	\bar{R}	σ										
Classe 1	0,02	4,41	0,12	4,35	-0,11	4,52	0,03	5,46	-0,14	4,80	0,10	4,50
Classe 2	-0,81	7,40	-0,79	12,60	-0,78	8,80	0,10	9,59	-0,80	9,31	-0,82	12,71
Classe 3	1,05	7,93	0,73	8,71	0,39	8,25	1,45	10,91	0,20	8,46	0,49	8,62
BeniVolutt.	-0,55	6,45	-0,62	12,01	-0,98	15,19	0,14	8,85	-0,69	9,68	-0,68	12,43
Energia	1,11	6,90	0,78	5,69	0,87	5,82	0,96	5,85	0,74	5,75	0,71	5,62
Finanza	0,15	5,86	-0,03	5,98	-0,11	6,25	0,30	7,58	-0,19	6,38	-0,10	5,96
Materiali	-0,32	5,99	0,07	6,43	-0,58	6,79	0,72	8,28	-0,48	7,19	0,14	6,80
Pubbl.Util.	0,22	6,26	-0,14	4,60	-0,12	4,60	0,52	5,46	-0,17	4,53	-0,18	4,66

Tabella 4.18: media (\bar{R}) e deviazione standard (σ) degli indici relativi alla nuova classificazione e a quella settoriale tradizionale

In figura 4.14, i valori dei numeri indici dei prezzi riportati in tabella 4.18 vengono rappresentati nello spazio media-deviazione standard. Dal confronto dei risultati riportati in tabella 4.18 e dei punti rappresentati nella figura 4.14, emerge nuovamente l'ottima definizione del profilo rischio-rendimento della nuova classificazione dei titoli finanziari. In particolare, si può osservare come, per tutti e sei i numeri indici eccetto *I5*, la Classe 1 presenti i valori di deviazione standard inferiori. La Classe 2 assume, invece, i valori più bassi per il rendimento medio e, nella maggior parte dei casi, la deviazione standard più alta. Infine, la Classe 3 risulta essere particolarmente influenzata dalla struttura di ponderazione: nonostante un livello di deviazione standard piuttosto simile tra i vari numeri indici, le medie dei rendimenti risultano particolarmente differenti, variando da un minimo di 0,20 ad un massimo di 1,05, fatta eccezione per l'indice *I4* che presenta i valori di \bar{R} e σ più alti. Come nel caso della Classe 3, anche i valori di alcuni settori GICS risultano particolarmente eterogenei rispetto alle varie strutture di ponderazione: si noti, ad esempio, la forte variabilità dei valori di \bar{R} per i Materiali e di σ per i Beni Voluttuari nei diversi indici.

Oltre alla composizione del paniere, è possibile osservare alcune differenze rilevanti tra i vari numeri indici dei prezzi. In primo luogo, gli indici a capitalizzazione del flottante (*I6*) hanno una performance molto simile agli indici tradizionali di Laspeyres a capitalizzazione fissa (*I2*). Questo risultato contribuisce al dibattito sul ruolo e l'importanza dell'approccio riferito al flottante. La questione che sorge alla luce di questi risultati è se valga la pena continuare ad usare *I6* come numero indice di riferimento principale del mercato azionario, quando *I2* ha il pregio di essere meno complicato da definire e non richiede un'elaborazione costante (e costosa) dei dati. In secondo luogo, gli indici con struttura di ponderazione definita sulle basi del numero di quote (*I3*) e sul numero di quote del flottante (*I5*) mostrano i valori medi più bassi in 5 casi su 8. Essi sono infatti influenzati più degli altri indici dalla performance delle società con molte quote azionarie emesse nonostante la loro importanza sul mercato sia moderata. Per questo motivo, la struttura di ponderazione basata esclusivamente sul numero delle quote viene usata raramente nella computazione dei numeri indici finanziari. Terzo, è interessante osservare come gli indici basati sui volumi (*I4*) assumono, per tutti i casi eccetto per la Classe 1, i valori più alti per il rendimento medio \bar{R} . In particolare, per la Classe 2 ed il settore dei Beni Voluttuari, il ricorso agli indici *I4* permette di ottenere un valore della media positivo, a differenza degli altri indici sintetici che forniscono un valore di \bar{R} molto inferiore a zero. Infine, vi sono numerosi casi in cui i numeri indici dei prezzi basati sulla media aritmetica semplice (*I1*) mostrano i rendimenti medi più alti ad eccezione dell'indice *I4* (Beni Voluttuari, Energia, Finanza e Pubblica Utilità) o la deviazione standard più contenuta (Classe 2, Classe 3, Beni Voluttuari, Finanza e Materiali).

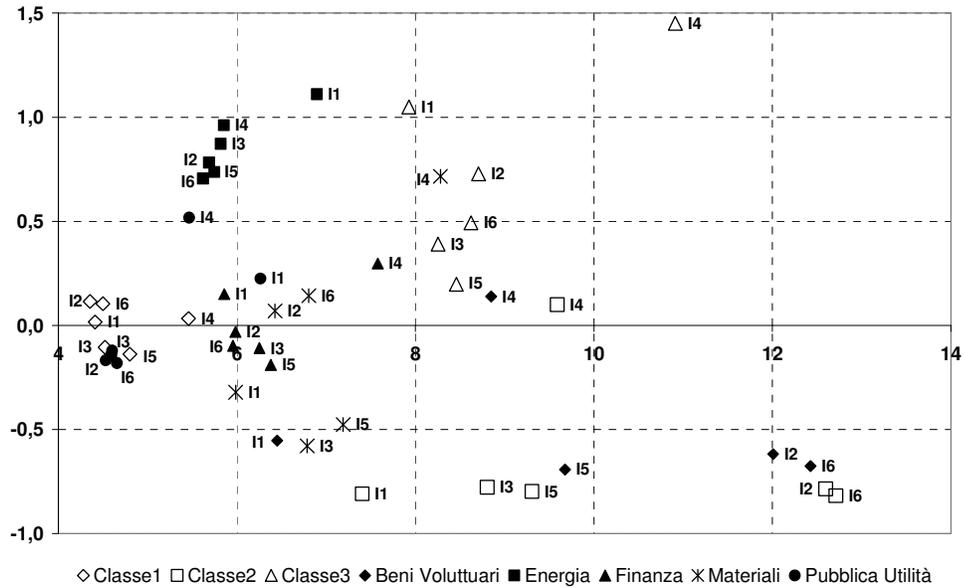


Figura 4.14: i numeri indici sintetici dei prezzi relativi alla nuova e tradizionale definizione settoriale

Al fine di valutare gli effetti dei valori estremi nella nuova classificazione azionaria ed in quella tradizionale, in tabella 4.19 vengono presentati il primo ed il quinto percentile dei diversi numeri indici proposti. La figura 4.15 mostra l'impatto che le diverse strutture di ponderazione hanno sulla volatilità dell'indice dei prezzi: il grafico a barre illustra la deviazione standard ed il primo percentile di 4 numeri indici sintetici.

Infine, in tabella 4.20 e in figura 4.16 vengono riportati i valori dei rapporti di Sharpe (Sharpe, 1994), i quali misurano il rendimento in eccesso (rispetto al BOT a 3 mesi) per unità di rischio, e permettono un confronto diretto tra le performance degli indici sintetici. Dalla analisi dei risultati riportati nelle tabelle 4.18-4.20 e rappresentati nelle figure 4.14-4.16 emerge che la Classe 1 presenta i valori della media, degli indicatori relativi al rischio (σ , τ_1 e τ_5) e del rapporto di Sharpe più simili per le varie tipologie di numeri indici sintetici dei prezzi, a differenza degli

scostamenti più marcati riscontrabili nei risultati riferiti ai settori tradizionali. Questa maggiore omogeneità della prima classe latente, composta dal 45% dei titoli considerati e particolarmente rappresentativa se valutata sul piano del profilo rischio-rendimento, è interpretabile come ulteriore conferma della potenzialità che la nuova classificazione ha nel rilevare gruppi di azioni con caratteristiche finanziarie molto simili.

	I1		I2		I3		I4		I5		I6	
	τ_1	τ_5										
Classe 1	-11,5	-9,8	-13,1	-9,8	-14,3	-10,7	-15,9	-11,1	-15,0	-10,7	-13,3	-10,4
Classe 2	-19,0	-13,1	-29,3	-16,0	-23,2	-13,4	-23,8	-15,3	-23,5	-13,6	-31,5	-16,4
Classe 3	-17,3	-10,4	-25,0	-11,5	-24,9	-11,6	-28,0	-12,0	-26,1	-11,8	-26,1	-11,7
Beni Volutt.	-16,1	-12,4	-29,1	-15,5	-43,1	-27,8	-22,8	-14,0	-23,8	-14,2	-30,8	-16,3
Energia	-17,1	-12,5	-12,7	-10,0	-13,1	-9,5	-12,6	-9,8	-12,7	-9,9	-12,8	-9,7
Finanza	-14,3	-10,2	-18,8	-10,9	-20,5	-11,0	-20,8	-13,7	-20,5	-11,9	-18,7	-11,4
Materiali	-16,2	-12,4	-16,6	-11,6	-16,9	-12,2	-18,8	-11,4	-16,4	-12,7	-16,7	-11,3
Pubbl. Util.	-12,4	-9,9	-13,5	-8,0	-12,4	-8,4	-15,2	-7,6	-13,3	-8,2	-14,1	-8,1

Tabella 4.19: primo (τ_1) e quinto (τ_5) percentile degli indici relativi alla classificazione settoriale nuova e tradizionale

	I1	I2	I3	I4	I5	I6
Classe 1	-0,057	-0,035	-0,082	-0,043	-0,084	-0,036
Classe 2	-0,145	-0,083	-0,119	-0,017	-0,114	-0,085
Classe 3	0,099	0,053	0,015	0,109	-0,008	0,026
Beni Voluttuari	-0,127	-0,074	-0,082	-0,014	-0,099	-0,076
Energia	0,122	0,090	0,104	0,119	0,082	0,078
Finanza	-0,020	-0,050	-0,060	-0,004	-0,072	-0,061
Materiali	-0,098	-0,031	-0,125	0,054	-0,103	-0,018
Pubblica Utilità	-0,007	-0,089	-0,084	0,046	-0,096	-0,096

Tabella 4.20: indice di Sharpe degli indici relativi alla classificazione settoriale nuova e tradizionale

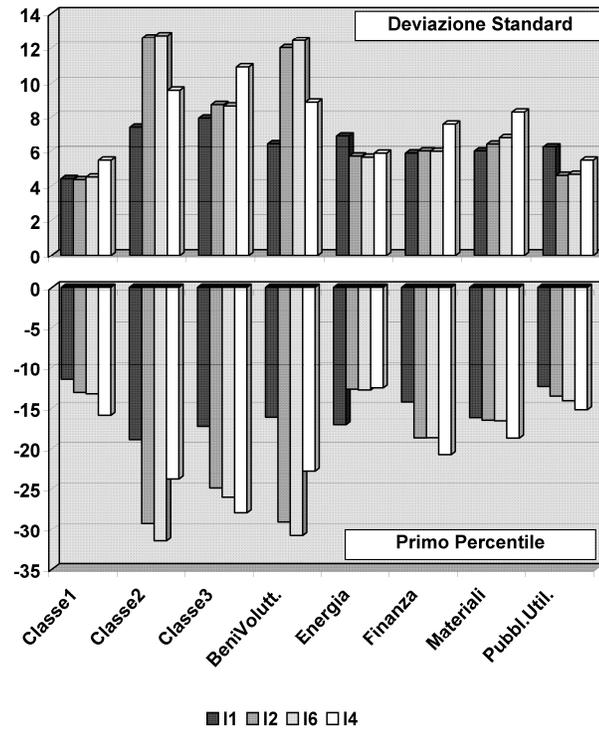


Figura 4.15: effetti della diversa struttura di ponderazione sulla volatilità degli indici dei prezzi: deviazione standard e primo percentile degli indici *I1*, *I2*, *I6* e *I4*

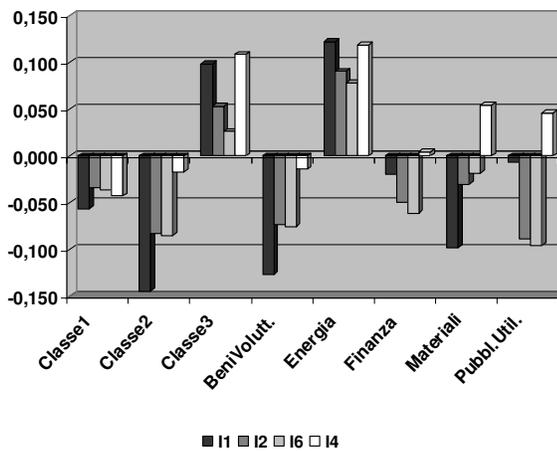


Figura 4.16: rapporto di Sharpe degli indici dei prezzi *I1*, *I2*, *I6* e *I4*

Inoltre, il confronto dei risultati riportati nelle tabelle 4.18-4.20 e nelle figure 4.14-4.16 permette di valutare le varie proposte metodologiche alternative per la costruzione dei numeri indici settoriali. In primo luogo, gli indici *I4* sono spesso caratterizzati dai valori più elevati, mentre viene confermato il fatto che gli indici *I1* sono generalmente caratterizzati dai valori inferiori di deviazione standard e primo percentile rispetto agli altri numeri indici dei prezzi, ad eccezione di quei settori costituiti da pochi titoli, come l'Energia e la Pubblica Utilità. Inoltre, gli indici dei prezzi *I2* ed *I6* conducono a risultati della deviazione standard e del primo percentile molto simili tra loro e, nella quasi totalità dei casi, l'indice riferito alla capitalizzazione totale (*I2*) fornisce dei valori più contenuti rispetto all'indice basato sul flottante (*I6*). Questo ultimo risultato viene confermato anche dal confronto dei valori del rapporto di Sharpe riportati in tabella 4.20 e in figura 4.16: l'indice *I2* assume valori superiori all'indice *I6* in tutti i casi analizzati ad esclusione del settore dei Materiali. Inoltre, l'indice basato sui volumi (*I4*) è caratterizzato dai valori del rapporto di Sharpe superiori rispetto a tutti gli altri indici di tutti i settori analizzati ad eccezione della Classe 1.

4.5.4 Le implicazioni per l'analisi del portafoglio

Nella teoria classica del portafoglio, la media e la deviazione standard della distribuzione dei rendimenti azionari sono i riferimenti principali per derivare i portafogli efficienti, ovvero quelle combinazioni di titoli finanziari che presentano minimo rischio per un dato livello di rendimento medio. La soluzione al problema di minimo ottenuta tramite il metodo dei moltiplicatori di Lagrange implica l'uso della matrice di correlazione, impiegando perciò anche l'informazione sulla struttura di interrelazione tra i titoli. L'insieme dei portafogli efficienti viene chiamato frontiera efficiente al quale, aggiungendo il titolo privo di rischio, si ottiene la Capital Market Line (CML): una semiretta nello spazio media-deviazione standard che sintetizza le migliori opportunità di investimento.

Confrontando le frontiere efficienti basate sui settori tradizionali a quelle relative alla nuova classificazione, è possibile valutare gli effetti della nuova classificazione azionaria sull'analisi del portafoglio.

In particolare, si considerano le combinazioni di portafoglio dei nuovi gruppi e dei settori tradizionali aggiungendo il Buono del Tesoro (BOT) a tre mesi come approssimazione del titolo privo di rischio, ottenendo due CML che permettono un confronto semplice e diretto delle due classificazioni.

In figura 4.17 sono illustrate le frontiere efficienti ottenute all'interno della nuova classificazione azionaria (linea continua) e quelle relative alla definizione settoriale tradizionale (linea tratteggiata). Per osservare gli effetti della struttura di ponderazione sulle opportunità di investimento, si calcolano inoltre le CML sulla base delle diverse proposte di numeri indici. La figura 4.17a mostra i risultati ottenuti usando gli indici semplici $I1$. La figura 4.17b si riferisce agli indici a capitalizzazione totale $I2$, mentre la figura 4.17c riporta il caso della capitalizzazione basata sul flottante. Infine, la figura 4.17d mostra le CML costruite sulla base degli indici ponderati per i volumi di scambio $I4$.

Le frontiere efficienti ottenute ricorrendo alle classi latenti dominano le frontiere efficienti calcolate sui settori GICS tradizionali se valutate rispetto agli indici $I1$ ed $I4$. Nei casi degli indici basati sulla capitalizzazione $I2$ ed $I6$, questa caratteristica della nuova classificazione viene meno. Questo risultato si discosta da quello ottenuto da Costa e De Angelis (2009): in quel lavoro, le CML riferite alle classi latenti dominano in tutti i casi quelle calcolate ricorrendo ai settori tradizionali fornendo delle opportunità di investimento molto più interessanti rispetto a quelle tradizionali, grazie ad una notevole riduzione del livello di rischio. La ragione di questa discrepanza può dipendere dall'orizzonte temporale scelto, in cui il rendimento medio e, quindi, il premio per il rischio sono negativi (si veda il paragrafo 4.2).

Vale anche la pena notare che gli indici *I1* conducono a frontiere efficienti con una performance di rischio-rendimento migliore, mentre i risultati ottenuti usando gli indici *I2* ed *I6* sono piuttosto simili, soprattutto per quanto riguarda i settori tradizionali.

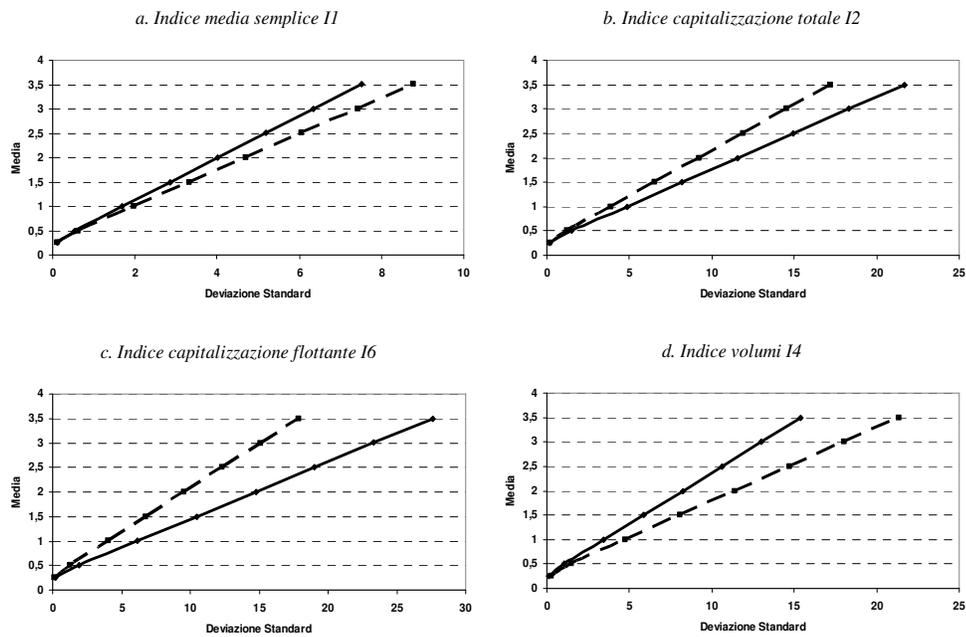


Figura 4.17: le Capital Market Line per la nuova classificazione azionaria (linee continue) e quella tradizionale (linee tratteggiate)

4.5.5 Osservazioni conclusive

Nelle analisi svolte nei paragrafi **4.3-4.5**, si sono proposte delle soluzioni metodologiche interessanti ad alcune problematiche nello studio dei fenomeni finanziari.

In primo luogo, si è mostrato come i metodi statistici a struttura latente, in particolare il modello a classi latenti e l'analisi a profilo latente, rappresentino un metodo appropriato per la definizione di una nuova classificazione, nella quale i titoli azionari vengono allocati in gruppi omogenei sotto il profilo (latente) del

rischio – rendimento atteso. Entrambe le analisi si basano su un rigoroso procedimento scientifico dove il numero dei gruppi viene individuato ricorrendo a numerosi criteri statistici e l'assegnazione delle unità nelle varie classi avviene in termini probabilistici, ovvero, con la consapevolezza del grado di errore nel quale si può incorrere.

La definizione del profilo rischio-rendimento dei vari gruppi di azioni permette una forte discriminazione delle classi latenti rispetto alle caratteristiche finanziari fondamentali alla base della teoria classica del portafoglio. Inoltre, consente di sfruttare ulteriori informazioni rilevanti, quali i metodi di classificazione tradizionali come il settore economico e la capitalizzazione, oppure le performance dei titoli durante i periodi di crisi finanziaria, aumentando l'affidabilità delle stime e indicando nuove opportunità di investimento sia a livello strategico che operativo. La valutazione in termini multivariati del livello di rischio permette di aumentare notevolmente il grado di precisione ed il dettaglio della misura di questa variabile che ricopre un ruolo fondamentale nelle decisioni di investimento ma che, solitamente, risulta molto difficile da calcolare.

La nuova classificazione azionaria conduce ad un miglioramento nei processi di diversificazione settoriale: dal confronto con i settori tradizionali, i tre gruppi ottenuti dal modello a classi latenti sono meglio definiti rispetto al profilo rischio-rendimento e permettono di minimizzare il livello di rischio, creando nuove opportunità di investimento particolarmente interessanti. Una ulteriore implicazione dell'analisi consiste nell'uso dei nove gruppi scaturiti dalla stima del modello a profilo latente per migliorare e semplificare i processi decisionali di investimento ed, in particolare, la scelta di un portafoglio efficiente.

Infine, si sono affrontati due problemi metodologici riguardanti i numeri indici dei prezzi settoriali nei mercati finanziari. In primo luogo, si è mostrato come la classificazione ottenuta dal modello a classi latenti rappresenti un metodo appropriato per definire il paniere sottostante i numeri indici dei prezzi finanziari.

Questa proposta permette di superare alcuni problemi relativi alla definizione dei settori tradizionali e suggerisce una soluzione metodologicamente corretta. L'uso delle classi latenti conduce ad un miglioramento nella qualità dei nuovi settori, in particolare rispetto alla copertura e alla rappresentatività del profilo rischio-rendimento. In secondo luogo, si è definita la composizione del paniere sottostante i numeri indici dei prezzi settoriali seguendo un processo metodologico rigoroso. Nell'analisi si sono costruiti i numeri indici dei prezzi con diverse strutture di ponderazione per entrambe le classificazioni settoriali. I risultati ottenuti sono consistenti con la teoria classica del portafoglio e mostrano che, in alcuni casi, la nuova classificazione azionaria permette una performance migliore rispetto a quella basata sui settori tradizionali, suggerendo perciò nuove ed avanzate opportunità di investimento. Inoltre, si è evidenziato come la classe latente più rappresentativa sul piano del rischio-rendimento risulta essere anche la meno eterogenea rispetto ai vari indici analizzati. Questo risultato può essere considerato come prova del fatto che la metodologia a classi latenti impiegata per definire la nuova classificazione azionaria sia robusta rispetto alla struttura di ponderazione dell'indice. Tuttavia, tale proprietà di robustezza necessita di ulteriori accertamenti, in quanto, i risultati ottenuti non sono del tutto coerenti con quelli ricavati in studi precedenti. Questa incongruenza può essere dovuta al fatto che, nel periodo analizzato ed in contrasto con la teoria economica, il premio per il rischio risulta essere negativo.

Si è inoltre mostrato che gli indici a capitalizzazione basati sul flottante, che hanno goduto di una grande popolarità negli ultimi anni e sono stati adottati come riferimento principale nei mercati finanziari, hanno una performance quasi identica rispetto ai meno complicati e più accessibili numeri indici dei prezzi di tipo Laspeyres basati sulla capitalizzazione di mercato. Infine, l'analisi degli effetti sulla volatilità e sui valori estremi degli indici sintetici sottolineano come la struttura dell'indice più semplice, la media aritmetica, sia caratterizzata, nella

maggior parte dei casi, da una ottima performance. Questo risultato emerge anche valutando le frontiere efficienti.

Una struttura di ponderazione riferita alla capitalizzazione del flottante rappresenta una conquista metodologica e teorica importante ma richiede anche una pesante attività di aggiustamento dei dati. L'introduzione e l'uso di nuovi numeri indici dei prezzi dovrebbe essere attentamente valutata bilanciando accuratamente il maggiore costo che si deve affrontare ed il contenuto informativo fornito che, dall'analisi effettuata, sembra essere molto simile a quello dei numeri indici dei prezzi tradizionali.

4.6 L'analisi della dinamica e la previsione del mercato azionario

In questa analisi, si illustra come il latent Markov model (LMM) presentato nel paragrafo **2.6** possa essere impiegato per riconoscere le diverse condizioni del mercato azionario, definite come regimi. La definizione dei regimi e dei cambiamenti tra gli stessi permettono di rilevare i periodi di crisi finanziaria, definire con precisione e tempismo la fine di una crisi e prevedere gli sviluppi futuri del mercato azionario.

4.6.1 Lo scopo dell'analisi

In tempi di crisi finanziaria, due delle domande più pressanti a cui si vorrebbe dare una risposta sono: cosa succederà tra uno o due mesi e quando la situazione economica tenderà a migliorare. Dall'inizio della crisi finanziaria cominciata alla fine del 2007, queste questioni sono state sollevate diverse volte. In questo paragrafo, si cercherà di rispondere a questi quesiti ricorrendo alle potenzialità del LMM. L'analisi si basa sulla serie storica dei rendimenti mensili del principale indice del mercato azionario italiano: il S&P-MIB.

Il LMM classifica le diverse osservazioni mensili in un insieme limitato di regimi sulla base dei rendimenti dell'indice S&P-MIB ed, inoltre, fornisce le probabilità di passaggio da un regime ad un altro. L'allocatione delle osservazioni mensili nei vari regimi di mercato e le probabilità di cambiamento tra i regimi possono essere sfruttati per individuare la fine di una crisi e prevedere l'andamento futuro dell'indice azionario.

Riprendendo le motivazioni illustrate nel paragrafo **1.4**, la decisione di ricorrere al LMM per ottenere gli scopi sopraelencati è supportata dal fatto che i mercati finanziari sono comunemente caratterizzati da frequenti cambi di regime e che ogni regime di mercato è caratterizzato da profili di rischio-rendimento differenti. Ad esempio, durante una crisi finanziaria, il rendimento medio del

mercato azionario risulta fortemente negativo e la deviazione standard elevata. Durante fasi di mercato più stabili, i rendimenti azionari fluttuano attorno ad una media costante e i valori della deviazione standard sono molto più ridotti. Nell'analisi delle serie storiche, quest'ultima fase rappresenta una serie stazionaria, mentre, nei periodi di turbolenza finanziaria, la proprietà di stazionarietà viene meno.

Come premesso nel paragrafo **1.4**, l'interesse principale dei ricercatori si è indirizzato allo studio della volatilità attraverso modelli per lo studio della varianza condizionata, trascurando però l'informazione riferita alla media condizionata e, quindi, al rendimento atteso. Il ricorso al LMM permette un'indagine approfondita dell'alternanza dei regimi caratterizzati da una alta volatilità da quelli stabili, in cui la variabile di interesse è però la media condizionata dei rendimenti della serie storica analizzata.

Nella teoria del portafoglio di Markowitz e nei suoi seguenti sviluppi, i rendimenti azionari sono assunti come distribuiti normalmente nonostante le analisi empiriche dimostrino chiaramente come questi siano caratterizzati da asimmetria e da una curtosi più elevata rispetto alla distribuzione gaussiana. Il LMM fornisce una soluzione per ovviare a questi problemi modellando le variazioni nei regimi tramite una mistura di distribuzioni normali. Il modello raggruppa in categorie omogenee e non osservabili (usualmente chiamati stati latenti) ogni osservazione temporale, contribuendo ulteriormente alla definizione delle procedure di clustering per serie storiche finanziarie (Frühwirth-Schnatter e Kaufmann, 2008). Gli stati latenti sono caratterizzati da diversi profili di rendimento medio e, perciò, possono essere interpretati come diversi regimi che il mercato azionario può fronteggiare. Inoltre, i modelli di mistura come il LMM forniscono la flessibilità necessaria per trattare casi di asimmetria o curtosi e catturare quasi ogni allontanamento dalla distribuzione normale (Dias et al., 2008).

Gli sviluppi dinamici del mercato azionario possono essere rappresentati dal LMM. Se l'andamento dell'indice di mercato è soggetto a cambiamenti discreti nei regimi, ovvero a periodi in cui la dinamica della serie è marcatamente diversa, allora è utile considerare un modello nonlineare che sfrutta l'andamento temporale della serie osservata per inferire su di un insieme di stati latenti discreti (Hamilton, 1989). Per esempio, il mercato azionario può essere in forte crescita, profondo declino o attraversare una fase di stabilità e i cambiamenti tra questi regimi possono essere modellati attraverso un processo markoviano. La specificazione a catena di Markov non solo offre la possibilità di individuare le variazioni tra una specifica fase di mercato ed un'altra ma permette anche di prevedere la dinamica futura del mercato azionario.

4.5.2 Il modello e i dati

Come mostrato in alcuni recenti lavori (Rydén et al., 1998; Hamilton e Raj, 2002; Dias et al., 2008; Frühwirth-Schnatter e Kaufmann, 2008), il latent Markov model offre la possibilità di analizzare in modo efficace e flessibile la dinamica di serie storiche finanziarie.

Il modello a cui si ricorre in questa analisi consiste nel LMM presentato nel paragrafo 2.5 e rappresentato nelle equazioni (2.23)-(2.25). Il modello considera le variabili latenti y_t varianti nel tempo per $t = 1, \dots, T$ e costituite da S stati latenti:

$$f(z) = \sum_{y_1=1}^S \sum_{y_2=1}^S \dots \sum_{y_T=1}^S f(y_1) \prod_{t=2}^T f(y_t | y_{t-1}) \prod_{t=1}^T f(z_t | y_t).$$

In questa analisi, z denota la serie storica dei rendimenti mensili percentuali dell'indice azionario S&P-MIB da gennaio 2000 fino a dicembre 2008 per un totale di $T = 108$ osservazioni. Come mostrato in figura 4.18 e nella precedente tabella 4.2, il periodo considerato è caratterizzato da due crisi finanziarie mondiali: la crisi del 2000/01 e la crisi cominciata a fine 2007. La figura 4.18

mostra come i periodi di crisi siano caratterizzati da un sensibile aumento della volatilità, ovvero da fluttuazioni più ampie e rapidi cambi tra picchi positivi e negativi. Nel mezzo di queste due crisi, tra la metà del 2003 e la fine del 2007, si osserva un periodo più stabile per il mercato azionario italiano. Come detto in precedenza, questi cambi di regime sono una delle cause principali che limitano la potenza previsiva della maggior parte delle specificazioni più tradizionali come i modelli ARIMA e ARCH.

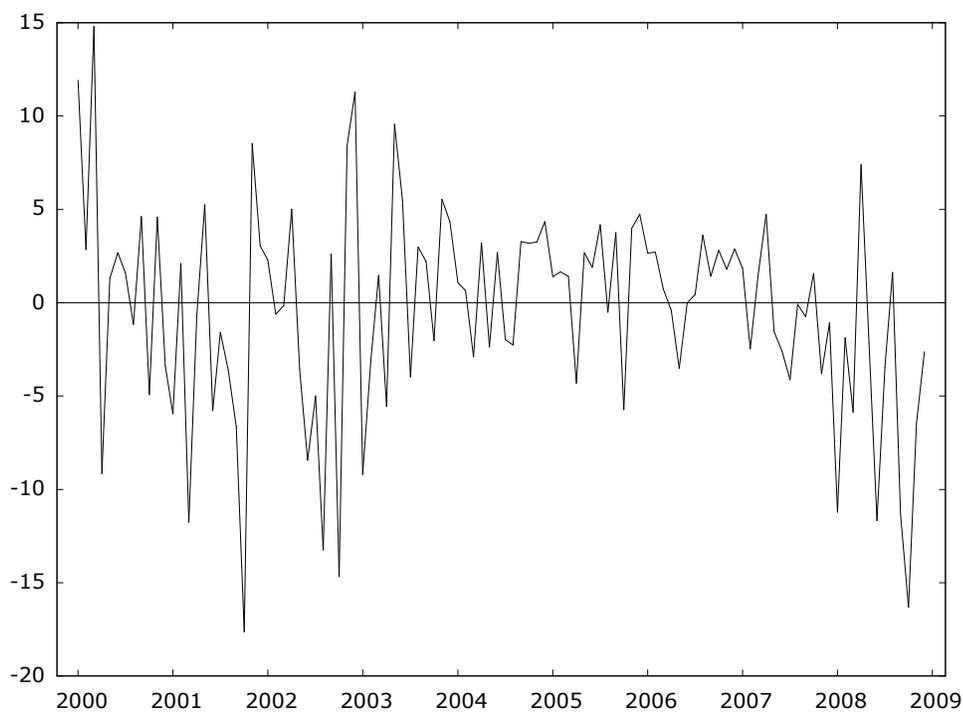


Figura 4.18: la serie storica dei rendimenti mensili dell'indice S&P-MIB da gennaio 2000 a dicembre 2008

La tabella 4.21 mostra i diversi valori dei rendimenti medi e delle deviazioni standard nei due periodi di crisi e in quello stabile. È interessante notare come il rendimento medio del periodo tra giugno 2003 e settembre 2007 sia in valore

assoluto maggiore rispetto a quello della crisi 2000/01. Questa caratteristica sottolinea il fatto che, a seguito di un forte ribasso, il mercato azionario tende a recuperare le perdite subite ed a creare nuova ricchezza. Inoltre, osservando i valori delle deviazioni standard in tabella 4.21, i tre periodi sono caratterizzati da livelli di variabilità marcatamente differenti. La differenza più accentuata si ha tra il periodo di stabilità e le due crisi, ma anche i valori di deviazione standard delle crisi risultano abbastanza diversi tra loro, sottolineando il fatto che ogni periodo di turbolenza finanziaria presenta le proprie peculiarità.

Il test di Jarque-Bera per il data set completo risulta significativo ad un livello di confidenza dell'uno per cento, sottolineando una differenza sostanziale tra la distribuzione osservata e quella normale. Tuttavia, suddividendo la serie storica rispetto ai periodi di stabilità e di crisi, l'assunzione di normalità non viene rifiutata. Questi risultati implicano il fatto che il LMM può essere una valida alternativa ai modelli econometrici finanziari tradizionali in quanto considera anche eventuali asimmetrie e curtosi diverse rispetto alla distribuzione normale senza la necessità di suddividere preliminarmente la serie storica in sotto-periodi omogenei.

Periodo	Rendimento Medio	Deviazione Standard	Asimmetria	Curtosi	Test di Jarque-Bera
Crisi 2000/01 (Gen-00 - Mag-03)	-0,781	7,360	-0,120	2,719	0,23
Crisi 2008 (Ott-07 - Dic-08)	-4,484	6,171	-0,144	2,604	0,15
Fase Stabile (Giu-03 - Set-07)	1,031	2,846	-0,514	2,309	3,32
Data Set Completo (Gen-00 - Dic-08)	-0,423	5,699	-0,549	3,924	9,26*

Tabella 4.21: media, deviazione standard, asimmetria e curtosi dell'indice FTSE-MIB in periodi diversi (* indica una significatività del test al 1%)

4.6.3 La stima del modello

La stima del modello LMM con diversi stati latenti permette di definire il numero dei diversi regimi che caratterizzano il mercato azionario italiano.

Si stima il LMM da 1 a 8 stati latenti ($S = 1, \dots, 8$) e la scelta del modello avviene ricorrendo al criterio CAIC. Dalla tabella 4.22 si evince che il modello con il miglior adattamento ai dati è il LMM con 5 stati latenti (CAIC = 656,00).

I profili dei 5 stati latenti che, in questa analisi corrispondono ai 5 diversi regimi del mercato azionario, sono determinati riferendosi ai rendimenti medi mostrati nella tabella 4.23. In particolare, l'indice S&P-MIB è caratterizzato da tre regimi positivi e due negativi. Riferendosi al test di Wald, i valori dei rendimenti medi differiscono significativamente tra loro ($Wald = 231,05$, $df = 4$, $p-value < 0,001$) evidenziando il fatto che il LMM rappresenta un valido metodo per definire i diversi regimi del mercato azionario italiano. Inoltre, le deviazioni standard di ogni stato latente risultano relativamente basse e piuttosto simili tra loro. Questi risultati sottolineano la capacità del modello di creare gruppi (gli stati latenti) formati da osservazioni tra loro omogenee.

Numero di Stati Latenti	LL	CAIC
1	-340,69	683,39
2	-335,85	677,70
3	-324,95	661,90
4	-318,87	657,75
5	-313,00	656,00
6	-309,27	661,90
7	-305,22	666,45
8	-301,02	674,03

Tabella 4.22: funzione di log-verosimiglianza e criterio CAIC per il LMM da 1 a 8 stati latenti

La dimensione di ogni stato latente misura la proporzione delle osservazioni mensili classificate in un determinato stato e, pertanto, definiscono la probabilità

di trovarsi in un particolare regime di mercato al tempo t . Dalla tabella 4.23, si nota come i due stati estremi risultano essere anche i meno probabili: 0,112 e 0,027 rispettivamente per lo Stato 1 e 5. Al contrario, lo Stato 3 è il più visitato ed intercorre il 54,3% delle volte nell'orizzonte temporale considerato, ovvero $T = 108 \times 0,543 \approx 59$ osservazioni.

I test di Jarque-Bera in tabella 4.23 mostrano il non rifiuto dell'assunzione di normalità distributiva per tutti i 5 stati latenti. Pertanto, la serie storica analizzata viene adeguatamente approssimata da una mistura di cinque distribuzioni normali con medie diverse e simili valori di deviazione standard.

La figura 4.19 illustra gli andamenti della serie storica osservata e della serie stimata dal LMM con 5 stati latenti. Questa ultima viene tracciata considerando le medie dei rendimenti di ogni stato latente. La figura 4.19 mostra come il LMM approssimi piuttosto accuratamente la serie storica dell'indice S&P-MIB. Inoltre, il modello rileva il periodo di stabilità tra le due crisi finanziarie, rappresentato in figura 4.19 dalla linea retta tra giugno 2003 e settembre 2007.

Stato Latente	Dimensione	Rendimento Medio	Deviazione Standard	Jarque-Bera Test
1	0,112	-12,07	2,97	0,77
2	0,245	-2,66	2,81	1,46
3	0,543	1,12	2,82	4,29
4	0,073	6,32	2,50	1,38
5	0,027	12,63	1,87	0,47
Data Set Completo	1,000	-0,42	5,70	9,26*

Tabella 4.23: dimensione, rendimento medio, deviazione standard e test di Jarque-Bera per i 5 stati latenti (* indica una significatività del test al 1%)

La tabella 4.24 riporta la matrice delle probabilità di transizione stimate dal LMM. In questa analisi, queste probabilità definiscono il cambiamento di regime del mercato azionario. I valori sulla diagonale rappresentano la persistenza di un

particolare stato latente, ovvero la probabilità di rimanere in un determinato regime. Il mercato azionario mostra uno stato con alta persistenza, che corrisponde anche allo stato modale: lo Stato 3 ($p_{33} = 0,945$). Questo stato latente rappresenta il regime stabile e, come mostrato in figura 4.19, evidenzia come il mercato azionario tenda a rimanere in questo regime per un tempo piuttosto lungo: $T = (1 - p_{33})^{-1} \approx 18$ mesi.

I valori p_{jk} al di fuori della diagonale indicano le probabilità del cambiamento di regime. Per esempio, è abbastanza probabile che l'indice S&P-MIB passi da un periodo di forte crescita ad una fase molto negativa ($p_{51} = 0,631$). Dai valori nella tabella 4.24, si può notare come il S&P-MIB tenda a cambiare regime piuttosto frequentemente. Infatti, tre probabilità sulla diagonale principale risultano uguali o inferiori a 0,10, indicando come il mercato azionario italiano sia caratterizzato da una limitata persistenza dei regimi.

Questi risultati sono abbastanza differenti rispetto a quelli riportati in altri lavori esistenti in letteratura, in cui le probabilità di persistenza risultano spesso molto elevate per buona parte o, addirittura, per tutti gli stati latenti analizzati e le probabilità al di fuori della diagonale principale molto basse. La spiegazione di questa discordanza è riconducibile allo scopo dell'analisi stessa: in questo lavoro, si è infatti interessati ad investigare le fluttuazioni nella serie dei rendimenti che avvengono durante i periodi di crisi finanziaria per riuscire ad interpretare meglio sia da un punto di vista descrittivo della dinamica temporale sia dal lato operativo, ovvero rispetto alle potenziali opportunità di investimento che si possono creare discriminando i passaggi tra due regimi con medie condizionate molto diverse tra loro. Ad esempio, ottenere una stima affidabile delle probabilità di transizione da un regime caratterizzato da un rendimento negativo ad uno positivo permette la determinazione di una strategia di investimento che può risultare particolarmente redditizia.

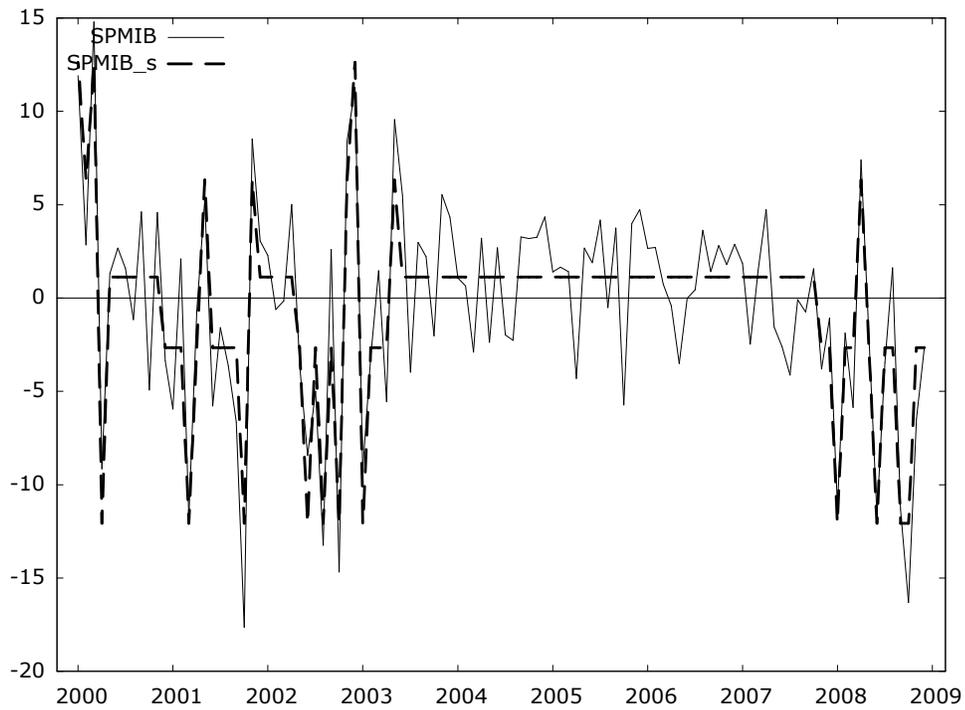


Figura 4.19: la serie storica dei rendimenti mensili dell'indice S&P-MIB osservata e quella stimata dal LMM ($SPMIB_s$)

$j \backslash k$	1	2	3	4	5
1	0,1016	0,5958	0,0869	0,2118	0,0040
2	0,3335	0,5042	0,0034	0,1564	0,0025
3	0,0012	0,0516	0,9450	0,0015	0,0007
4	0,0162	0,3733	0,3171	0,0166	0,2768
5	0,6314	0,0134	0,0134	0,3294	0,0125

Tabella 4.24: matrice di transizione latente P per l'indice S&P-MIB

4.6.4 Le implicazioni per l'analisi finanziaria

La stima del LMM permette di valutare due interessanti questioni: il riconoscimento del periodo di stabilità e la previsione del regime di mercato per il mese successivo.

Come visto nel paragrafo **4.5.3**, lo stato latente 3, caratterizzato da un rendimento medio moderatamente positivo, risulta essere il regime modale e presenta la probabilità di persistenza più alta (si vedano le tabelle 4.23 e 4.24). Quest'ultima caratteristica identifica lo Stato 3 come il regime che corrisponde alla fase più stabile del mercato azionario.

Per valutare la capacità del modello di rilevare il periodo di stabilità, si stima il LMM con 5 stati latenti considerando la serie storica accorciata. Il regime stabile per la serie storica completa comincia nel giugno 2003 e termina nell'ottobre 2007. Si vuole valutare quanti mesi occorrono per rilevare la fine di un periodo di crisi, considerando la crisi del 2000/01. Per raggiungere questo scopo, per prima cosa, si stima il LMM con la serie storica dei rendimenti mensili dell'indice S&P-MIB da gennaio 2000 fino a giugno 2003. Quindi, si riesegue la stima del modello aggiungendo una ulteriore osservazione e, pertanto, analizzando la serie da gennaio 2000 fino ad agosto 2003, e così via. La rilevazione del periodo stabile avviene quando diversi mesi consecutivi vengono classificati nello stato latente 3.

Dall'analisi risulta che il LMM riesce a riconoscere il regime stabile del mercato piuttosto prontamente. La figura 4.20 confronta la serie storica originale con quella derivata dal LMM per l'intero data set e la stima del LMM con 5 stati latenti applicato alla serie accorciata. Ovviamente, i rendimenti medi delle stime del modello basato sulla serie più corta differiscono leggermente rispetto a quelle del LMM per la serie con tutte le osservazioni. Ciononostante, l'allocatione negli stati latenti derivata dalla serie accorciata è quasi la medesima di quella risultante dal LMM per l'intera serie storica. La figura 4.20 mostra che servono 4 mesi per rilevare il regime di stabilità (linea tratteggiata). Bisogna sottolineare il fatto che questo analizzato è il primo ed unico periodo di recupero dopo una crisi presente nel data set ed inoltre che, per questa analisi, si considerano meno di 50 osservazioni. Ripetendo questa indagine su una serie storica con più di un periodo

di stabilità al suo interno, è presumibile che la potenza del LMM di riconoscere un regime stabile migliori ulteriormente. Questa caratteristica del LMM può essere utile per rilevare la fine della crisi finanziaria cominciata a fine 2007 ed ancora in corso come mostrano gli alti valori della deviazione standard degli ultimi mesi del 2008.

Inoltre, è interessante notare dalla tabella 4.23 come il periodo di stabilità possa essere raggiunto dagli stati latenti 4 e 1 ($p_{43} = 0,32$ e $p_{13} = 0,09$) e, con probabilità inferiore, anche dallo stato 5 ($p_{53} = 0,01$). Tuttavia, è bene tener presente che lo stato di stabilità può essere frainteso. Osservando la figura 4.19, infatti, è possibile notare come l'andamento dell'indice S&P-MIB presenti due altri periodi assai più brevi classificati nello stato latente 3: il primo da maggio 2000 a settembre 2000 ed il secondo da dicembre 2001 ad aprile 2002. Questo numero sensibilmente inferiore di mesi allocati nello stato stabile ed i frequenti cambi di regime registrati prima e dopo rendono questi periodi difficilmente interpretabili. Considerare un orizzonte temporale più lungo potrebbe evitare queste potenziali misclassificazioni delle osservazioni.

Come ulteriore analisi, sfruttando le informazioni fornite dalle probabilità di transizione riportate nella tabella 4.24, si valuta la precisione previsiva del LMM. Per raggiungere questo obiettivo, si assume che le probabilità di transizione non cambino nel tempo (Paas et al., 2007). Per verificare questa assunzione, si stima un LMM con probabilità di transizione varianti nel tempo. L'adattamento ai dati fornito da questo modello risulta tuttavia molto peggiore ($CAIC = 2550,50$) rispetto a quello con p_{jk} costanti. Il valore della statistica CAIC è infatti molto più elevato se confrontato con quello del modello con matrice di transizione fissa (si veda la tabella 4.22), a causa del forte incremento nel numero dei parametri da stimare che scaturisce dal rilassamento dell'assunzione di invarianza nel tempo delle probabilità di transizione con $T = 108$.

La tabella 4.24 mostra che alcuni cambiamenti nei regimi possono essere previsti abbastanza precisamente, in quanto la relativa probabilità p_{jk} è alta. Ad esempio, la persistenza del regime stabile è altamente prevedibile. Al contrario, vi sono stati latenti caratterizzati da almeno tre probabilità di transizione superiori a 0,10, rendendo la previsione un passo avanti più difficoltosa. Ad esempio, gli Stati 2 e 4 presentano tre probabilità di transizione superiori a 0,15.

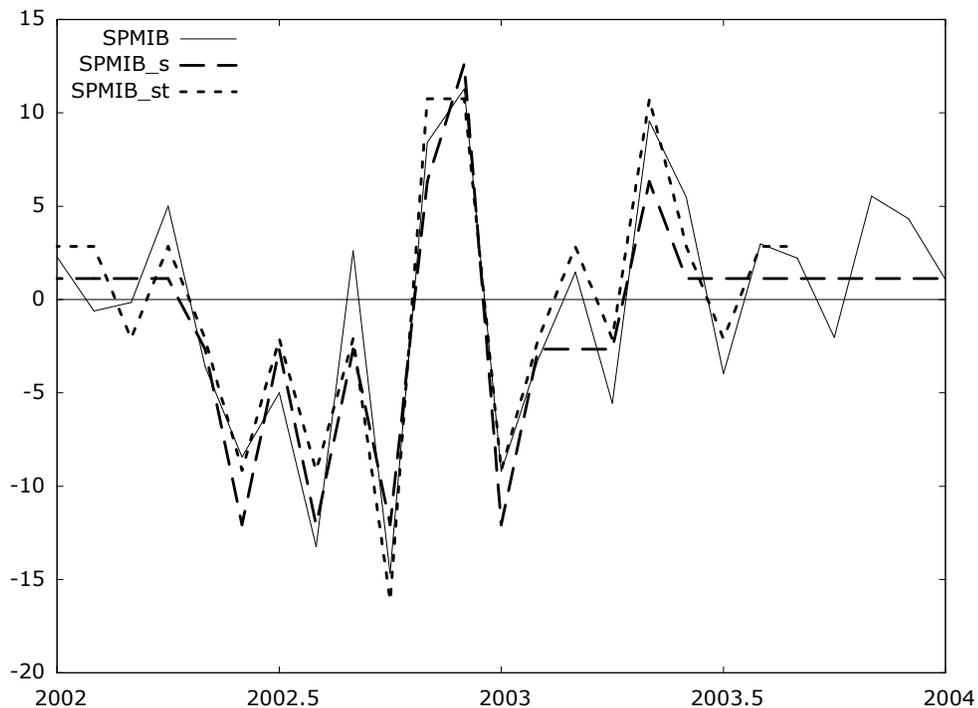


Figura 4.20: la serie storica dei rendimenti mensili dell'indice S&P-MIB osservata, quella stimata dal LMM per l'intera serie (*SPMIB_s*) e quella stimata dal LMM con serie accorciata a settembre 2003 (*SPMIB_st*)

La precisione della previsione del LMM può essere valutata più accuratamente. Nel LMM ogni regime ha una specifica probabilità di verificarsi. Usando queste probabilità, possiamo determinare la potenza predittiva del

modello riferendosi alle previsioni un passo avanti (Paas et al., 2007). I risultati vengono riassunti nella tabella 4.25. Nella colonna 1 della tabella 4.25 si riporta il numero delle volte che il LMM è in grado di prevedere il regime del mese successivo riferendosi alla p_{jk} più alta nella matrice P , la colonna 2 contiene il numero di volte che la previsione del LMM risulta corretta valutando la seconda probabilità di transizione più probabile, e così via. Ad esempio, l'osservazione di novembre 2007 è stata classificata nello stato latente 2, mentre quella riferita a ottobre 2007 nello stato 3. La probabilità di transizione dallo Stato 3 al 2 è $p_{32} = 0,05$ che è la seconda più alta per lo stato latente 3, dopo p_{33} , per cui, questo cambiamento viene conteggiato nella colonna 2 della tabella 4.25. L'ultima colonna fornisce il numero delle volte che il modello non è in grado di prevedere il regime del mese successivo ricorrendo alle tre probabilità di transizione maggiori. Si può notare che la percentuale della colonna "-", che, in un certo senso, può essere considerata come la proporzione di volte che il LMM fallisce nel prevedere il regime del mese successivo, è molto bassa: 0,9%. La percentuale riportata nella colonna 1 risulta essere la più alta (78,5%), mentre basandosi sulle colonne 1 e 2 congiuntamente, la percentuale supera il 90%, indicando una buona performance previsiva un passo avanti del modello.

	1	2	3	-	Total
#	84	16	6	1	107
%	78,5	15,0	5,6	0,9	100

Tabella 4.25: numero delle volte (#) e percentuali (%) in cui il LMM riesce a prevedere correttamente il regime del mese successivo riferendosi alle tre probabilità di transizione maggiori

I risultati ottenuti sono particolarmente utili per la scelta di una proficua strategia di investimento. Un aggiornamento costante dell'analisi dinamica

attraverso il LMM può suggerire la decisione finanziaria più opportuna per il mese successivo. Ad esempio, se le osservazioni riferite ai mesi precedenti sono state classificate nello stato latente 3 allora può essere proficuo mantenere o aumentare l'ammontare dell'investimento in quanto la probabilità di rimanere in questo regime è molto alta. Al contrario, se il mese precedente è stato classificato nello stato 2 allora può essere più opportuno ridurre la quota investita ed evitare ulteriori esposizioni sul mercato in quanto le probabilità di rimanere in uno stato negativo sono maggiori rispetto ad un passaggio ad un regime positivo ($p_{21} + p_{22} = 0,84$ e $p_{24} = 0,16$). Inoltre, è opportuno osservare che, se il mercato si trova nel regime altamente positivo (Stato 5) al tempo $t - 1$, il mese successivo è probabile che avvenga un crollo oppure che continui la fase positiva: $p_{51} = 0,63$ e $p_{54} = 0,33$. In questo caso, risulta difficile decidere quale decisione adottare, tuttavia, valutando le intensità delle due percentuali, può essere più corretto prestare una maggiore cautela e disinvestire.

4.6.5 Osservazioni conclusive e sviluppi futuri

In questa analisi, si è valutato l'andamento dinamico del mercato azionario ricorrendo alle potenzialità del LMM per definire i diversi regimi del mercato e fornire le probabilità di transizione dei cambiamenti dei regimi. Per l'indice S&P-MIB, si sono trovati 5 stati latenti che rappresentano 5 diversi regimi del mercato azionario italiano. Questi ultimi sono chiaramente definiti e caratterizzati da diversi livelli di rendimento medio. Questo risultato è particolarmente interessante in quanto, in letteratura, la maggior parte dei lavori si soffermano principalmente sullo studio della componente di rischio, ottenendo valori dei rendimenti medi condizionati non significativamente diversi tra loro (Ang e Bekaert, 2002). Il LMM è in grado di riconoscere la crisi avvenuta nel 2000/01, quella cominciata alla fine del 2007 ed il lungo periodo di stabilità che intercorre tra queste due crisi.

Il regime stabile è definito da un particolare stato latente caratterizzato da un moderato rendimento positivo e da un'alta probabilità di persistenza.

La caratterizzazione dei regimi e le probabilità di transizione permettono di ottenere due importanti traguardi nell'analisi finanziaria. In primo luogo, il LMM permette di riconoscere prontamente l'inizio del periodo di stabilità in pochi mesi. Questa caratteristica può fornire l'opportunità di individuare la fine della crisi finanziaria cominciata nel 2007 che, a dicembre 2008, non è ancora terminata. Inoltre, il modello fornisce indicazioni su quando il passaggio ad un periodo di stabilità finanziaria è più probabile che avvenga, ad esempio dopo lo stato latente positivo 4. In secondo luogo, il LMM permette di prevedere quale regime caratterizzerà il mercato azionario il mese successivo.

Il latent Markov model fornisce un'analisi della dinamica dell'indice azionario dei prezzi che sarebbe piuttosto difficile da effettuare semplicemente osservando il grafico della serie storica. Il modello permette l'individuazione della fase di stabilità nonostante le fluttuazioni presenti anche durante questo periodo. In altre parole, il modello riesce a discriminare tra le fluttuazioni più limitate del periodo stabile e quelle più ampie che si sviluppano durante una crisi. In particolare, l'andamento della serie nei periodi di crisi sono caratterizzate da quattro prototipi (gli stati latenti) che permettono una migliore comprensione delle fasi di turbolenza finanziaria e della alternanza di rendimenti medi molto differenti che definiscono l'andamento del mercato azionario italiano.

Il presente contributo permette il miglioramento delle opportunità di investimento sia a livello strategico che operativo, consentendo di prendere decisioni di investimento basandosi su di un processo metodologico avanzato. Un limite del presente studio è quello di aver analizzato un periodo di soli 108 mesi. Uno sviluppo futuro può consistere nell'analizzare un maggior numero di osservazioni per valutare se gli stati latenti ottenuti e le analisi fatte sono valide anche in circostanze differenti. Inoltre, si potrebbe verificare la possibile

dipendenza delle probabilità di transizione da variabili esterne, ad esempio variabili macroeconomiche come indicatori dello stato dell'economia reale, per determinare le potenziali implicazioni che tali covariate possono avere sulla dinamica del processo latente sottostante a quello osservato ed, in particolare, sulla matrice P in cui potrebbe essere rilassata l'assunzione di invarianza temporale delle probabilità p_{jk} . Una seconda limitazione riguarda l'approccio in generale: in questa analisi, non si è cercato di valutare e prevedere precisamente le variazioni dei prezzi azionari su base giornaliera. Piuttosto, si è preferito modellare le variazioni dei regimi con dati mensili, assumendo che il LMM sia adatto per analisi di investimento a medio - lungo termine e senza voler ricavare previsioni un passo avanti puntuali ma con il principale interesse di determinare, con un certo grado di probabilità, il regime che caratterizzerà il mercato azionario il mese successivo.

Conclusioni

Metodo statistico ed analisi finanziaria rappresentano due mondi che, pur in presenza di rilevanti tratti comuni, sono caratterizzati da un forte grado di incomunicabilità. Da un lato gli analisti finanziari guardano con sospetto a formule e teoremi che vengono percepiti come ostici ed astrusi, e, soprattutto, di dubbia utilità. A questo proposito è facile ricordare come anche la recente crisi dei mercati finanziari sia stata una ulteriore occasione per attribuire a formule e modelli statistici la responsabilità dei crolli di borsa. Dall'altro lato gli statistici cercano spesso di evitare il confronto con una realtà fenomenica che non sempre rientra docilmente nei canoni previsti. Vengono preferiti risultati asintotici indubbiamente eleganti ma non sempre adatti ad orizzonti temporali finiti, studi su dati simulati e metodi sviluppati in contesti assolutamente generali ed asettici. In questo lavoro si vuole sottolineare come l'interazione tra metodologia statistica ed analisi dei fenomeni finanziari possa essere proficua per entrambe le parti e portare a risultati di profondo interesse sul piano sia teorico sia empirico.

Il filo conduttore della ricerca è rappresentato dai metodi statistici per le variabili latenti, per i quali viene riproposta una rilettura che consente di affrontare e risolvere alcune delle principali problematiche presenti nell'analisi delle variabili finanziarie. La tesi propone una antologia dei principali sviluppi finora formulati in riferimento ai modelli a struttura latente, organizzandoli secondo una chiave di lettura costituita dallo studio di due particolari variabili latenti, il rendimento atteso ed il rischio, che svolgono un ruolo di assoluto primo piano nei mercati finanziari.

Il profilo (latente) di rischio e rendimento atteso delle attività finanziarie viene definito in base ad alcuni indicatori osservabili e valutato considerando i risultati della stima dei modelli a variabili latenti. Il ricorso a tali modelli permette, in primo luogo, di classificare le unità osservate in gruppi omogenei, consentendo di

rendere meno arbitraria la scelta del criterio di classificazione delle unità nelle classi. Le notevoli capacità dei modelli a classi latenti nel contesto della classificazione rappresentano una opportunità estremamente interessante per gli studi finanziari, in quanto i titoli vengono classificati sulla base del proprio profilo latente di rischio e rendimento che sovrasta i criteri di classificazione tradizionali, quali il settore economico oppure la capitalizzazione.

Una valida classificazione delle attività finanziarie rispetto al profilo rischio-rendimento rappresenta anche il presupposto per lo sviluppo di una nuova proposta relativa alla definizione e alla costruzione di numeri indici sintetici dei prezzi. Soprattutto in riferimento al contesto settoriale, la classificazione tradizionale non rappresenta una soluzione soddisfacente sotto diversi aspetti, in particolare sotto il profilo metodologico. I modelli a struttura latente consentono sia di proporre una soluzione metodologicamente corretta per la costruzione dei numeri indici settoriali, sia di ottenere opportunità di investimento più efficaci rispetto ai processi di diversificazione tradizionali, riuscendo a limitare ulteriormente il livello di rischio. Ancora, la classificazione dei titoli azionari permette di facilitare i processi decisionali di selezione di un portafoglio finanziario efficiente basandosi su una metodologia potenziata ed innovativa per la gestione del rischio e dei processi di diversificazione del portafoglio.

Infine, il ricorso ai modelli a variabili latenti consente una importante generalizzazione al contesto dinamico. Risulta così possibile definire delle probabilità di transizione interpretabili come le probabilità di passaggio tra le differenti fasi, o regimi, che contraddistinguono l'andamento dei mercati finanziari. L'introduzione di una valutazione probabilistica della dinamica futura dei prezzi delle attività finanziarie, viene, inoltre, affiancata sia dalla probabilità di permanenza nella fase stessa, sia dalle probabilità di passaggio a fasi differenti. Le caratteristiche dei diversi regimi, infine, permettono una valutazione differenziata del profilo latente di rischio – rendimento delle attività finanziarie rispetto alle

dinamiche del mercato finanziario. Tale insieme di informazioni è particolarmente utile per riconoscere le fasi di crisi e quelle di stabilità che il mercato azionario attraversa, riconoscendo tempestivamente la fine di un periodo di turbolenza finanziaria. Ancora, la specificazione a catena di Markov del modello e l'individuazione del processo stocastico latente sottostante quello osservato favorisce la previsione dell'andamento futuro della serie, migliorando le opportunità di investimento sia a livello strategico che operativo.

I risultati ottenuti a proposito della classificazione delle attività finanziarie, della valutazione del profilo di rischio – rendimento, della misura sintetica sulla base dei numeri indici e dell'analisi dinamica delle variabili finanziarie rappresentano passi importanti ed innovativi nella direzione di valutazioni prese in un contesto metodologico corretto, con una minor influenza di criteri soggettivi e un maggior ricorso a principi scientifici. Il quadro metodologico così delineato non deve, però, implicare automatismi e, soprattutto, non può prescindere dalle conoscenze individuali, dall'intuizione e dall'esperienza soggettiva di operatori e analisti finanziari, il contributo dei quali rimane indispensabile e fondamentale.

Il dibattito sull'interazione tra metodologia statistica ed analisi finanziaria è tuttora in pieno svolgimento, testimone e conseguenza di un panorama ancora in movimento, e, quindi, necessariamente sfumato e incompleto rispetto a tanti particolari e ad alcune linee guida. La tesi si è voluta ispirare e ha cercato di contribuire a questo dibattito, suggerendo soluzioni metodologiche a problematiche finora solo parzialmente affrontate e proponendo direzioni di lavoro per possibili ricerche future.

Bibliografia

Aitkin M., Anderson D., Hinde J. (1981): Statistical Modelling of Data on Teaching Styles. *Journal of the Royal Statistical Society, A*, 144, 419-461.

Akaike H. (1974): A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, 19, 6, 716-723.

Ang A., Bekaert G. (2002): International Asset Allocation with Regime Shifts. *Review of Financial Studies*, 15(4), 1137-1187.

Banfield J.D., Raftery A.E. (1993): Model-based Gaussian and non-Gaussian Clustering. *Biometrics*, 49, 803-821.

Bartholomew D.J., Knott M. (1999): *Latent Variable Models and Factor Analysis*. Kendall's Library of Statistics 7. Oxford University Press, New York.

Bartholomew D.J., Steele F., Moustaki I., Galbraith J. (2002): *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Chapman and Hall, London.

Bartolucci F., Pennoni F., Francis B. (2007): A Latent Markov Model for Detecting Patterns of Criminal Activity. *Journal of Royal Statistical Society A*, 170, 115-132.

Baum L.E., Petrie T., Soules G., Weiss N. (1970): A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematical Statistics*, 41, 164-171.

Bijmolt T.H.A., Paas L.J., Vermunt J.K. (2004): Country and Consumer Segmentation: Multi-level Latent Class Analysis of Financial Product Ownership. *International Journal of Research in Marketing*, 21, 323-340.

Bishop C.M. (1995): *Neural Networks for Pattern Recognition*. Oxford University Press.

Bock R.D. (1972): Estimating Item Parameters and Latent Ability When Responses Are Scored in Two or More Nominal Categories. *Psychometrika*, 37, 29-51.

Bock R.D., Aitkin M. (1981): Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, 46, 443-459.

- Bozdogan H. (1987). Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions. *Psychometrika*, 52(3), 345-370.
- Cappé O., Moulines E., Rydén T. (2005): *Inference in Hidden Markov Models*. Springer - Series in Statistics.
- Castellano R., Scaccia L. (2007): Bayesian Hidden Markov Models for Financial Data. Book of Short Papers, Proceedings CLADAG 2007, 12-14 Settembre, Macerata, Italy, 417-420.
- Clogg C.C. (1981): New Developments in Latent Structure Analysis. In D.J. Jackson and E.F. Borgotta (eds.), *Factor Analysis and Measurement in Sociological Research*. Sage Publications, Beverly Hills, 215-246.
- Croon M. (1990): Latent Class Analysis with Ordered Latent Classes. *British Journal of Mathematical and Statistical Psychology*, 43, 171-192.
- De Menezes L.M. (1999): On Fitting Latent Class Models for Binary Data: the Estimation of Standard Errors. *British Journal of Mathematical and Statistical Psychology*, 52, 149-168.
- Dempster A.P., Laird N.M., Rubin D.B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *Journal of the Royal Statistical Society B*, 39, 1-38.
- Di Zio M., Guarnera U., Rocci R. (2007): A Mixture of Mixture Models for a Classification Problem: The Unity Measure Error. *Computational Statistics & Data Analysis*, 51, 2573-2585.
- Dias J.G., Vermunt J.K. (2007): Latent Class Modelling of Website Users' Search Patterns: Implications for Online Market Segmentation. *Journal of Retailing and Consumer Services*, 14(6), 359-368.
- Dias J.G., Vermunt J.K., Ramos S. (2007): Analysis of Heterogeneous Financial Time Series Using a Mixture Gaussian Hidden Markov Model. Forthcoming.
- Dias J.G., Vermunt J.K., Ramos S. (2008): Mixture Hidden Markov Models in Finance Research. *Advances in Data Analysis, Data Handling and Business Intelligence*. Springer, Berlin. Forthcoming.

- Duekel M.J. (1997): Markov Switching in GARCH Processes and Mean-Reverting Stock-Market Volatility. *Journal of Business & Economic Statistics*, American Statistical Association, 15(1), 26-34.
- Everitt B.S. (1993): *Cluster Analysis*. Edward Arnold, London.
- Fama E.F. (1965): Random Walks in Stock Market Prices. *Financial Analysts Journal* 51(1).
- Fama E.F. (1970): Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25(2), 383–417.
- Formann A.K. (1992): Linear Logistic Latent Class Analysis for Polytomous Data. *Journal of the American Statistical Association*, 87, 476-486.
- Francq C., Zakoïan J.M. (2001): Stationarity of Multivariate Markov-Switching ARMA Models. *Journal of Econometrics*, 102, 339-364.
- Frühwirth-Schnatter S., Kaufmann S. (2008): Model-Based Clustering of Multiple Time Series. *Journal of Business and Economic Statistics* 26, 78–89.
- Goodman L.A. (1974): Explanatory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, 61, 215-231.
- Goodman L.A. (1978): *Analyzing Qualitative\Categorical Data: Log-Linear Models and Latent Structure Analysis*. Addison-Wesley, London.
- Goldfeld S.M., Quandt R.E. (1973): A Markov Model for Switching Regressions. *Journal of Econometrics*, 1, 3-16.
- Gordon A.D. (1999): *Classification*. 2nd Edition. Chapman and Hall/CRC, London.
- Haas M., Mittnik S., Paolella M. (2004). A New Approach to Markov-Switching GARCH Models. *Journal of Financial Econometrics*, 2, 27-62.
- Haberman S.J. (1979): *Analysis of Qualitative Data, Vol. 2, New Developments*. Academic Press, New York.
- Hagenaars J.A. (1988): Latent Structure Models with Direct Effects Between Indicators: Local Dependence Models. *Sociological Methods and Research*, 16, 379-405.

- Hagenaars J.A. (1990): *Categorical Longitudinal Data - Loglinear Analysis of Panel, Trend and Cohort Data*. Sage, Newbury Park.
- Hagenaars J.A. (1993): *Loglinear Models with Latent Variables*. Sage, Newbury Park.
- Hamilton J.D. (1989): A New Approach to the Economic-Analysis of Non-Stationary Time-Series and the Business-Cycle. *Econometrica*, 57, 357-384.
- Hamilton J.D. (2008): Regime-Switching Models. In S.N. Durlaf, L.E. Blume (eds.), *The New Palgrave Dictionary of Economics*, 2nd Edn.
- Hamilton J.D., Susmel R. (1994): Autoregressive Conditional Heteroskedasticity and Changes in Regime. *Journal of Econometrics*, 64, 307-333.
- Hamilton J.D., Raj B. (2002): *Advances in Markov-Switching Models*. Springer-Verlag, Berlin.
- Heinen T. (1996): *Latent Class And Discrete Latent Trait Models: Similarities And Differences*. Sage, Newbury Park.
- Hodson F.R., Sneath P.H.A., Doran J.E. (1966): Some Experiments in the Numerical Analysis of Archaeological Data. *Biometrika*, 53, 311-324.
- Jöreskog K.G. (1971): Statistical Analysis of Sets of Congeneric Tests. *Psychometrika*, 36, 109-133.
- Kaufman L., Rousseeuw P.J. (1990): *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, Inc., New York.
- Lamoureux C.G., Lastrapes W.D. (1993): Forecasting Stock Return Variance: Toward an Understanding of Stochastic Implied Volatilities. *Review of Financial Studies*, 5, 293-326.
- Langeheine R., Pannekoek J., Van de Pol F. (1996): Bootstrapping Goodness-of-fit Measures in Categorical Data Analysis. *Sociological Methods and Research*, 24, 492-516.
- Lazarsfeld P.F. (1950a): The Logical and Mathematical Foundation of Latent Structure Analysis. In S.A. Stouffer et al. (eds.), *Measurement and Prediction*, Princeton University Press, Princeton, NJ, 362-412.

Lazarsfeld P.F. (1950b): The Interpretation and Mathematical Foundation of Latent Structure Analysis. In S.A. Stoufer et al. (eds.), *Measurement and Prediction*, Princeton University Press, Princeton, NJ, 413-472.

Lazarsfeld P.F. (1959): Latent Structure Analysis. In S. Koch (ed.), *Psychology: A Study of a Science*, Vol. 3. McGraw-Hill, New York.

Lazarsfeld P.F., Henry N.W. (1968): *Latent Structure Analysis*. Houghton Mill, Boston.

Lisi F., Otranto E. (2008): Clustering Mutual Funds by Return and Risk Levels. Working Paper CRENoS 200813, Centre for North South Economic Research, University of Cagliari and Sassari, Sardinia.

Lisi, F., Mortandello, F. (2004): Numeri indici di borsa: flottante e volatilità. *Statistica Applicata*, 1, 17–37.

MacKay R.J. (2003): *Hidden Markov Models: Multiple Processes and Model Selection*. Ph.D. thesis, University of British Columbia, Vancouver.

MacKay Altman R.J. (2007): Mixed Hidden Markov Models: an Extension of the Hidden Markov Model to the Longitudinal Data Setting. *Journal of the American Statistical Association*, 102, 201-210.

MacQueen J.B. (1967): Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 281–297.

Magidson J., Vermunt J.K. (2001): Latent Class Factor and Cluster Models, Bi-Plots and Related Graphics Displays. *Sociological Methodology*, 31, 223-264.

Magidson J., Vermunt J.K. (2002a): Latent Class Modeling as a Probabilistic Extension of K-means Clustering. *Quirk's Marketing Research Review*, March.

Magidson J., Vermunt J.K. (2002b): Latent Class Models for Clustering: A Comparison with K-Means. *Canadian Journal of Marketing Research*, 20, 37-44.

Markowitz H. (1952): Portfolio Selection. *Journal of Finance*, 8, 77-91.

Masters G.N. (1982): A Rasch Model for Partial Credit Scoring. *Psychometrika*, 47, 149-174.

- Masters G.N. (1985): A Comparison of Latent Trait and Latent Class Analyses of Likert-type Data. *Psychometrika*, 49, 69-82.
- McCulloch C.E. (1997): Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 92, 162-170.
- McCulloch C.E., Searle S.R. (2001): *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.
- McDonald R.P. (1967): Non-linear Factor Analysis. *Psychometric Monographs*, 15.
- McLachan G.J., Peel D. (2000): *Finite Mixture Models*. John Wiley & Sons, New York.
- Molenaar P.C.W., Von Eye A. (1994): On the Arbitrary Nature of Latent Variables. In A. Von Eye and C.C. Clogg (eds.), *Latent Variables Analysis*, Sage Publications, Thousand Oaks, 226-242.
- Muenz L.R., Rubinstein L.V. (1985): Markov Models for Covariate Dependence of Binary Sequences. *Biometrics*, 41, 91-101.
- Otranto, E. (2008): Clustering Heteroskedastic Time Series by Model-Based Procedures. *Computational Statistics and Data Analysis*, 52, 4685–4698.
- Paas L.J., Bijmolt T.H.A., Vermunt J.K. (2007): Acquisition Patterns of Financial Products: A Longitudinal Investigation. *Journal of Economic Psychology*, 28, 229–241.
- Paas L.J., Vermunt J.K., Bijmolt T.H.A. (2007): Discrete Time, Discrete State Latent Markov Modelling for Assessing and Predicting Household Acquisitions of Financial Products. *Journal of the Royal Statistical Society A* 170, 955–974.
- Raftery A.E. (1986): Choosing Models for Cross-Classifications. *American Sociological Review*, 51, 1, 145-146.
- Rasch G. (1960): Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute of Educational Research, Copenhagen.
- Rydén T., Teräsvirta T., Åsbrink S. (1998): Stylized Facts of Daily Return Series and the Hidden Markov Model. *Journal of Applied Econometrics* 13, 217–244.

- Samejima F. (1969): Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika Monograph No. 17*, 34 (4, Pt. 2).
- Sharpe W.F. (1994): The Sharpe Ratio. *Journal of Portfolio Management*, 21(1), 49–58.
- Smith W.R. (1956): Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *The Journal of Marketing*, 21(1), 3-8.
- Sokal R.R., Sneath P.H.A. (1963): *Principles of Numerical Taxonomy*. W.H. Freeman, San Francisco.
- Sörbom D. (1975): Detection of Correlated Errors in Longitudinal Data. *British Journal of Mathematical and Statistical Psychology*, 28, 138-151.
- Spearman C. (1904): “General intelligence” objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Standard and Poor’s (2008): Frequently Asked Questions About GICS. www2.standardandpoors.com, August 2008.
- Thede S.M., Harper M.P. (1999): A Second-Order Hidden Markov Model for Part-of-Speech Tagging. In *Proceedings of the 37th Annual Meeting of the ACL*, 20-26 June 1999, University of Maryland, USA, 175-182.
- Thissen D., Steinberg L. (1986): A Taxonomy of Item Response Models. *Psychometrika*, 51, 501-519.
- Trivellato U. (1990): Modelli di Comportamento e Problemi di Misura nelle Scienze Sociali: Alcune Riflessioni. In *Società Italiana di Statistica, Atti della XXXV Riunione Scientifica*, 1, Cedam, Padova, 11-34.
- Turner C.M., Startz R., Nelson C.R. (1989): A Markov Model of Heteroskedasticity, Risk, and Learning in the Stock Market. NBER Working Paper, W2818.
- Van de Pol F., Langeheine R. (1990): Mixed Markov Latent Class Models. *Sociological Methodology*, 20, 213-247.
- Vermunt J.K. (2003): Multilevel Latent Class Models. *Sociological Methodology*, 33, 213-239.

Vermunt J.K. (2007): A Hierarchical Mixture Model for Clustering Three-way Data Sets. *Computational Statistics & Data Analysis*, 51, 5368-5376.

Vermunt J.K., Magidson J. (2003): Latent Class Models for Classification. *Computational Statistics & Data Analysis*, 41, 531 – 537.

Vermunt J.K., Langeheine R., Böckenholt U. (1999): Discrete-time Discrete-state Latent Markov Models with Time-constant and Time-varying Covariates. *Journal of Educational and Behavioral Statistics*, 24, 178-205.

Vermunt J.K., Tran B, Magidson J. (2007): Latent Class Models in Longitudinal Research. In S. Menard (ed.), *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, Elsevier, Burlington, MA, 373-385.

Wedel M., DeSarbo W.S. (1994): A Review of Recent Developments in Latent Class Regression Models. In R.P. Bagozzi (ed.), *Advanced Methods of Marketing*, Blackwell, Cambridge, MA, 352–388.

Appendice A – Mercato azionario italiano: le statistiche descrittive dei rendimenti mensili dal 2000 al 2008 dei titoli azionari analizzati

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
A2A	S&P-MIB	Pubblica Utilità	109	-0,254	9,453	-0,112	-17,627	-13,985	-19,68	43,51
ACEA	Midex	Pubblica Utilità	109	0,290	9,486	0,223	-21,553	-13,908	-21,95	35,35
ACEGAS-APS	Standard	Pubblica Utilità	94	-0,456	7,855	-1,439	-15,887	-11,646	-35,20	25,00
ACOTEL GROUP	All-Stars	Informatica	100	-0,051	16,016	-0,953	-38,429	-24,126	-39,31	74,26
ACQUE POTABILI	Standard	Pubblica Utilità	109	-0,121	12,883	-0,936	-24,738	-15,994	-34,23	76,43
ACSM-AGAM	Standard	Pubblica Utilità	109	-1,182	9,318	-1,253	-20,634	-16,670	-22,94	23,14
ACTELIOS	All-Stars	Pubblica Utilità	82	3,727	33,255	-0,716	-24,211	-14,104	-33,03	279,20
AEROPORTO DI FIRENZE	Standard	Beni Voluttuari	101	0,237	8,612	-0,787	-20,769	-10,271	-23,08	32,14
ALERION INDUSTRIES	Standard	Finanza	109	0,945	13,328	-0,967	-20,069	-11,472	-25,14	114,43
ALLEANZA	S&P-MIB	Finanza	109	-0,144	7,682	-0,322	-19,332	-13,602	-22,99	21,11
AMPLIFON	All-Stars	Sanità	90	-0,364	11,857	0,722	-35,832	-20,983	-46,67	29,38
ANSALDO STS	All-Stars	Industriali	33	0,511	5,571	0,407	-8,635	-7,918	-8,91	10,05
ARENA	Standard	Beni Prima Necessità	109	-2,659	12,896	-3,509	-30,369	-21,431	-37,24	48,04
AS ROMA	Standard	Beni Voluttuari	103	-0,488	18,643	-2,556	-31,043	-22,575	-38,63	127,72
ASCOPIAVE	All-Stars	Pubblica Utilità	24	-1,578	5,258	-1,252	-9,169	-8,561	-9,30	11,24
ASTALDI	All-Stars	Industriali	78	0,896	8,900	1,558	-19,912	-15,142	-23,35	22,42
ATLANTIA	S&P-MIB	Industriali	109	0,824	6,876	0,889	-17,409	-10,123	-20,46	21,92
AUTOGRILL	S&P-MIB	Beni Voluttuari	109	-0,238	8,324	0,264	-21,947	-14,769	-34,07	25,39
AUTOSTRADA TO-MI	Midex	Industriali	109	0,172	8,365	-0,059	-28,076	-12,331	-30,25	24,20
AZIMUT HOLDING	Midex	Finanza	53	0,551	9,603	1,812	-23,436	-16,049	-24,49	19,05
BANCA CARIGE	Midex	Finanza	109	0,390	4,899	0,627	-16,537	-6,956	-23,39	12,17
BANCA FINNAT	All-Stars	Finanza	109	1,429	10,384	0,191	-19,574	-11,838	-21,24	56,11

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
BANCA GENERALI	All-Stars	Finanza	25	-4,139	10,247	-5,306	-23,157	-19,400	-23,93	20,81
BANCA IFIS	All-Stars	Finanza	109	2,072	13,146	-0,480	-11,820	-9,895	-16,81	72,20
BANCA INTERMOBILIARE	Standard	Finanza	109	0,620	12,613	-0,207	-16,658	-12,864	-29,20	94,92
BANCA ITALEASE	Midex	Finanza	42	-1,666	18,368	-4,275	-36,042	-28,524	-37,02	31,12
BANCA MONTE DEI PASCHI	S&P-MIB	Finanza	109	-0,221	7,554	0,000	-18,703	-13,639	-28,60	20,65
BANCA POPOLARE DI MILANO	S&P-MIB	Finanza	109	0,074	8,577	-0,125	-18,271	-12,316	-28,95	37,71
BANCA POPOLARE ETRURIA	All-Stars	Finanza	109	-0,459	7,084	-0,793	-16,895	-11,075	-16,97	30,86
BANCA PPO.DI SPOLETO	Standard	Finanza	109	-0,231	6,811	0,000	-15,629	-13,908	-27,14	20,14
BANCA PROFILO	Standard	Finanza	109	0,598	22,269	-0,641	-26,522	-19,137	-34,95	155,72
BANCO DI SARDEGNA RSP	Standard	Finanza	109	-0,562	7,271	0,220	-21,224	-12,135	-25,30	16,37
BANCO POPOLARE	S&P-MIB	Finanza	109	-0,371	7,989	0,388	-20,453	-12,952	-37,64	19,95
BASICNET	Standard	Beni Voluttuari	109	-0,030	16,049	-1,883	-28,324	-18,780	-38,44	99,44
BASTOGI	Standard	Finanza	109	3,255	23,988	-1,008	-25,139	-16,881	-43,27	180,99
BB BIOTECH (MIL)	All-Stars	Finanza	98	-0,821	7,601	-0,113	-19,814	-12,650	-22,60	20,51
BEGHELLI	Standard	Beni Voluttuari	109	-0,306	13,566	-1,923	-28,078	-16,328	-29,38	72,04
BENETTON	Midex	Beni Voluttuari	109	-0,622	9,488	0,000	-23,347	-17,552	-29,44	19,05
BENI STABILI	Midex	Finanza	109	0,781	8,662	0,634	-23,023	-12,676	-24,52	29,21
BIESSE	All-Stars	Industriali	90	-0,028	12,493	-0,179	-30,791	-22,520	-31,50	31,39
BNC.DI DESIO E DELB.	Standard	Finanza	109	0,693	8,024	0,209	-17,549	-11,400	-17,98	23,23
BOERO BARTOLOMEO	Standard	Industriali	109	1,030	5,941	0,000	-9,526	-7,766	-10,36	22,34
BOLZONI	All-Stars	Industriali	30	-1,450	8,731	-0,152	-20,719	-15,330	-22,74	12,27
BONIFICHE FERRARESI	Standard	Beni Prima Necessità	109	1,205	5,712	0,707	-14,867	-6,371	-17,52	16,86
BREMBO	All-Stars	Beni Voluttuari	109	-0,298	7,757	0,386	-17,539	-12,207	-23,54	22,00
BRIOSCHI	Standard	Finanza	109	0,947	15,473	-0,532	-26,510	-15,159	-40,00	113,24
BULGARI	S&P-MIB	Beni Voluttuari	109	-0,021	9,951	0,251	-28,764	-16,176	-34,51	26,63
BUONGIORNO	All-Stars	Informatica	98	-1,704	22,652	-2,732	-37,147	-30,775	-60,37	163,75

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
BUZZI UNICEM	S&P-MIB	Materiali	109	0,474	9,374	-0,293	-20,378	-14,671	-23,12	26,21
CAD IT	All-Stars	Informatica	98	-1,401	8,496	-0,701	-19,912	-14,751	-22,61	24,08
CAIRO COMMUNICATION	All-Stars	Beni Voluttuari	101	-0,427	11,740	-0,415	-27,463	-19,642	-32,74	36,72
CALTAGIRONE	Standard	Finanza	109	0,655	11,440	0,720	-21,200	-13,684	-24,49	78,73
CALTAGIRONE EDITORE	Standard	Beni Voluttuari	101	-1,675	6,662	-1,045	-17,751	-13,657	-25,70	16,01
CAM-FIN	Standard	Energia	109	-0,475	11,421	-1,279	-28,513	-17,111	-35,21	49,01
CARRARO	Standard	Industriali	109	0,092	9,809	0,588	-22,972	-16,482	-37,32	31,25
CATTOLICA ASSICURAZIONI	Midex	Finanza	97	-0,108	6,506	0,444	-14,128	-11,001	-17,70	17,07
CEMBRE	All-Stars	Beni Voluttuari	109	0,263	6,927	0,000	-14,660	-10,316	-18,09	24,32
CEMENTIR HOLDING	Standard	Materiali	109	0,574	9,721	1,196	-23,501	-18,330	-25,72	25,20
CENTRALE DEL LATTE DI TRO.	All-Stars	Beni Prima Necessità	97	-0,678	8,772	-0,209	-19,250	-13,495	-33,16	46,85
CICCOLELLA	Standard	Finanza	109	1,205	16,680	-1,749	-26,962	-15,537	-35,93	84,91
CIR	Midex	Industriali	109	-0,044	15,562	0,465	-32,522	-22,002	-35,49	114,91
CLASS EDITORI	Standard	Beni Voluttuari	109	-1,121	15,324	-2,000	-35,247	-23,182	-48,04	73,73
COBRA	All-Stars	Beni Voluttuari	24	-4,042	12,565	-3,549	-29,396	-22,542	-31,43	19,62
COFIDE	Standard	Industriali	109	1,201	25,241	0,216	-36,887	-20,441	-45,48	228,24
CREDITO ARTIGIANO	Standard	Finanza	109	-0,092	5,234	-0,905	-12,163	-7,984	-14,77	16,08
CREDITO EMILIANO	Midex	Finanza	109	0,311	9,057	1,226	-20,448	-14,813	-31,81	25,33
CREDITO VALTELLINES	Midex	Finanza	109	0,215	5,578	-0,325	-9,774	-7,856	-12,91	22,09
CRESPI	Standard	Materiali	109	-0,904	8,307	-1,667	-19,688	-16,775	-22,93	26,85
CSP INTERNATIONAL	Standard	Beni Voluttuari	109	-1,190	9,547	-1,200	-20,554	-16,369	-23,32	31,25
DADA	All-Stars	Beni Voluttuari	102	-0,368	16,469	0,320	-33,678	-11,878	-24,51	20,01
DANIELI	Standard	Industriali	109	1,078	12,688	-0,474	-34,302	-16,678	-41,33	37,01
DATALOGIC	All-Stars	Informatica	93	0,247	7,019	-3,296	-20,903	-19,867	-38,81	57,46
DAVIDE CAMPARI	Midex	Beni Prima Necessità	89	0,750	6,550	1,294	-15,742	-11,346	-22,02	15,38
DE LONGHI	Standard	Beni Voluttuari	89	-0,453	9,913	-0,677	-20,304	-16,891	-28,88	29,19

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
DEA CAPITAL	All-Stars	Finanza	105	-1,527	13,388	0,345	-29,078	-8,730	-9,00	13,60
DIGITAL BROS	All-Stars	Beni Voluttuari	98	-1,135	13,608	-1,681	-23,557	-19,717	-31,26	82,13
DMAIL GROUP	All-Stars	Beni Voluttuari	96	-0,683	13,688	-2,349	-31,012	-19,768	-31,04	40,00
DMT	All-Stars	Informatica	54	0,046	13,749	0,322	-31,031	-24,819	-32,78	36,13
EDISON	Midex	Pubblica Utilità	109	0,486	8,498	0,163	-18,522	-12,966	-22,86	32,13
EEMS ITALIA	All-Stars	Informatica	32	-5,901	13,261	-5,543	-34,762	-29,469	-37,00	28,36
EL EN	All-Stars	Beni Voluttuari	96	-0,119	8,530	-0,505	-17,850	-14,397	-26,78	34,29
ELICA	All-Stars	Beni Voluttuari	25	-6,204	14,208	-8,996	-39,225	-31,370	-40,71	21,55
EMAK	All-Stars	Beni Voluttuari	109	0,886	6,126	0,254	-14,564	-7,511	-15,79	25,00
ENEL	S&P-MIB	Pubblica Utilità	109	-0,366	5,262	0,070	-17,167	-9,122	-19,38	12,13
ENGINEERING I I	All-Stars	Informatica	96	-0,571	9,491	-0,415	-23,456	-15,408	-23,64	24,05
ENI	S&P-MIB	Energia	109	0,563	5,503	1,546	-13,016	-8,959	-15,56	13,38
ERG	Midex	Energia	109	1,446	9,275	0,809	-18,100	-14,816	-19,41	29,01
ERG RENEW	Standard	Pubblica Utilità	109	0,028	11,433	-1,245	-25,828	-12,350	-45,44	60,93
ERGO PREVIDENZA	Standard	Finanza	109	0,250	10,738	-0,094	-30,715	-13,272	-42,36	37,50
ESPRINET	All-Stars	Informatica	89	1,726	12,276	1,546	-27,546	-17,060	-32,62	43,03
EUROTECH	All-Stars	Informatica	37	-0,276	16,376	-3,017	-25,834	-20,231	-28,12	64,37
EVEREL GROUP	Standard	Beni Voluttuari	109	-2,978	16,061	-3,912	-34,665	-21,675	-35,52	110,30
EXPRIVIA	All-Stars	Informatica	101	-1,371	16,680	-2,209	-31,959	-21,963	-39,07	67,17
FASTWEB	S&P-MIB	Telecomunicazione	105	-1,200	14,795	-3,124	-33,960	-22,580	-40,72	56,15
FIAT	S&P-MIB	Beni Voluttuari	109	-0,986	10,215	0,529	-27,513	-20,009	-30,77	22,44
FIDIA	All-Stars	Industriali	97	-0,842	13,041	-2,148	-24,704	-19,233	-25,02	61,14
FIERA MILANO	All-Stars	Industriali	72	-0,466	8,140	-0,318	-17,613	-12,235	-18,19	46,26
FILATURA DI POLLONE	Standard	Beni Voluttuari	109	-0,775	12,147	-2,453	-27,087	-13,419	-32,53	68,42
FINARTE-SEMENZATO CASA D'ASTE	Standard	Beni Voluttuari	109	-2,127	11,420	-4,068	-22,691	-17,041	-36,30	45,94
FINMECCANICA	S&P-MIB	Industriali	109	-0,113	9,696	0,695	-26,128	-15,206	-27,85	24,36

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
FONDIARIA-SAI	S&P-MIB	Finanza	109	0,064	9,187	-0,257	-24,426	-14,670	-29,72	27,86
GABETTI	Standard	Finanza	109	0,217	11,860	-1,250	-26,469	-16,224	-32,14	48,27
GEFRAN	All-Stars	Beni Voluttuari	109	0,028	7,588	0,000	-14,953	-10,691	-24,71	42,39
GEMINA	Midex	Industriali	109	0,277	13,180	0,548	-35,646	-18,426	-36,75	70,25
GENERALI	S&P-MIB	Finanza	109	-0,044	6,952	0,328	-20,862	-9,349	-23,57	22,43
GEOX	S&P-MIB	Beni Voluttuari	49	0,095	10,200	2,285	-24,490	-17,772	-29,02	19,24
GEWISS	Standard	Industriali	109	-0,304	9,323	-1,515	-25,927	-12,581	-28,09	35,43
GRANDI VIAGGI	Standard	Beni Voluttuari	109	0,088	11,634	-0,665	-23,995	-17,207	-39,97	37,22
GRANITIFIANDRE	All-Stars	Industriali	90	-0,688	7,106	-0,342	-21,686	-9,136	-33,65	16,72
GRUPPO CERAMICHE RICCHETTI	Standard	Industriali	109	-1,026	8,404	-1,220	-18,662	-14,198	-19,16	24,07
GRUPPO COIN	Standard	Beni Voluttuari	109	-0,269	12,875	0,105	-28,812	-18,593	-40,30	53,64
GRUPPO EDIT.L'ESPRESSO	S&P-MIB	Beni Voluttuari	109	-0,859	14,386	-1,992	-24,509	-20,840	-36,20	89,54
HERA	Midex	Pubblica Utilità	66	0,510	6,301	0,650	-16,395	-11,715	-16,46	11,23
I VIAGGI DEL VENTAGLIO	Standard	Beni Voluttuari	91	-2,118	13,988	-3,942	-31,957	-22,094	-44,77	52,05
IFI PV	Midex	Finanza	109	-0,502	11,830	1,406	-35,144	-21,123	-49,64	21,91
IFIL	Midex	Finanza	109	-0,607	10,716	-0,278	-27,756	-18,921	-28,30	29,48
IMA	All-Stars	Industriali	109	0,695	5,244	0,290	-9,990	-7,313	-16,49	10,56
IMMOBILIARE GRANDE DISTRIBUZIONE	All-Stars	Finanza	46	-0,441	10,595	-0,429	-27,586	-17,644	-31,34	20,47
IMMSI	Standard	Beni Voluttuari	106	-0,504	10,553	-1,603	-19,974	-16,702	-21,15	46,72
IMPREGILO	S&P-MIB	Industriali	109	0,475	11,983	0,000	-27,698	-18,392	-30,30	33,81
INDESIT COMPANY	Midex	Beni Voluttuari	109	0,446	9,009	0,697	-20,499	-12,884	-29,21	32,17
INTEK	Standard	Finanza	109	-0,266	9,105	-1,675	-21,427	-10,988	-28,45	29,33
INTERPUMP GROUP	All-Stars	Industriali	109	0,411	7,005	-0,441	-12,435	-9,383	-21,77	18,39
INTESA SANPAOLO	S&P-MIB	Finanza	109	0,051	9,660	0,393	-29,109	-14,630	-29,73	31,92
INVEST E SVILUPPO	Standard	Finanza	109	-1,558	9,522	-0,735	-21,738	-16,028	-24,89	39,71
IPI	Standard	Finanza	109	-0,092	9,097	0,000	-19,609	-9,710	-50,97	32,04

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
IRCE	All-Stars	Beni Voluttuari	109	-0,565	7,167	-0,362	-20,314	-11,055	-21,88	18,63
IRIDE	Midex	Pubblica Utilità	97	-0,779	8,462	0,867	-23,123	-16,561	-31,74	16,95
ISAGRO	All-Stars	Industriali	61	0,684	16,017	-0,894	-35,737	-23,412	-45,92	57,11
IT HOLDING	All-Stars	Beni Voluttuari	109	-1,278	12,744	-2,074	-31,176	-22,003	-43,83	51,27
IT WAY	All-Stars	Informatica	89	-0,800	9,539	-1,553	-25,117	-16,332	-32,62	28,81
ITALCEMENTI	S&P-MIB	Materiali	109	0,028	7,204	0,922	-17,579	-13,665	-19,91	16,69
ITALMOBILIARE	Midex	Materiali	109	0,558	8,230	0,904	-27,096	-10,567	-29,05	21,75
JUVENTUS FOOTBALL CLUB	All-Stars	Beni Voluttuari	84	-0,882	12,462	-1,714	-27,207	-16,706	-45,56	52,74
KINEXIA	Standard	Materiali	109	-0,295	11,871	-0,588	-28,751	-14,549	-35,56	63,64
KME GROUP	Standard	Materiali	109	-1,259	10,670	-1,751	-22,304	-17,835	-25,66	40,84
LA DORIA	All-Stars	Beni Prima Necessità	109	-0,590	7,394	-0,818	-18,946	-12,911	-27,77	22,17
LAVORWASH	Standard	Beni Voluttuari	101	-0,637	10,206	-2,618	-22,222	-13,043	-28,57	43,56
LAZIO	Standard	Beni Voluttuari	109	-2,377	20,390	-4,412	-45,993	-26,912	-67,82	131,19
LOTTOMATICA	S&P-MIB	Beni Voluttuari	91	1,321	7,025	1,575	-14,353	-9,124	-14,58	24,72
LUXOTTICA	S&P-MIB	Beni Voluttuari	97	0,005	7,642	1,024	-17,219	-11,990	-18,58	18,02
MARCOLIN	Standard	Beni Voluttuari	109	0,090	12,329	-1,309	-23,181	-15,991	-31,71	50,76
MARIELLA BURANI	All-Stars	Beni Voluttuari	101	0,647	7,069	0,640	-17,355	-11,410	-23,32	22,17
MARR	All-Stars	Beni Prima Necessità	42	-0,261	6,260	-0,230	-15,157	-10,567	-15,54	17,39
MEDIASET	S&P-MIB	Beni Voluttuari	109	-0,441	10,508	-1,511	-22,489	-16,202	-26,22	50,58
MEDIOBANCA	S&P-MIB	Finanza	109	0,030	7,493	0,000	-18,637	-12,444	-20,37	18,72
MEDIOLANUM	S&P-MIB	Finanza	109	-0,128	13,061	-0,481	-28,908	-19,413	-44,72	58,07
MEDITERRANEA DELLE ACQUE	Standard	Pubblica Utilità	109	0,235	10,052	-0,964	-18,406	-10,423	-22,89	60,00
MILANO ASSICURAZIONI	Midex	Finanza	109	0,200	8,688	0,316	-21,680	-15,991	-23,82	24,63
MIRATO	All-Stars	Materiali	109	0,078	6,727	0,357	-17,171	-9,168	-25,22	23,66
MITTEL	Standard	Finanza	109	0,703	9,134	-0,285	-16,700	-11,344	-33,48	41,55
MONDADORI EDITORE	S&P-MIB	Beni Voluttuari	109	-0,260	12,346	-0,633	-21,100	-17,544	-42,14	67,65

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
MONDO TV	All-Stars	Beni Voluttuari	102	-2,272	9,912	-1,880	-28,808	-16,138	-35,61	30,30
MONRIF	Standard	Finanza	109	-0,667	10,776	-1,566	-23,913	-15,601	-24,25	39,77
MONTEFIBRE	Standard	Materiali	109	-0,417	13,919	-0,635	-32,415	-19,059	-35,17	67,90
NEGRI BOSSI	All-Stars	Industriali	85	-0,651	9,882	-1,338	-27,224	-14,579	-30,55	50,22
NICE	All-Stars	Informatica	31	-3,616	9,668	-1,724	-31,763	-20,379	-33,06	21,27
OLIDATA	Standard	Beni Voluttuari	109	-1,723	14,301	-3,303	-27,825	-19,241	-31,52	77,66
PANARIA GROUP	All-Stars	Materiali	49	-1,593	6,871	-1,071	-18,405	-14,636	-18,79	12,64
PARMALAT	S&P-MIB	Beni Prima Necessità	38	-1,475	7,768	1,024	-17,328	-14,202	-18,77	13,44
PERMASTEELISA	Standard	Industriali	109	0,637	9,529	0,625	-20,725	-18,016	-33,48	27,45
PININFARINA	Standard	Beni Voluttuari	109	-1,598	11,571	-0,866	-34,779	-21,496	-44,58	36,11
PIRELLI	S&P-MIB	Beni Voluttuari	109	-0,432	9,323	0,216	-27,398	-16,502	-33,97	22,80
PIRELLI REAL ESTATE	Midex	Finanza	78	-1,481	11,053	0,532	-35,768	-21,457	-40,89	17,88
POLIGRAFICA S F	All-Stars	Beni Voluttuari	109	-1,310	16,860	-2,679	-21,794	-13,506	-28,02	154,82
POLIGRAFICI EDITORIALE	Standard	Beni Voluttuari	109	-1,116	11,878	-0,324	-25,290	-19,739	-30,82	45,65
POLTRONA FRAU	All-Stars	Beni Voluttuari	25	-4,870	8,443	-2,722	-25,397	-17,338	-27,91	8,36
PREMAFIN-HLDG.DI PART.	Standard	Finanza	109	0,532	8,913	-0,384	-14,522	-11,918	-16,45	42,35
PREMUDA	Standard	Energia	109	1,660	10,186	1,124	-16,004	-10,175	-24,40	69,69
PRIMA INDUSTRIE	All-Stars	Industriali	109	0,414	12,600	-0,464	-28,520	-17,773	-34,96	43,97
RATTI	Standard	Beni Voluttuari	109	-1,357	9,541	-1,801	-16,949	-14,536	-17,06	40,79
RCS MEDIAGROUP	Midex	Beni Voluttuari	109	-0,085	13,330	-0,644	-26,776	-17,354	-42,77	57,14
RECORDATI	Midex	Sanità	109	1,750	10,076	0,718	-19,379	-10,637	-31,29	50,00
RENO DE MEDICI	All-Stars	Materiali	109	-1,957	9,146	-1,435	-26,022	-18,423	-29,16	20,49
REPLY	All-Stars	Informatica	96	0,638	9,491	0,796	-16,852	-13,743	-20,38	53,56
SABAF	All-Stars	Beni Voluttuari	109	0,741	6,923	0,363	-18,116	-8,461	-19,14	29,29
SADI SERVIZI INDUSTRIALI	Standard	Industriali	109	-0,421	11,328	-1,678	-19,678	-14,779	-26,21	68,72
SAES GETTERS	All-Stars	Informatica	109	-0,171	9,870	0,180	-27,212	-16,834	-30,40	24,24

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
SAES GETTERS RNC	All-Stars	Informatica	109	0,523	10,159	0,878	-24,384	-15,204	-40,25	24,00
SAIPEM	S&P-MIB	Energia	109	1,450	9,147	2,760	-24,708	-15,789	-26,31	22,93
SARAS	Midex	Energia	31	-2,028	8,216	-2,844	-17,107	-15,850	-17,24	14,98
SEAT PAGINE GIALLE	S&P-MIB	Beni Voluttuari	64	-2,200	12,324	-0,479	-36,660	-19,112	-38,94	44,07
SIAS	Midex	Industriali	82	0,611	7,084	1,344	-16,748	-10,150	-25,49	20,26
SNAI	Standard	Beni Voluttuari	109	0,630	19,649	-1,044	-36,686	-25,288	-37,00	111,11
SNAM RETE GAS	S&P-MIB	Pubblica Utilità	84	0,707	3,785	0,857	-9,951	-5,302	-12,94	8,79
SNIA ORD	Standard	Industriali	109	-1,232	12,373	-1,692	-37,239	-21,688	-39,31	36,36
SOCOTHERM	All-Stars	Energia	72	-0,651	11,446	-0,424	-33,678	-21,115	-38,22	29,64
SOGEFI	All-Stars	Beni Voluttuari	109	0,063	8,850	0,658	-26,092	-13,693	-36,46	32,35
SOL	Standard	Materiali	109	0,538	6,826	-0,495	-15,435	-7,502	-23,28	19,15
SOPAF	Standard	Finanza	109	-0,790	14,049	-2,798	-35,276	-15,889	-37,19	76,69
SORIN	Standard	Sanità	59	-2,298	8,825	-2,246	-23,748	-16,356	-26,73	21,38
STMICROELECTRONICS (MIL)	S&P-MIB	Informatica	109	-1,125	13,372	-0,942	-32,476	-20,561	-35,74	49,67
TELECOM ITALIA	S&P-MIB	Telecomunicazione	109	-0,619	10,388	-0,579	-24,728	-19,442	-32,84	36,30
TENARIS	S&P-MIB	Energia	72	2,754	12,142	2,764	-31,738	-18,409	-36,17	31,57
TERNA	S&P-MIB	Pubblica Utilità	54	0,579	3,904	1,200	-10,552	-6,192	-10,98	6,78
TISCALI	Midex	Informatica	109	-0,797	22,636	-5,089	-33,846	-27,141	-42,79	130,43
TOD'S	Midex	Beni Voluttuari	97	-0,044	8,561	-0,208	-20,157	-13,109	-33,72	16,20
TREVI FIN INDUSTRIALE	Standard	Industriali	109	2,094	14,881	0,334	-25,795	-16,599	-34,31	77,66
TREVISAN COMETAL	All-Stars	Industriali	61	-1,146	13,370	-1,188	-39,256	-23,596	-44,19	40,36
TXT E-SOLUTION	All-Stars	Informatica	101	-1,120	19,128	-2,718	-29,175	-22,394	-33,32	131,25
UBI BANCA	S&P-MIB	Finanza	66	-0,232	5,219	0,506	-15,081	-10,374	-15,23	8,85
UNI LAND	All-Stars	Finanza	109	0,285	18,894	-1,530	-32,261	-19,554	-34,25	143,24
UNICREDIT	S&P-MIB	Finanza	109	-0,614	7,137	0,505	-21,583	-13,067	-35,61	12,74
UNIPOL	S&P-MIB	Finanza	109	-0,446	5,717	-0,071	-16,603	-9,142	-20,58	16,08

Azione	Indice	Settore GICS	Oss.	\bar{R}	σ	Me	τ_1	τ_5	Min	Max
VIANINI INDR.	Standard	Industriali	109	0,494	11,189	-0,597	-13,902	-9,722	-31,03	91,77
VIANINI LAVORI	Standard	Industriali	109	0,649	8,367	0,262	-22,612	-10,296	-32,70	44,18
VITTORIA ASSICURAZIONI	All-Stars	Finanza	109	0,830	7,507	0,000	-15,409	-10,553	-16,67	25,25
ZUCCHI	Standard	Beni Voluttuari	109	-1,626	8,144	-1,020	-32,343	-12,807	-39,00	20,00

Legenda:

Oss. – numero di osservazioni mensili disponibili

\bar{R} – media dei rendimenti mensili

σ – scarto quadratico medio dei rendimenti mensili

Me – mediana dei rendimenti mensili

τ_1 – primo percentile dei rendimenti mensili

τ_5 – quinto percentile dei rendimenti mensili

Min – minimo dei rendimenti mensili

Max – massimo dei rendimenti mensili

Appendice B – Mercato azionario italiano: la classificazione dei titoli azionari ottenuta dall’analisi a classi latenti

Azione	Settore GICS	\bar{R}	σ	τ_1	M	S	P	CL1	CL2	CL3	CL#	Cluster
A2A	Pubblica Utilità	-0,254	9,453	-17,627	1	2	2	1,00	0,00	0,00	1	1
ACEA	Pubblica Utilità	0,290	9,486	-21,553	2	2	2	0,98	0,00	0,02	1	1
ACEGAS-APS	Pubblica Utilità	-0,456	7,855	-15,887	1	2	2	1,00	0,00	0,00	1	1
ACQUE POTABILI	Pubblica Utilità	-0,121	12,883	-24,738	2	1	1	0,00	0,01	0,99	3	2
ACSM-AGAM	Pubblica Utilità	-1,182	9,318	-20,634	1	2	2	1,00	0,00	0,00	1	1
ACTELIOS	Pubblica Utilità	3,727	33,255	-24,211	2	1	1	0,00	0,01	0,99	3	3
AEROPORTO DI FIRENZE	Beni Voluttuari	0,237	8,612	-20,769	2	2	2	0,99	0,01	0,00	1	1
ALERION INDUSTRIES	Finanza	0,945	13,328	-20,069	2	1	2	0,02	0,03	0,95	3	1
ALLEANZA	Finanza	-0,144	7,682	-19,332	2	2	2	0,96	0,00	0,04	1	1
AS ROMA	Beni Voluttuari	-0,488	18,643	-31,043	1	1	1	0,00	1,00	0,00	2	2
ASCOPIAVE	Pubblica Utilità	-1,578	5,258	-9,169	1	2	2	1,00	0,00	0,00	1	1
AUTOGRILL	Beni Voluttuari	-0,238	8,324	-21,947	1	2	2	0,96	0,04	0,00	1	1
AZIMUT HOLDING	Finanza	0,551	9,603	-23,436	2	1	1	0,00	0,16	0,84	3	1
BANCA CARIGE	Finanza	0,390	4,899	-16,537	2	2	2	0,96	0,00	0,04	1	1
BANCA FINNAT	Finanza	1,429	10,384	-19,574	2	1	2	0,02	0,03	0,95	3	1
BANCA GENERALI	Finanza	-4,139	10,247	-23,157	1	1	1	0,00	0,98	0,02	2	1
BANCA IFIS	Finanza	2,072	13,146	-11,820	2	1	2	0,02	0,03	0,95	3	1
BANCA INTERMOBILIARE	Finanza	0,620	12,613	-16,658	2	1	2	0,02	0,03	0,95	3	1
BANCA ITALEASE	Finanza	-1,666	18,368	-36,042	1	1	1	0,00	0,98	0,02	2	2
BANCA MONTE DEI PASCHI	Finanza	-0,221	7,554	-18,703	2	2	2	0,96	0,00	0,04	1	1
BANCA POPOLARE DI MILANO	Finanza	0,074	8,577	-18,271	2	2	2	0,96	0,00	0,04	1	1
BANCA POPOLARE ETRURIA	Finanza	-0,459	7,084	-16,895	1	2	2	0,99	0,01	0,00	1	1

Azione	Settore GICS	\bar{R}	σ	τ_1	M	S	P	CL1	CL2	CL3	CL#	Cluster
BANCA PPO.DI SPOLETO	Finanza	-0,231	6,811	-15,629	2	2	2	0,96	0,00	0,04	1	1
BANCA PROFILO	Finanza	0,598	22,269	-26,522	2	1	1	0,00	0,16	0,84	3	3
BANCO DI SARDEGNA RSP	Finanza	-0,562	7,271	-21,224	1	2	2	0,99	0,01	0,00	1	1
BANCO POPOLARE	Finanza	-0,371	7,989	-20,453	1	2	2	0,99	0,01	0,00	1	1
BASICNET	Beni Voluttuari	-0,030	16,049	-28,324	2	1	1	0,00	0,98	0,02	2	2
BASTOGI	Finanza	3,255	23,988	-25,139	2	1	1	0,00	0,16	0,84	3	3
BB BIOTECH (MIL)	Finanza	-0,821	7,601	-19,814	1	2	2	0,99	0,01	0,00	1	1
BEGHELLI	Beni Voluttuari	-0,306	13,566	-28,078	1	1	1	0,00	1,00	0,00	2	2
BENETTON	Beni Voluttuari	-0,622	9,488	-23,347	1	2	1	0,14	0,86	0,00	2	1
BENI STABILI	Finanza	0,781	8,662	-23,023	2	2	1	0,50	0,07	0,43	1	1
BNC.DI DESIO E DELB.	Finanza	0,693	8,024	-17,549	2	2	2	0,96	0,00	0,04	1	1
BREMBO	Beni Voluttuari	-0,298	7,757	-17,539	1	2	2	0,96	0,04	0,00	1	1
BRIOSCHI	Finanza	0,947	15,473	-26,510	2	1	1	0,00	0,16	0,84	3	2
BULGARI	Beni Voluttuari	-0,021	9,951	-28,764	2	1	1	0,00	0,98	0,02	2	2
BUZZI UNICEM	Materiali	0,474	9,374	-20,378	2	2	2	1,00	0,00	0,00	1	1
CAIRO COMMUNICATION	Beni Voluttuari	-0,427	11,740	-27,463	1	1	1	0,00	1,00	0,00	2	2
CALTAGIRONE	Finanza	0,655	11,440	-21,200	2	1	2	0,02	0,03	0,95	3	1
CALTAGIRONE EDITORE	Beni Voluttuari	-1,675	6,662	-17,751	1	2	2	0,96	0,04	0,00	1	1
CAM-FIN	Energia	-0,475	11,421	-28,513	1	1	1	0,00	0,98	0,02	2	2
CATTOLICA ASSICURAZIONI	Finanza	-0,108	6,506	-14,128	2	2	2	0,96	0,00	0,04	1	1
CEMBRE	Beni Voluttuari	0,263	6,927	-14,660	2	2	2	0,99	0,01	0,00	1	1
CEMENTIR HOLDING	Materiali	0,574	9,721	-23,501	2	1	1	0,00	0,76	0,23	2	1
CICCOLELLA	Finanza	1,205	16,680	-26,962	2	1	1	0,00	0,16	0,84	3	2
CLASS EDITORI	Beni Voluttuari	-1,121	15,324	-35,247	1	1	1	0,00	1,00	0,00	2	2
COBRA	Beni Voluttuari	-4,042	12,565	-29,396	1	1	1	0,00	1,00	0,00	2	2
CREDITO ARTIGIANO	Finanza	-0,092	5,234	-12,163	2	2	2	0,96	0,00	0,04	1	1

Azione	Settore GICS	\bar{R}	σ	τ_1	M	S	P	CL1	CL2	CL3	CL#	Cluster
CREDITO EMILIANO	Finanza	0,311	9,057	-20,448	2	2	2	0,96	0,00	0,04	1	1
CREDITO VALTELLINES	Finanza	0,215	5,578	-9,774	2	2	2	0,96	0,00	0,04	1	1
CRESPI	Materiali	-0,904	8,307	-19,688	1	2	2	0,99	0,01	0,00	1	1
CSP INTERNATIONAL	Beni Voluttuari	-1,190	9,547	-20,554	1	2	2	0,96	0,04	0,00	1	1
DADA	Beni Voluttuari	-0,368	16,469	-33,678	1	1	1	0,00	1,00	0,00	2	2
DE LONGHI	Beni Voluttuari	-0,453	9,913	-20,304	1	1	2	0,01	0,98	0,00	2	1
DEA CAPITAL	Finanza	-1,527	13,388	-29,078	1	1	1	0,00	0,98	0,02	2	2
DIGITAL BROS	Beni Voluttuari	-1,135	13,608	-23,557	1	1	1	0,00	1,00	0,00	2	2
DMAIL GROUP	Beni Voluttuari	-0,683	13,688	-31,012	1	1	1	0,00	1,00	0,00	2	2
EDISON	Pubblica Utilità	0,486	8,498	-18,522	2	2	2	0,98	0,00	0,02	1	1
EL EN	Beni Voluttuari	-0,119	8,530	-17,850	2	2	2	0,99	0,01	0,00	1	1
ELICA	Beni Voluttuari	-6,204	14,208	-39,225	1	1	1	0,00	1,00	0,00	2	2
EMAK	Beni Voluttuari	0,886	6,126	-14,564	2	2	2	0,99	0,01	0,00	1	1
ENEL	Pubblica Utilità	-0,366	5,262	-17,167	1	2	2	1,00	0,00	0,00	1	1
ENI	Energia	0,563	5,503	-13,016	2	2	2	0,95	0,00	0,04	1	1
ERG	Energia	1,446	9,275	-18,100	2	2	2	0,95	0,00	0,04	1	1
ERG RENEW	Pubblica Utilità	0,028	11,433	-25,828	2	1	1	0,00	0,01	0,99	3	2
ERGO PREVIDENZA	Finanza	0,250	10,738	-30,715	2	1	1	0,00	0,16	0,84	3	2
EVEREL GROUP	Beni Voluttuari	-2,978	16,061	-34,665	1	1	1	0,00	1,00	0,00	2	2
FIAT	Beni Voluttuari	-0,986	10,215	-27,513	1	1	1	0,00	1,00	0,00	2	2
FILATURA DI POLLONE	Beni Voluttuari	-0,775	12,147	-27,087	1	1	1	0,00	1,00	0,00	2	2
FINARTE-SEMENZATO CASA D'ASTE	Beni Voluttuari	-2,127	11,420	-22,691	1	1	1	0,00	1,00	0,00	2	1
FONDIARIA-SAI	Finanza	0,064	9,187	-24,426	2	2	1	0,50	0,07	0,43	1	1
GABETTI	Finanza	0,217	11,860	-26,469	2	1	1	0,00	0,16	0,84	3	2
GEFRAN	Beni Voluttuari	0,028	7,588	-14,953	2	2	2	0,99	0,01	0,00	1	1
GENERALI	Finanza	-0,044	6,952	-20,862	2	2	2	0,96	0,00	0,04	1	1

Azione	Settore GICS	\bar{R}	σ	τ_1	M	S	P	CL1	CL2	CL3	CL#	Cluster
GEOX	Beni Voluttuari	0,095	10,200	-24,490	2	1	1	0,00	0,98	0,02	2	1
GRANDI VIAGGI	Beni Voluttuari	0,088	11,634	-23,995	2	1	1	0,00	0,98	0,02	2	1
GRUPPO COIN	Beni Voluttuari	-0,269	12,875	-28,812	1	1	1	0,00	1,00	0,00	2	2
GRUPPO EDIT.L'ESPRESSO	Beni Voluttuari	-0,859	14,386	-24,509	1	1	1	0,00	1,00	0,00	2	2
HERA	Pubblica Utilità	0,510	6,301	-16,395	2	2	2	0,98	0,00	0,02	1	1
I VIAGGI DEL VENTAGLIO	Beni Voluttuari	-2,118	13,988	-31,957	1	1	1	0,00	1,00	0,00	2	2
IFI PV	Finanza	-0,502	11,830	-35,144	1	1	1	0,00	0,98	0,02	2	2
IFIL	Finanza	-0,607	10,716	-27,756	1	1	1	0,00	0,98	0,02	2	2
IMMOBILIARE GRANDE DISTRIBUZIONE	Finanza	-0,441	10,595	-27,586	1	1	1	0,00	0,98	0,02	2	2
IMMSI	Beni Voluttuari	-0,504	10,553	-19,974	1	1	2	0,01	0,98	0,00	2	1
INDESIT COMPANY	Beni Voluttuari	0,446	9,009	-20,499	2	2	2	0,99	0,01	0,00	1	1
INTEK	Finanza	-0,266	9,105	-21,427	1	2	2	0,99	0,01	0,00	1	1
INTESA SANPAOLO	Finanza	0,051	9,660	-29,109	2	1	1	0,00	0,16	0,84	3	2
INVEST E SVILUPPO	Finanza	-1,558	9,522	-21,738	1	2	2	0,99	0,01	0,00	1	1
IPI	Finanza	-0,092	9,097	-19,609	2	2	2	0,96	0,00	0,04	1	1
IRCE	Beni Voluttuari	-0,565	7,167	-20,314	1	2	2	0,96	0,04	0,00	1	1
IRIDE	Pubblica Utilità	-0,779	8,462	-23,123	1	2	1	0,95	0,03	0,02	1	1
IT HOLDING	Beni Voluttuari	-1,278	12,744	-31,176	1	1	1	0,00	1,00	0,00	2	2
ITALCEMENTI	Materiali	0,028	7,204	-17,579	2	2	2	1,00	0,00	0,00	1	1
ITALMOBILIARE	Materiali	0,558	8,230	-27,096	2	2	1	0,77	0,17	0,06	1	2
JUVENTUS FOOTBALL CLUB	Beni Voluttuari	-0,882	12,462	-27,207	1	1	1	0,00	1,00	0,00	2	2
KINEXIA	Materiali	-0,295	11,871	-28,751	1	1	1	0,00	1,00	0,00	2	2
KME GROUP	Materiali	-1,259	10,670	-22,304	1	1	1	0,00	1,00	0,00	2	1
LAVORWASH	Beni Voluttuari	-0,637	10,206	-22,222	1	1	1	0,00	1,00	0,00	2	1
LAZIO	Beni Voluttuari	-2,377	20,390	-45,993	1	1	1	0,00	1,00	0,00	2	2
LOTTOMATICA	Beni Voluttuari	1,321	7,025	-14,353	2	2	2	0,99	0,01	0,00	1	1

Azione	Settore GICS	\bar{R}	σ	τ_1	M	S	P	CL1	CL2	CL3	CL#	Cluster
LUXOTTICA	Beni Voluttuari	0,005	7,642	-17,219	2	2	2	0,99	0,01	0,00	1	1
MARCOLIN	Beni Voluttuari	0,090	12,329	-23,181	2	1	1	0,00	0,98	0,02	2	1
MARIELLA BURANI	Beni Voluttuari	0,647	7,069	-17,355	2	2	2	0,99	0,01	0,00	1	1
MEDIASET	Beni Voluttuari	-0,441	10,508	-22,489	1	1	1	0,00	1,00	0,00	2	1
MEDIOBANCA	Finanza	0,030	7,493	-18,637	2	2	2	0,96	0,00	0,04	1	1
MEDIOLANUM	Finanza	-0,128	13,061	-28,908	2	1	1	0,00	0,16	0,84	3	2
MEDITERRANEA DELLE ACQUE	Pubblica Utilità	0,235	10,052	-18,406	2	1	2	0,03	0,00	0,97	3	1
MILANO ASSICURAZIONI	Finanza	0,200	8,688	-21,680	2	2	2	0,96	0,00	0,04	1	1
MIRATO	Materiali	0,078	6,727	-17,171	2	2	2	1,00	0,00	0,00	1	1
MITTEL	Finanza	0,703	9,134	-16,700	2	2	2	0,96	0,00	0,04	1	1
MONDADORI EDITORE	Beni Voluttuari	-0,260	12,346	-21,100	1	1	2	0,01	0,98	0,00	2	1
MONDO TV	Beni Voluttuari	-2,272	9,912	-28,808	1	1	1	0,00	1,00	0,00	2	2
MONRIF	Finanza	-0,667	10,776	-23,913	1	1	1	0,00	0,98	0,02	2	1
MONTEFIBRE	Materiali	-0,417	13,919	-32,415	1	1	1	0,00	1,00	0,00	2	2
OLIDATA	Beni Voluttuari	-1,723	14,301	-27,825	1	1	1	0,00	1,00	0,00	2	2
PANARIA GROUP	Materiali	-1,593	6,871	-18,405	1	2	2	0,99	0,01	0,00	1	1
PININFARINA	Beni Voluttuari	-1,598	11,571	-34,779	1	1	1	0,00	1,00	0,00	2	2
PIRELLI	Beni Voluttuari	-0,432	9,323	-27,398	1	2	1	0,14	0,86	0,00	2	2
PIRELLI REAL ESTATE	Finanza	-1,481	11,053	-35,768	1	1	1	0,00	0,98	0,02	2	2
POLIGRAFICA S F	Beni Voluttuari	-1,310	16,860	-21,794	1	1	2	0,01	0,98	0,00	2	3
POLIGRAFICI EDITORIALE	Beni Voluttuari	-1,116	11,878	-25,290	1	1	1	0,00	1,00	0,00	2	2
POLTRONA FRAU	Beni Voluttuari	-4,870	8,443	-25,397	1	2	1	0,14	0,86	0,00	2	2
PREMAFIN-HLDG.DI PART.	Finanza	0,532	8,913	-14,522	2	2	2	0,96	0,00	0,04	1	1
PREMUDA	Energia	1,660	10,186	-16,004	2	1	2	0,02	0,04	0,94	3	1
RATTI	Beni Voluttuari	-1,357	9,541	-16,949	1	2	2	0,96	0,04	0,00	1	1
RCS MEDIAGROUP	Beni Voluttuari	-0,085	13,330	-26,776	2	1	1	0,00	0,98	0,02	2	2

Azione	Settore GICS	\bar{R}	σ	τ_1	M	S	P	CL1	CL2	CL3	CL#	Cluster
RENO DE MEDICI	Materiali	-1,957	9,146	-26,022	1	2	1	0,36	0,64	0,00	2	2
SABAF	Beni Voluttuari	0,741	6,923	-18,116	2	2	2	0,99	0,01	0,00	1	1
SAIPEM	Energia	1,450	9,147	-24,708	2	2	1	0,43	0,10	0,46	3	1
SARAS	Energia	-2,028	8,216	-17,107	1	2	2	0,99	0,01	0,00	1	1
SEAT PAGINE GIALLE	Beni Voluttuari	-2,200	12,324	-36,660	1	1	1	0,00	1,00	0,00	2	2
SNAI	Beni Voluttuari	0,630	19,649	-36,686	2	1	1	0,00	0,98	0,02	2	2
SNAM RETE GAS	Pubblica Utilità	0,707	3,785	-9,951	2	2	2	0,98	0,00	0,02	1	1
SOCOTHERM	Energia	-0,651	11,446	-33,678	1	1	1	0,00	0,98	0,02	2	2
SOGEFI	Beni Voluttuari	0,063	8,850	-26,092	2	2	1	0,56	0,43	0,01	1	2
SOL	Materiali	0,538	6,826	-15,435	2	2	2	1,00	0,00	0,00	1	1
SOPAF	Finanza	-0,790	14,049	-35,276	1	1	1	0,00	0,98	0,02	2	2
TENARIS	Energia	2,754	12,142	-31,738	2	1	1	0,00	0,21	0,79	3	2
TERNA	Pubblica Utilità	0,579	3,904	-10,552	2	2	2	0,98	0,00	0,02	1	1
TOD'S	Beni Voluttuari	-0,044	8,561	-20,157	2	2	2	0,99	0,01	0,00	1	1
UBI BANCA	Finanza	-0,232	5,219	-15,081	2	2	2	0,96	0,00	0,04	1	1
UNI LAND	Finanza	0,285	18,894	-32,261	2	1	1	0,00	0,16	0,84	3	2
UNICREDIT	Finanza	-0,614	7,137	-21,583	1	2	2	0,99	0,01	0,00	1	1
UNIPOL	Finanza	-0,446	5,717	-16,603	1	2	2	0,99	0,01	0,00	1	1
VITTORIA ASSICURAZIONI	Finanza	0,830	7,507	-15,409	2	2	2	0,96	0,00	0,04	1	1
ZUCCHI	Beni Voluttuari	-1,626	8,144	-32,343	1	2	1	0,14	0,86	0,00	2	2

Legenda:

\bar{R} – media dei rendimenti mensili

σ – scarto quadratico medio dei rendimenti mensili

τ_1 – primo percentile dei rendimenti mensili

M – indicatore categorico dicotomico per la media

S – indicatore categorico dicotomico per lo scarto quadratico medio

P – indicatore categorico dicotomico per il primo percentile

CL1 – stima della probabilità a posteriori di appartenenza alla Classe 1

CL2 – stima della probabilità a posteriori di appartenenza alla Classe 2

CL3 – stima della probabilità a posteriori di appartenenza alla Classe 3

CL# – classe modale

Cluster – classificazione con il metodo delle k-medie: appartenenza al cluster

Appendice C – Mercato azionario italiano: la classificazione dei titoli azionari ottenuta dall’analisi a profilo latente

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
A2A	S&P-MIB	Pubblica Utilità	-0,254	9,453	-17,627	3,048	0,96	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
ACEA	Midex	Pubblica Utilità	0,290	9,486	-21,553	2,954	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
ACEGAS-APS	Standard	Pubblica Utilità	-0,456	7,855	-15,887	2,610	0,03	0,90	0,00	0,00	0,00	0,07	0,00	0,00	0,00	2
ACOTEL GROUP	All-Stars	Informatica	-0,051	16,016	-38,429	4,406	0,00	0,00	0,00	0,98	0,00	0,00	0,02	0,00	0,00	4
ACQUE POTABILI	Standard	Pubblica Utilità	-0,121	12,883	-24,738	3,375	0,00	0,00	0,96	0,00	0,00	0,00	0,00	0,00	0,03	3
ACSM-AGAM	Standard	Pubblica Utilità	-1,182	9,318	-20,634	3,116	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
ACTELIOS	All-Stars	Pubblica Utilità	3,727	33,255	-24,211	3,454	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	9
AEROPORTO DI FIRENZE	Standard	Beni Voluttuari	0,237	8,612	-20,769	3,079	0,93	0,06	0,01	0,00	0,00	0,00	0,00	0,00	0,00	1
ALERION INDUSTRIES	Standard	Finanza	0,945	13,328	-20,069	3,092	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	5
ALLEANZA	S&P-MIB	Finanza	-0,144	7,682	-19,332	3,270	0,01	0,99	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
AMPLIFON	All-Stars	Sanità	-0,364	11,857	-35,832	3,747	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	8
ANSALDO STS	All-Stars	Industriali	0,511	5,571	-8,635	3,335	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
ARENA	Standard	Beni Prima Necessità	-2,659	12,896	-30,369	4,673	0,00	0,00	0,00	0,98	0,00	0,00	0,02	0,00	0,00	4
AS ROMA	Standard	Beni Voluttuari	-0,488	18,643	-31,043	3,667	0,00	0,00	0,00	0,87	0,00	0,00	0,13	0,00	0,00	4
ASCOPIAVE	All-Stars	Pubblica Utilità	-1,578	5,258	-9,169	2,363	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
ASTALDI	All-Stars	Industriali	0,896	8,900	-19,912	3,118	0,92	0,05	0,00	0,00	0,03	0,00	0,00	0,00	0,00	1
ATLANTIA	S&P-MIB	Industriali	0,824	6,876	-17,409	2,734	0,00	0,53	0,00	0,00	0,00	0,47	0,00	0,00	0,00	2
AUTOGRILL	S&P-MIB	Beni Voluttuari	-0,238	8,324	-21,947	3,534	0,38	0,13	0,49	0,00	0,00	0,00	0,00	0,00	0,00	3
AUTOSTRADA TO-MI	Midex	Industriali	0,172	8,365	-28,076	2,874	0,88	0,00	0,10	0,00	0,02	0,00	0,00	0,00	0,00	1
AZIMUT HOLDING	Midex	Finanza	0,551	9,603	-23,436	4,593	0,00	0,00	0,02	0,01	0,97	0,00	0,00	0,00	0,00	5
BANCA CARIGE	Midex	Finanza	0,390	4,899	-16,537	3,011	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
BANCA FINNAT	All-Stars	Finanza	1,429	10,384	-19,574	2,551	0,01	0,00	0,00	0,00	0,98	0,00	0,00	0,00	0,00	5

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
BANCA GENERALI	All-Stars	Finanza	-4,139	10,247	-23,157	2,854	0,00	0,00	0,00	0,00	0,02	0,00	0,98	0,00	0,00	7
BANCA IFIS	All-Stars	Finanza	2,072	13,146	-11,820	2,449	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	5
BANCA INTERMOBILIARE	Standard	Finanza	0,620	12,613	-16,658	2,632	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	5
BANCA ITALEASE	Midex	Finanza	-1,666	18,368	-36,042	5,580	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
BANCA MONTE DEI PASCHI	S&P-MIB	Finanza	-0,221	7,554	-18,703	3,045	0,00	0,99	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
BANCA POPOLARE DI MILANO	S&P-MIB	Finanza	0,074	8,577	-18,271	3,751	0,09	0,75	0,01	0,00	0,16	0,00	0,00	0,00	0,00	2
BANCA POPOLARE ETRURIA	All-Stars	Finanza	-0,459	7,084	-16,895	2,215	0,00	0,98	0,00	0,00	0,00	0,02	0,00	0,00	0,00	2
BANCA PPO.DI SPOLETO	Standard	Finanza	-0,231	6,811	-15,629	2,851	0,00	0,92	0,00	0,00	0,01	0,06	0,00	0,00	0,00	2
BANCA PROFILO	Standard	Finanza	0,598	22,269	-26,522	4,507	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,99	9
BANCO DI SARDEGNA RSP	Standard	Finanza	-0,562	7,271	-21,224	2,404	0,00	0,95	0,00	0,00	0,05	0,00	0,00	0,00	0,00	2
BANCO POPOLARE	S&P-MIB	Finanza	-0,371	7,989	-20,453	3,695	0,03	0,88	0,02	0,00	0,06	0,00	0,00	0,00	0,00	2
BASICNET	Standard	Beni Voluttuari	-0,030	16,049	-28,324	4,688	0,00	0,00	0,01	0,98	0,00	0,00	0,01	0,00	0,00	4
BASTOGI	Standard	Finanza	3,255	23,988	-25,139	4,728	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	9
BB BIOTECH (MIL)	All-Stars	Finanza	-0,821	7,601	-19,814	3,268	0,00	0,99	0,00	0,00	0,01	0,00	0,00	0,00	0,00	2
BEGHELLI	Standard	Beni Voluttuari	-0,306	13,566	-28,078	3,130	0,00	0,00	0,98	0,00	0,00	0,00	0,01	0,00	0,00	3
BENETTON	Midex	Beni Voluttuari	-0,622	9,488	-23,347	3,411	0,92	0,00	0,08	0,00	0,00	0,00	0,00	0,00	0,00	1
BENI STABILI	Midex	Finanza	0,781	8,662	-23,023	3,813	0,67	0,00	0,01	0,00	0,32	0,00	0,00	0,00	0,00	1
BIESSE	All-Stars	Industriali	-0,028	12,493	-30,791	3,169	0,00	0,00	0,95	0,00	0,04	0,00	0,00	0,01	0,00	3
BNC.DI DESIO E DELB.	Standard	Finanza	0,693	8,024	-17,549	2,304	0,00	0,89	0,00	0,00	0,10	0,01	0,00	0,00	0,00	2
BOERO BARTOLOMEO	Standard	Industriali	1,030	5,941	-9,526	2,779	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
BOLZONI	All-Stars	Industriali	-1,450	8,731	-20,719	2,821	0,78	0,19	0,00	0,00	0,03	0,00	0,00	0,00	0,00	1
BONIFICHE FERRARESI	Standard	Beni Prima Necessità	1,205	5,712	-14,867	2,539	0,00	0,01	0,00	0,00	0,00	0,99	0,00	0,00	0,00	6
BREMBO	All-Stars	Beni Voluttuari	-0,298	7,757	-17,539	2,744	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
BRIOSCHI	Standard	Finanza	0,947	15,473	-26,510	3,262	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,97	9
BULGARI	S&P-MIB	Beni Voluttuari	-0,021	9,951	-28,764	3,974	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	3
BUONGIORNO	All-Stars	Informatica	-1,704	22,652	-37,147	4,265	0,00	0,00	0,00	0,98	0,00	0,00	0,02	0,00	0,00	4

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
BUZZI UNICEM	S&P-MIB	Materiali	0,474	9,374	-20,378	3,242	0,97	0,02	0,00	0,00	0,01	0,00	0,00	0,00	0,00	1
CAD IT	All-Stars	Informatica	-1,401	8,496	-19,912	3,017	0,92	0,08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
CAIRO COMMUNICATION	All-Stars	Beni Voluttuari	-0,427	11,740	-27,463	3,251	0,00	0,00	0,98	0,00	0,00	0,00	0,02	0,00	0,00	3
CALTAGIRONE	Standard	Finanza	0,655	11,440	-21,200	3,199	0,00	0,00	0,02	0,00	0,98	0,00	0,00	0,00	0,00	5
CALTAGIRONE EDITORE	Standard	Beni Voluttuari	-1,675	6,662	-17,751	2,833	0,00	0,99	0,00	0,00	0,00	0,00	0,01	0,00	0,00	2
CAM-FIN	Standard	Energia	-0,475	11,421	-28,513	3,089	0,00	0,00	0,02	0,00	0,98	0,00	0,00	0,00	0,00	5
CARRARO	Standard	Industriali	0,092	9,809	-22,972	3,590	0,62	0,00	0,29	0,00	0,09	0,00	0,00	0,00	0,00	1
CATTOLICA ASSICURAZIONI	Midex	Finanza	-0,108	6,506	-14,128	2,569	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
CEMBRE	All-Stars	Beni Voluttuari	0,263	6,927	-14,660	2,374	0,00	0,78	0,00	0,00	0,00	0,22	0,00	0,00	0,00	2
CEMENTIR HOLDING	Standard	Materiali	0,574	9,721	-23,501	2,921	0,95	0,00	0,00	0,00	0,05	0,00	0,00	0,00	0,00	1
CENTRALE DEL LATTE DI TRO.	All-Stars	Beni Prima Necessità	-0,678	8,772	-19,250	2,818	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
CICCOLELLA	Standard	Finanza	1,205	16,680	-26,962	3,648	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	9
CIR	Midex	Industriali	-0,044	15,562	-32,522	3,905	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	8
CLASS EDITORI	Standard	Beni Voluttuari	-1,121	15,324	-35,247	4,624	0,00	0,00	0,00	0,99	0,00	0,00	0,01	0,00	0,00	4
COBRA	All-Stars	Beni Voluttuari	-4,042	12,565	-29,396	3,667	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	7
COFIDE	Standard	Industriali	1,201	25,241	-36,887	3,440	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	8
CREDITO ARTIGIANO	Standard	Finanza	-0,092	5,234	-12,163	1,943	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
CREDITO EMILIANO	Midex	Finanza	0,311	9,057	-20,448	3,595	0,94	0,00	0,00	0,00	0,06	0,00	0,00	0,00	0,00	1
CREDITO VALTELLINES	Midex	Finanza	0,215	5,578	-9,774	2,346	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
CRESPI	Standard	Materiali	-0,904	8,307	-19,688	2,875	0,48	0,50	0,00	0,00	0,02	0,00	0,00	0,00	0,00	2
CSP INTERNATIONAL	Standard	Beni Voluttuari	-1,190	9,547	-20,554	2,959	0,98	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	1
DADA	All-Stars	Beni Voluttuari	-0,368	16,469	-33,678	4,065	0,00	0,00	0,00	0,91	0,00	0,00	0,09	0,00	0,00	4
DANIELI	Standard	Industriali	1,078	12,688	-34,302	3,675	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,99	0,00	8
DATALOGIC	All-Stars	Informatica	0,247	7,019	-20,903	2,907	0,02	0,98	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
DAVIDE CAMPARI	Midex	Beni Prima Necessità	0,750	6,550	-15,742	2,965	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
DE LONGHI	Standard	Beni Voluttuari	-0,453	9,913	-20,304	3,016	0,92	0,00	0,08	0,00	0,00	0,00	0,00	0,00	0,00	1

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
DEA CAPITAL	All-Stars	Finanza	-1,527	13,388	-29,078	3,841	0,00	0,00	0,01	0,51	0,07	0,00	0,40	0,00	0,00	4
DIGITAL BROS	All-Stars	Beni Voluttuari	-1,135	13,608	-23,557	5,591	0,00	0,00	0,00	0,98	0,00	0,00	0,01	0,00	0,00	4
DMAIL GROUP	All-Stars	Beni Voluttuari	-0,683	13,688	-31,012	4,304	0,00	0,00	0,07	0,85	0,00	0,00	0,08	0,00	0,00	4
DMT	All-Stars	Informatica	0,046	13,749	-31,031	4,694	0,00	0,00	0,00	0,99	0,00	0,00	0,00	0,00	0,00	4
EDISON	Midex	Pubblica Utilità	0,486	8,498	-18,522	3,418	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
EEMS ITALIA	All-Stars	Informatica	-5,901	13,261	-34,762	4,509	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	7
EL EN	All-Stars	Beni Voluttuari	-0,119	8,530	-17,850	3,243	0,21	0,78	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
ELICA	All-Stars	Beni Voluttuari	-6,204	14,208	-39,225	4,252	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	7
EMAK	All-Stars	Beni Voluttuari	0,886	6,126	-14,564	1,882	0,00	0,03	0,00	0,00	0,00	0,97	0,00	0,00	0,00	6
ENEL	S&P-MIB	Pubblica Utilità	-0,366	5,262	-17,167	3,265	0,00	0,01	0,00	0,00	0,00	0,99	0,00	0,00	0,00	6
ENGINEERING I I	All-Stars	Informatica	-0,571	9,491	-23,456	2,793	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
ENI	S&P-MIB	Energia	0,563	5,503	-13,016	3,780	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
ERG	Midex	Energia	1,446	9,275	-18,100	3,185	0,82	0,00	0,00	0,00	0,18	0,00	0,00	0,00	0,00	1
ERG RENEW	Standard	Pubblica Utilità	0,028	11,433	-25,828	3,229	0,00	0,00	0,96	0,00	0,00	0,00	0,00	0,00	0,03	3
ERGO PREVIDENZA	Standard	Finanza	0,250	10,738	-30,715	3,408	0,00	0,00	0,20	0,01	0,77	0,00	0,01	0,00	0,00	5
ESPRINET	All-Stars	Informatica	1,726	12,276	-27,546	3,603	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,98	9
EUROTECH	All-Stars	Informatica	-0,276	16,376	-25,834	3,775	0,00	0,00	0,00	0,30	0,00	0,00	0,08	0,00	0,61	9
EVEREL GROUP	Standard	Beni Voluttuari	-2,978	16,061	-34,665	4,362	0,00	0,00	0,00	0,41	0,00	0,00	0,59	0,00	0,00	7
EXPRIVIA	All-Stars	Informatica	-1,371	16,680	-31,959	4,877	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
FASTWEB	S&P-MIB	Telecomunicazione	-1,200	14,795	-33,960	4,311	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
FIAT	S&P-MIB	Beni Voluttuari	-0,986	10,215	-27,513	3,733	0,00	0,00	0,99	0,00	0,00	0,00	0,00	0,00	0,00	3
FIDIA	All-Stars	Industriali	-0,842	13,041	-24,704	4,459	0,00	0,00	0,89	0,00	0,09	0,00	0,00	0,00	0,01	3
FIERA MILANO	All-Stars	Industriali	-0,466	8,140	-17,613	3,406	0,14	0,84	0,00	0,00	0,01	0,00	0,00	0,00	0,00	2
FILATURA DI POLLONE	Standard	Beni Voluttuari	-0,775	12,147	-27,087	3,922	0,00	0,00	0,95	0,03	0,00	0,00	0,02	0,00	0,00	3
FINARTE-SEMENZATO CASA D'ASTE	Standard	Beni Voluttuari	-2,127	11,420	-22,691	3,653	0,00	0,00	0,00	0,01	0,00	0,00	0,98	0,00	0,00	7
FINMECCANICA	S&P-MIB	Industriali	-0,113	9,696	-26,128	3,489	0,08	0,00	0,91	0,00	0,01	0,00	0,00	0,00	0,00	3

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
FONDIARIA-SAI	S&P-MIB	Finanza	0,064	9,187	-24,426	3,394	0,51	0,00	0,37	0,00	0,11	0,00	0,00	0,00	0,00	1
GABETTI	Standard	Finanza	0,217	11,860	-26,469	3,070	0,00	0,00	0,38	0,00	0,55	0,00	0,00	0,00	0,07	5
GEFRAN	All-Stars	Beni Voluttuari	0,028	7,588	-14,953	2,655	0,00	0,98	0,00	0,00	0,00	0,02	0,00	0,00	0,00	2
GEMINA	Midex	Industriali	0,277	13,180	-35,646	3,868	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	8
GENERALI	S&P-MIB	Finanza	-0,044	6,952	-20,862	2,973	0,00	0,99	0,00	0,00	0,00	0,01	0,00	0,00	0,00	2
GEOX	S&P-MIB	Beni Voluttuari	0,095	10,200	-24,490	3,462	0,04	0,00	0,96	0,00	0,00	0,00	0,00	0,00	0,00	3
GEWISS	Standard	Industriali	-0,304	9,323	-25,927	3,243	0,70	0,00	0,28	0,00	0,03	0,00	0,00	0,00	0,00	1
GRANDI VIAGGI	Standard	Beni Voluttuari	0,088	11,634	-23,995	3,973	0,00	0,00	0,99	0,00	0,00	0,00	0,01	0,00	0,00	3
GRANITIFIANDRE	All-Stars	Industriali	-0,688	7,106	-21,686	2,454	0,00	0,98	0,00	0,00	0,01	0,01	0,00	0,00	0,00	2
GRUPPO CERAMICHE RICCHET	Standard	Industriali	-1,026	8,404	-18,662	2,722	0,51	0,45	0,00	0,00	0,05	0,00	0,00	0,00	0,00	1
GRUPPO COIN	Standard	Beni Voluttuari	-0,269	12,875	-28,812	3,101	0,00	0,00	0,99	0,00	0,00	0,00	0,01	0,00	0,00	3
GRUPPO EDIT.L'ESPRESSO	S&P-MIB	Beni Voluttuari	-0,859	14,386	-24,509	3,833	0,00	0,00	0,99	0,01	0,00	0,00	0,00	0,00	0,00	3
HERA	Midex	Pubblica Utilità	0,510	6,301	-16,395	3,959	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
I VIAGGI DEL VENTAGLIO	Standard	Beni Voluttuari	-2,118	13,988	-31,957	3,476	0,00	0,00	0,00	0,17	0,00	0,00	0,83	0,00	0,00	7
IFI PV	Midex	Finanza	-0,502	11,830	-35,144	4,643	0,00	0,00	0,00	0,98	0,01	0,00	0,00	0,00	0,00	4
IFIL	Midex	Finanza	-0,607	10,716	-27,756	4,369	0,00	0,00	0,33	0,36	0,31	0,00	0,00	0,00	0,00	4
IMA	All-Stars	Industriali	0,695	5,244	-9,990	1,685	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
IMMOBILIARE GRANDE DISTRIBUZIONE	All-Stars	Finanza	-0,441	10,595	-27,586	5,165	0,00	0,00	0,02	0,50	0,47	0,00	0,00	0,00	0,00	4
IMMSI	Standard	Beni Voluttuari	-0,504	10,553	-19,974	4,174	0,00	0,00	0,95	0,00	0,01	0,00	0,04	0,00	0,00	3
IMPREGILO	S&P-MIB	Industriali	0,475	11,983	-27,698	3,836	0,00	0,00	0,98	0,00	0,02	0,00	0,00	0,00	0,00	3
INDESIT COMPANY	Midex	Beni Voluttuari	0,446	9,009	-20,499	3,332	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
INTEK	Standard	Finanza	-0,266	9,105	-21,427	3,153	0,85	0,02	0,01	0,00	0,12	0,00	0,00	0,00	0,00	1
INTERPUMP GROUP	All-Stars	Industriali	0,411	7,005	-12,435	2,417	0,00	0,02	0,00	0,00	0,00	0,98	0,00	0,00	0,00	6
INTESA SANPAOLO	S&P-MIB	Finanza	0,051	9,660	-29,109	4,629	0,00	0,00	0,36	0,16	0,48	0,00	0,00	0,00	0,00	5
INVEST E SVILUPPO	Standard	Finanza	-1,558	9,522	-21,738	4,242	0,00	0,00	0,00	0,00	0,97	0,00	0,02	0,00	0,00	5
IPI	Standard	Finanza	-0,092	9,097	-19,609	3,363	0,77	0,05	0,00	0,00	0,18	0,00	0,00	0,00	0,00	1

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
IRCE	All-Stars	Beni Voluttuari	-0,565	7,167	-20,314	2,464	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
IRIDE	Midex	Pubblica Utilità	-0,779	8,462	-23,123	3,064	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
ISAGRO	All-Stars	Industriali	0,684	16,017	-35,737	3,683	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	8
IT HOLDING	All-Stars	Beni Voluttuari	-1,278	12,744	-31,176	5,345	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
IT WAY	All-Stars	Informatica	-0,800	9,539	-25,117	3,463	0,98	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	1
ITALCEMENTI	S&P-MIB	Materiali	0,028	7,204	-17,579	3,326	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
ITALMOBILIARE	Midex	Materiali	0,558	8,230	-27,096	3,221	0,99	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	1
JUVENTUS FOOTBALL CLUB	All-Stars	Beni Voluttuari	-0,882	12,462	-27,207	2,693	0,00	0,00	0,94	0,00	0,00	0,00	0,06	0,00	0,00	3
KINEXIA	Standard	Materiali	-0,295	11,871	-28,751	4,748	0,00	0,00	0,01	0,49	0,50	0,00	0,00	0,00	0,00	5
KME GROUP	Standard	Materiali	-1,259	10,670	-22,304	2,217	0,01	0,00	0,00	0,00	0,99	0,00	0,00	0,00	0,00	5
LA DORIA	All-Stars	Beni Prima Necessità	-0,590	7,394	-18,946	2,416	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
LAVORWASH	Standard	Beni Voluttuari	-0,637	10,206	-22,222	3,179	0,59	0,00	0,40	0,00	0,00	0,00	0,00	0,00	0,00	1
LAZIO	Standard	Beni Voluttuari	-2,377	20,390	-45,993	3,888	0,00	0,00	0,00	0,02	0,00	0,00	0,98	0,00	0,00	7
LOTTOMATICA	S&P-MIB	Beni Voluttuari	1,321	7,025	-14,353	2,904	0,00	0,36	0,00	0,00	0,00	0,64	0,00	0,00	0,00	6
LUXOTTICA	S&P-MIB	Beni Voluttuari	0,005	7,642	-17,219	3,281	0,01	0,99	0,00	0,00	0,00	0,01	0,00	0,00	0,00	2
MARCOLIN	Standard	Beni Voluttuari	0,090	12,329	-23,181	2,480	0,00	0,00	0,99	0,00	0,00	0,00	0,00	0,00	0,00	3
MARIELLA BURANI	All-Stars	Beni Voluttuari	0,647	7,069	-17,355	2,096	0,00	0,94	0,00	0,00	0,00	0,06	0,00	0,00	0,00	2
MARR	All-Stars	Beni Prima Necessità	-0,261	6,260	-15,157	2,278	0,00	0,40	0,00	0,00	0,00	0,60	0,00	0,00	0,00	6
MEDIASET	S&P-MIB	Beni Voluttuari	-0,441	10,508	-22,489	3,332	0,04	0,00	0,96	0,00	0,00	0,00	0,00	0,00	0,00	3
MEDIOBANCA	S&P-MIB	Finanza	0,030	7,493	-18,637	2,748	0,00	0,99	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
MEDIOLANUM	S&P-MIB	Finanza	-0,128	13,061	-28,908	4,228	0,00	0,00	0,78	0,14	0,07	0,00	0,00	0,00	0,00	3
MEDITERRANEA DELLE ACQUE	Standard	Pubblica Utilità	0,235	10,052	-18,406	2,965	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
MILANO ASSICURAZIONI	Midex	Finanza	0,200	8,688	-21,680	3,106	0,98	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00	1
MIRATO	All-Stars	Materiali	0,078	6,727	-17,171	2,580	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
MITTEL	Standard	Finanza	0,703	9,134	-16,700	3,439	0,47	0,05	0,00	0,00	0,48	0,00	0,00	0,00	0,00	5
MONDADORI EDITORE	S&P-MIB	Beni Voluttuari	-0,260	12,346	-21,100	3,461	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	3

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
MONDO TV	All-Stars	Beni Voluttuari	-2,272	9,912	-28,808	3,857	0,00	0,00	0,00	0,06	0,00	0,00	0,94	0,00	0,00	7
MONRIF	Standard	Finanza	-0,667	10,776	-23,913	3,109	0,06	0,00	0,34	0,00	0,59	0,00	0,00	0,00	0,00	5
MONTEFIBRE	Standard	Materiali	-0,417	13,919	-32,415	3,716	0,00	0,00	0,00	0,07	0,01	0,00	0,00	0,91	0,00	8
NEGRI BOSSI	All-Stars	Industriali	-0,651	9,882	-27,224	2,167	0,00	0,00	0,66	0,00	0,34	0,00	0,00	0,00	0,00	3
NICE	All-Stars	Informatica	-3,616	9,668	-31,763	2,997	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	7
OLIDATA	Standard	Beni Voluttuari	-1,723	14,301	-27,825	5,517	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
PANARIA GROUP	All-Stars	Materiali	-1,593	6,871	-18,405	3,136	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
PARMALAT	S&P-MIB	Beni Prima Necessità	-1,475	7,768	-17,328	3,386	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
PERMASTEELISA	Standard	Industriali	0,637	9,529	-20,725	3,282	0,96	0,00	0,00	0,00	0,03	0,00	0,00	0,00	0,00	1
PININFARINA	Standard	Beni Voluttuari	-1,598	11,571	-34,779	4,796	0,00	0,00	0,00	0,97	0,00	0,00	0,03	0,00	0,00	4
PIRELLI	S&P-MIB	Beni Voluttuari	-0,432	9,323	-27,398	3,440	0,05	0,00	0,95	0,00	0,00	0,00	0,00	0,00	0,00	3
PIRELLI REAL ESTATE	Midex	Finanza	-1,481	11,053	-35,768	4,031	0,00	0,00	0,00	0,95	0,03	0,00	0,00	0,01	0,00	4
POLIGRAFICA S F	All-Stars	Beni Voluttuari	-1,310	16,860	-21,794	3,999	0,00	0,00	0,00	0,11	0,00	0,00	0,88	0,00	0,00	7
POLIGRAFICI EDITORIALE	Standard	Beni Voluttuari	-1,116	11,878	-25,290	3,108	0,00	0,00	0,96	0,00	0,00	0,00	0,04	0,00	0,00	3
POLTRONA FRAU	All-Stars	Beni Voluttuari	-4,870	8,443	-25,397	3,004	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	7
PREMAFIN-HLDG.DI PART.	Standard	Finanza	0,532	8,913	-14,522	3,516	0,22	0,06	0,00	0,00	0,72	0,00	0,00	0,00	0,00	5
PREMUDA	Standard	Energia	1,660	10,186	-16,004	3,133	0,03	0,00	0,00	0,00	0,97	0,00	0,00	0,00	0,00	5
PRIMA INDUSTRIE	All-Stars	Industriali	0,414	12,600	-28,520	3,391	0,00	0,00	0,92	0,00	0,06	0,00	0,00	0,00	0,01	3
RATTI	Standard	Beni Voluttuari	-1,357	9,541	-16,949	3,593	0,96	0,01	0,00	0,00	0,00	0,00	0,03	0,00	0,00	1
RCS MEDIAGROUP	Midex	Beni Voluttuari	-0,085	13,330	-26,776	3,602	0,00	0,00	0,99	0,01	0,00	0,00	0,00	0,00	0,00	3
RECORDATI	Midex	Sanità	1,750	10,076	-19,379	3,165	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
RENO DE MEDICI	All-Stars	Materiali	-1,957	9,146	-26,022	3,177	0,96	0,00	0,00	0,00	0,04	0,00	0,00	0,00	0,00	1
REPLY	All-Stars	Informatica	0,638	9,491	-16,852	2,686	0,96	0,04	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
SABAF	All-Stars	Beni Voluttuari	0,741	6,923	-18,116	2,276	0,00	0,96	0,00	0,00	0,00	0,04	0,00	0,00	0,00	2
SADI SERVIZI INDUSTRIALI	Standard	Industriali	-0,421	11,328	-19,678	2,250	0,00	0,00	0,12	0,00	0,88	0,00	0,00	0,00	0,00	5
SAES GETTERS	All-Stars	Informatica	-0,171	9,870	-27,212	3,202	0,97	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,01	1

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
SAES GETTERS RNC	All-Stars	Informatica	0,523	10,159	-24,384	3,449	0,96	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,03	1
SAIPEM	S&P-MIB	Energia	1,450	9,147	-24,708	4,426	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	5
SARAS	Midex	Energia	-2,028	8,216	-17,107	4,002	0,04	0,00	0,00	0,00	0,96	0,00	0,00	0,00	0,00	5
SEAT PAGINE GIALLE	S&P-MIB	Beni Voluttuari	-2,200	12,324	-36,660	4,807	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
SIAS	Midex	Industriali	0,611	7,084	-16,748	2,515	0,00	0,01	0,00	0,00	0,00	0,99	0,00	0,00	0,00	6
SNAI	Standard	Beni Voluttuari	0,630	19,649	-36,686	5,139	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
SNAM RETE GAS	S&P-MIB	Pubblica Utilità	0,707	3,785	-9,951	1,621	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
SNIA ORD	Standard	Industriali	-1,232	12,373	-37,239	3,692	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	8
SOCOTHERM	All-Stars	Energia	-0,651	11,446	-33,678	4,935	0,00	0,00	0,00	0,91	0,09	0,00	0,00	0,00	0,00	4
SOGEFI	All-Stars	Beni Voluttuari	0,063	8,850	-26,092	3,623	0,22	0,00	0,76	0,00	0,00	0,00	0,02	0,00	0,00	3
SOL	Standard	Materiali	0,538	6,826	-15,435	2,895	0,00	0,99	0,00	0,00	0,01	0,00	0,00	0,00	0,00	2
SOPAF	Standard	Finanza	-0,790	14,049	-35,276	4,185	0,00	0,00	0,00	0,96	0,01	0,00	0,03	0,00	0,00	4
SORIN	Standard	Sanità	-2,298	8,825	-23,748	3,577	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
STMICROELECTRONICS (MIL)	S&P-MIB	Informatica	-1,125	13,372	-32,476	4,508	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
TELECOM ITALIA	S&P-MIB	Telecomunicazione	-0,619	10,388	-24,728	3,812	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	3
TENARIS	S&P-MIB	Energia	2,754	12,142	-31,738	5,880	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	5
TERNA	S&P-MIB	Pubblica Utilità	0,579	3,904	-10,552	2,758	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	6
TISCALI	Midex	Informatica	-0,797	22,636	-33,846	5,411	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
TOD'S	Midex	Beni Voluttuari	-0,044	8,561	-20,157	3,103	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1
TREVI FIN INDUSTRIALE	Standard	Industriali	2,094	14,881	-25,795	3,665	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,99	9
TREVISAN COMETAL	All-Stars	Industriali	-1,146	13,370	-39,256	4,129	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,99	0,00	8
TXT E-SOLUTION	All-Stars	Informatica	-1,120	19,128	-29,175	5,614	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	4
UBI BANCA	S&P-MIB	Finanza	-0,232	5,219	-15,081	3,357	0,00	0,02	0,00	0,00	0,00	0,98	0,00	0,00	0,00	6
UNI LAND	All-Stars	Finanza	0,285	18,894	-32,261	4,440	0,00	0,00	0,00	0,99	0,00	0,00	0,01	0,00	0,00	4
UNICREDIT	S&P-MIB	Finanza	-0,614	7,137	-21,583	4,760	0,00	0,01	0,04	0,01	0,95	0,00	0,00	0,00	0,00	5
UNIPOL	S&P-MIB	Finanza	-0,446	5,717	-16,603	2,303	0,00	0,29	0,00	0,00	0,00	0,71	0,00	0,00	0,00	6

Azione	Indice	Settore GICS	\bar{R}	σ	τ_1	σ_C	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL#
VIANINI INDR.	Standard	Industriali	0,494	11,189	-13,902	2,268	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	5
VIANINI LAVORI	Standard	Industriali	0,649	8,367	-22,612	2,708	0,85	0,03	0,01	0,00	0,11	0,00	0,00	0,00	0,00	1
VITTORIA ASSICURAZIONI	All-Stars	Finanza	0,830	7,507	-15,409	2,762	0,00	0,94	0,00	0,00	0,01	0,05	0,00	0,00	0,00	2
ZUCCHI	Standard	Beni Voluttuari	-1,626	8,144	-32,343	3,102	0,02	0,00	0,00	0,03	0,00	0,00	0,94	0,00	0,00	7

Legenda:

\bar{R} – media dei rendimenti mensili

σ – scarto quadratico medio dei rendimenti mensili

τ_1 – primo percentile dei rendimenti mensili

σ_C – scarto quadratico medio dei rendimenti giornalieri nei periodi di crisi (set-ott 2001, lug-ott 2002 e set-dic 2008)

CLx – stima della probabilità a posteriori di appartenenza alla Classe x (x = 1, ..., 9)

CL# – classe modale

