

Capitolo 2

Modelli Latent Markov per dati longitudinali: utilizzo in ambito econometrico e possibili estensioni

Francesco Dotto

Abstract

Our contribution is focused on Latent Markov Models (LM) and some empirical applications in the econometrics context. In particular, we introduce the mathematical and statistical basic notions needed for both interpreting the output of a LM model and understanding the role of the parameters of the model. Furthermore, we present the issues that may arise whenever covariates (discrete or continuous) are inserted within the model. Finally, we present an empirical econometric application tailored at facing the problem of estimating material deprivation.

Keywords: Latent Variables, Markov Chain, DIF, covariates.

2.1 I Modelli a Classi Latenti: introduzione e motivazione

Con il termine ‘Latent Variable Models’, ovvero modelli a Variabili Latenti, si fa riferimento a un insieme di metodologie atte a stimare un tratto latente sottostante i dati, ovvero una variabile non direttamente osservabile ma che viene stimata a partire dall’insieme delle variabili manifeste. Un esempio di variabile latente è ‘la bravura’ o ‘l’intelligenza’ di un campione di studenti che può essere ricostruita a partire dalle loro valutazioni nei diversi corsi universitari, oppure lo status di ricchezza/povertà di un insieme di famiglie sulla base di un set di indicatori. Concettualmente parlando, una variabile latente è quindi un costrutto ipotetico per il quale non esiste un corrispettivo empirico e che viene stimato sulla base di un insieme di variabili osservate. A seconda della natura dei dati e del tratto latente si possono individuare quattro macrocategorie di modelli, le cui caratteristiche sono brevemente sintetizzate nella tabella sottostante (si rimanda, per una trattazione approfondita, a Collins e Lanza, 2009).

Tabella 2.1 – La nomenclatura dei modelli a Variabili Latenti al variare della natura delle variabili osservate e delle variabili latenti

| Variabili Osservate/Latenti | Continue | Discrete |
|-----------------------------|----------------------|--------------------------|
| Continue | Analisi Fattoriale | Modelli a mistura Finita |
| Discrete | Item response Theory | Modelli a Classi Latenti |

Fonte: Collins e Lanza (2009)

L’uso delle variabili latenti trova applicazione in diversi tipi di analisi in quanto permette di:

- 1) Eliminare eterogeneità non osservabile che comporta distorsione nella stima dei parametri stimati (siano essi semplici statistiche descrittive o parametri associati a modelli statistici più sofisticati).
- 2) Analizzare separatamente i dati in raggruppamenti massimamente omogenei al loro interno e massimamente eterogenei tra di loro.

Esistono innumerevoli esempi che hanno dimostrato come ignorare la presenza di una variabile raggruppamento sottostante i dati possa portare a formulare conclusioni drasticamente lontane dalla realtà o addirittura paradossali. In ambito metodologico si definisce ‘paradosso di Simpson’ quella situazione in cui una relazione tra due fenomeni appare modificata, o persino invertita, dai dati in possesso a causa di altri fenomeni non presi in considerazione.

A scopi illustrativi si riportano, nella Tabella 2.2, i tassi di mortalità

relativi alla Variante Delta del Covid-19 in relazione alle vaccinazioni. Da una prima analisi dei dati riportati nella Tabella 2.2 risulta che il tasso di mortalità per i pazienti vaccinati è nettamente maggiore del tasso di mortalità rilevato nei pazienti non vaccinati (circa tre volte maggiore). Tuttavia, stratificando opportunamente le unità statistiche si giunge a conclusioni diametralmente opposte. Separando, infatti, i pazienti over 50 (etichettati come ‘50+’) dai pazienti aventi meno di cinquanta anni (etichettati come ‘U 50’) le conclusioni cambiano drasticamente. La spiegazione di tale fenomeno apparentemente paradossale è che la variabile ‘vaccinazione’ presenta un forte grado di associazione con la variabile ‘età’. La vaccinazione, infatti, è più diffusa nei pazienti più anziani i quali, naturalmente, sperimentano una mortalità più elevata dei pazienti giovani. In conclusione, la variabile ‘età’ è un fattore che maschera l’effetto della vaccinazione e pertanto, occorre tenerne conto al fine di poter valutare correttamente l’effetto della vaccinazione sui pazienti.

Tabella 2.2 – Tassi di Mortalità $\times 1.000$ in relazione all’età e allo status vaccinale

| Status Vaccinale | Mortalità x 1.000 |
|--------------------|-------------------|
| Vaccinati | 4.11 |
| Non Vaccinati | 1.67 |
| Vaccinati U 50 | 0.23 |
| Non Vaccinati U 50 | 0.33 |
| Vaccinati 50+ | 16.84 |
| Non Vaccinati 50+ | 59.59 |

Fonte: elaborazione dell’autore su dati reperibili da
https://www.openintro.org/data/index.php?data=simpsons_paradox_covid.

Nell’esempio di cui sopra il fattore di confondimento (l’età) è facilmente osservabile ma nelle applicazioni empiriche questo non sempre si verifica. Pertanto, la stima dei fattori latenti nelle applicazioni empiriche si rivela di fondamentale importanza poiché una variabile latente può agire da fattore di confondimento sulla variabile di interesse. Tuttavia, poiché per stimare i parametri di un modello a variabili latenti occorrono procedure computazionalmente onerose, il loro utilizzo è relativamente recente. Ad oggi questo tipo di modelli trovano la loro implementazione nei più comuni software di analisi dei dati come Stata, R, Python e MatLab. Per una revisione, seppur non esaustiva, dell’utilizzo di tali metodologie in ambito econometrico si invita il lettore alla consultazione di Brzezińska (2016).

2.2 I modelli a classi latenti per dati longitudinali: i modelli Latent Markov

2.2.1 Introduzione e strumenti metodologici

I modelli Latent Markov (Wiggins 1973) costituiscono uno strumento di analisi atto a misurare l'evoluzione di un tratto latente sottostante i dati nell'evolversi del tempo, e si possono quindi classificare come un'estensione dei modelli a classi latenti per dati longitudinali.

Come osservato in Bartolucci *et. al.* (2014): 'Sotto molti punti di vista, i dati longitudinali sono simili ai dati delle serie storiche. La differenza principale risiede nel fatto che nel contesto delle serie storiche solitamente si ha un numero elevato di osservazioni ripetute nel tempo associate a una singola unità statistica, mentre, nel contesto dei dati longitudinali, si ha un numero generalmente contenuto di osservazioni ripetute nel tempo associate ad un cospicuo numero di unità statistiche'. Di conseguenza, gli approcci inferenziali utilizzati all'interno delle serie storiche non possono essere estesi direttamente ai dati longitudinali. Nell'analisi delle serie storiche, infatti, le proprietà asintotiche di uno stimatore sono studiate assumendo che il numero di misure ripetute tenda all'infinito, mentre, nel contesto dei dati longitudinali, le proprietà asintotiche vengono studiate assumendo che la dimensione del campione tenda ad infinito avendo un numero di osservazioni nel tempo che rimane fissato.

Pertanto, alla luce di quanto detto, i modelli Latent Markov possono essere visti come una generalizzazione di un modello basato su una catena di Markov discreta atta a tenere conto di errori di misurazione nelle variabili manifeste o, in alternativa, come una generalizzazione di un modello a classi latenti per dati longitudinali in cui ogni unità statistica può eventualmente sperimentare una transizione tra le classi latenti.

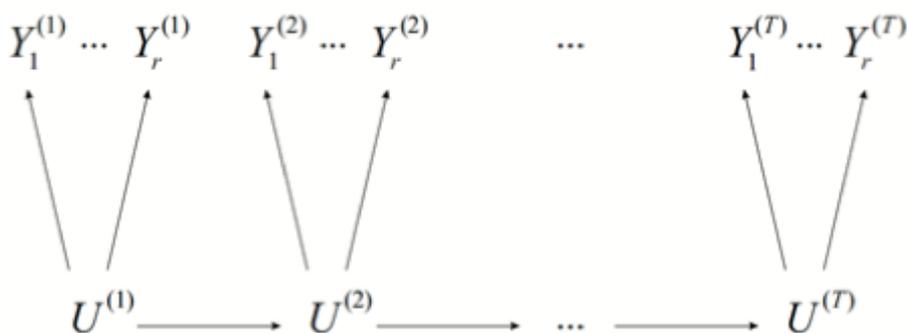
Matematicamente parlando, sia $(Y_{i1}^t, Y_{i2}^t, \dots, Y_{iR}^t)$ un vettore contenente i valori di R variabili risposta categoriche (solitamente dicotomiche) associate ad una generica i -esima unità statistica al tempo t dove n è il numero di osservazioni, R il numero di variabili, T il numero delle occasioni. Al fine modellizzare la distribuzione congiunta delle variabili $Y_{i1}^t, Y_{i2}^t, \dots, Y_{iR}^t$ si assume l'esistenza di una variabile latente U discreta avente uno spazio degli stati composto da k elementi. Si assume inoltre che:

- 1) le variabili risposta $Y_{i1}^t, Y_{i2}^t, \dots, Y_{iR}^t$ siano localmente indipendenti condizionatamente agli stati latenti (*assunzione di indipendenza locale*);
- 2) le realizzazioni della variabile aleatoria U formino una catena di Markov omogenea del primo ordine (*assunzione di Markovianità*).

La prima delle due assunzioni, come verrà specificato nei paragrafi seguenti, ricopre un ruolo cruciale e, come si vedrà in seguito, violazioni di tale asserto implicano che le stime dei parametri del modello siano fortemente distorte e talvolta non consistenti. Da un punto di vista applicativo l'assunzione di indipendenza locale implica che il tratto latente può spiegare completamente (a meno di una componente di errore idiosincratICA) le caratteristiche osservabili di una determinata unità statistica (Figura 2.1).

Figura 2.1

Il grafo associato a un insieme di indicatori aventi nessi di causalità con un processo markoviano latente osservato in T occasioni



Fonte: Bartolucci *et al.* (2012)

Bartolucci *et al.* (2012), riporta il grafo associato a un'ipotetica variabile latente U ed il suo nesso causale con le variabili osservate $Y_{i1}, Y_{i2}, \dots, Y_{ir}$. Si noti come le frecce del grafo partano dalla variabile U al fine di evidenziare la direzionalità del nesso causale che intercorre tra il tratto latente e le variabili osservate.

La seconda assunzione invece riguarda la struttura di dipendenza tra le realizzazioni della variabile latente che segue la distribuzione di una catena di Markov omogenea del primo ordine e questo implica, sulle variabili osservate $Y_{i1}, Y_{i2}, \dots, Y_{iR}$, che la loro distribuzione congiunta è a sua volta caratterizzata da memoria Markoviana, ovvero, in formule:

$$f_{Y^{(t)}|Y^{(1)}, \dots, Y^{(t-1)}}(y^{(t)}|y^{(1)}, \dots, y^{(t-1)}) = f_{Y^{(t)}|Y^{(t-0)}, \dots, Y^{(t-1)}}(y^{(t)}|y^{(t-0)}, \dots, y^{(t-1)}) \quad (2.1)$$

Si osservi come, tramite l'equazione di cui sopra, al variare del

termine o si possano esplicitare diverse forme di dipendenza. Questo, infatti, indica l'ordine della catena e, di conseguenza, quando $o = 0$ si ha indipendenza mentre, nel contesto dei modelli LM, vale la relazione $o = 1$.

Stante le assunzioni di cui sopra sia $\phi_{h|k} = \Pr(Y_{it} = 1 | U_{it} = k)$, ovvero la probabilità che l' h -esimo indicatore associato all' i -esima unità statistica che appartiene al k -esimo stato latente assuma valore pari a 1. L'insieme delle probabilità di cui sopra, comunemente chiamate in letteratura con il nome di *response probabilities*, formano la cosiddetta distribuzione manifesta. Di contro invece, per quanto riguarda la variabile latente U_{it} si assume che questa segua la distribuzione di una catena di Markov del primo ordine non omogenea. La sua distribuzione viene costruita a partire dalle probabilità iniziali definite da $P(U_{i1} = j) = \pi_j$ e dalle probabilità di transizione definite da $P(U_{it} = j | U_{i,t-1} = h) = \pi_{jth}$. Le due quantità descritte entrano direttamente nella funzione di verosimiglianza del modello data da

$$L(\theta) = \prod_{i=1}^n \left(\sum_{U_{i1}=1}^k \sum_{U_{i2}=1}^k \cdots \sum_{U_{iT_i}=1}^k \Pr(U_{i1}) \prod_{t=2}^{T_i} \Pr(U_{it} | U_{i,t-1}) \prod_{t=1}^{T_i} \prod_{r=1}^R \Pr(Y_{itr} | U_{it}) \right) \quad (2.2)$$

dove, nell'equazione (2.2), θ rappresenta il vettore dei parametri ed è pertanto composto dalle probabilità iniziali π_j , dalle probabilità di transizione π_{jth} e dalle response probabilities $\phi_{h|k}$.

Massimizzando la funzione di verosimiglianza si ottengono gli stimatori di massima verosimiglianza delle *response probabilities*, delle probabilità iniziali e di transizione. Tuttavia, la massimizzazione della funzione di verosimiglianza richiede l'utilizzo di metodi di ottimizzazione numerica il che, chiaramente, rende i modelli LM particolarmente onerosi da un punto di vista computazionale. Un possibile approccio è quello delineato in Bartolucci *et al.* (2012) che propone la massimizzazione della funzione di verosimiglianza tramite una opportuna variante dell'algoritmo EM (Dempster *et al.* 1977).

2.2.2 I modelli Latent Markov per la risoluzione di problemi econometrici

Il problema dell'eterogeneità non osservata affligge un cospicuo numero di casistiche di interesse per i modelli econometrici. Per tale ragione l'utilizzo dei modelli LM ha trovato spazio nelle applicazioni dalla natura più disparata: dalla modellizzazione di fenomeni di marketing (Poulsen, 1990) allo studio delle dinamiche legate alla composizione del portafoglio finanziario delle famiglie (Di Mari *et al.*, 2016) allo studio della deprivazione

materiale (Dotto *et al.*, 2019). Le tre applicazioni brevemente menzionate, che costituiscono un prospetto non esaustivo delle applicazioni dei modelli LM ai problemi econometrici, hanno tutte quante un denominatore comune: la necessità dello studio di una variabile non direttamente osservabile tramite le realizzazioni di un insieme di variabili osservate in diversi istanti temporali. Al fine di chiarire in maniera il più possibile esaustiva si entrerà, nel contesto di questo contributo, in alcuni dettagli applicativi della procedura proposta in Dotto *et al.* 2019 atta a misurare lo status non direttamente osservabile di deprivazione materiale.

2.2.2.1 *La deprivazione materiale: che cos'è e come viene quantificata*

Il concetto di deprivazione materiale viene definito in maniera esaustiva da (Townsend 1987) che afferma che una famiglia deve essere classificata come materialmente deprivata quando all'interno di questa ci sia la mancanza di standard 'materiali, nutrizionali, di abbigliamento, di alloggio, di strutture domestiche, di condizioni e strutture lavorative, ambientali e di localizzazione che sono generalmente disponibili nella società in cui i suoi membri vivono'. Pertanto, possiamo definire la deprivazione materiale come un concetto relativo, direttamente collegato con la povertà e che riguarda la collocazione di un individuo o di una famiglia, all'interno della società in cui vive.

Per quanto riguarda la statistica ufficiale il tasso di deprivazione materiale è stato definito dall'Eurostat (Guio *et al.*, 2012) calcolando la percentuale di famiglie che risultano manchevoli in almeno tre dei seguenti $R = 9$ attributi (item) in un determinato anno:

- 1) Possibilità di tenere la propria abitazione sufficientemente calda
- 2) Possibilità di affrontare una settimana di ferie lontano da casa
- 3) Possibilità di fare un pasto proteico (carne pesce o equivalente vegetariano) almeno una volta ogni due giorni
- 4) Possesso di un televisore a colori
- 5) Potersi permettere una lavatrice
- 6) Potersi permettere un'automobile
- 7) Potersi permettere un telefono
- 8) Potersi permettere una macchina
- 9) Non essere in arretrato con il pagamento di bollette, affitto, mutuo o altro tipo di prestito

Inoltre, sempre in accordo con le soglie Eurostat, gli individui che risultano manchevoli in almeno quattro attributi (item) vengono definiti come fortemente deprivati. Lo studio della deprivazione materiale basato

sul conteggio degli item presenta alcune evidenti limitazioni, qui brevemente sintetizzate e studiate in maniera più approfondita in Dotto *et al.* 2019.

In particolare, la scelta di imporre $k = 3$ è del tutto arbitraria. La seconda questione riguarda invece l'importanza relativa degli item. Due famiglie aventi lo stesso score di deprivazione (ovvero lo stesso numero di item in cui sono manchevoli) vengono considerate equivalenti sebbene i diversi item possano avere ragionevolmente un peso differente per individuare lo stato di deprivazione materiale. Per cercare di sopperire a queste due limitazioni, in Dotto *et al.* (2019), viene proposto un approccio basato sui modelli LM. Questo, in primis, evita di dover fissare una soglia sul numero di item e inoltre permette di studiare il fenomeno latente in maniera dinamica, ovvero, nella sua evoluzione nel tempo.

2.2.2.2 *Lo studio della deprivazione materiale tramite i modelli LM*

I dati utilizzati fanno parte dell'indagine EU-SILC del 2013 rilasciata nell'Agosto del 2016 ed hanno una struttura longitudinale che prevede che le misurazioni siano ripetute in $T=4$ anni a partire dal 2010 e comprendono al suo interno famiglie appartenenti a tutti quanti i Paesi dell'Unione Europea. Al fine di snellire la trattazione, in questa sede l'analisi sarà focalizzata su tre Paesi differenti: Italia, Regno Unito e Grecia.

Riprendendo la notazione del paragrafo introduttivo, sia $Y = (Y_{i1}^t, Y_{i2}^t, \dots, Y_{iR}^t)$ il vettore che rappresenta la configurazione dei nove item di deprivazione associati all' i -esima famiglia al tempo t . Prendendo come esempio l'item numero 1, Y_{i1}^t , il suo valore viene determinato sulla base della risposta alla seguente domanda: 'Puoi permetterti di riscaldare la tua abitazione?' Se la risposta è affermativa allora $Y_{i1}^t = 0$, altrimenti $Y_{i1}^t = 1$. Similmente, i rimanenti item sono codificati con la stessa logica: assumono valore pari ad 1 se la famiglia risulta deprivata in quell'item e valore 0 in caso contrario.

Per quanto riguarda invece la variabile latente U_{it} , questa rappresenta lo stato di deprivazione di una famiglia. Essendo la deprivazione materiale uno status non direttamente osservabile occorre imporre, nel modello, il numero degli stati latenti che si vogliono stimare. Operativamente parlando questo implica fissare $k = 2$ stati latenti, e di conseguenza, la corrispettiva variabile latente può assumere valori pari a 1 o al 2 a seconda che una determinata famiglia venga classificata come materialmente non deprivata o materialmente deprivata. Per quanto riguarda invece le realizzazioni di U_{it} al variare di t queste non sono tra di loro indipendenti ma bensì intercorre tra di loro una memoria di tipo Markoviano. Più semplicemente, la

probabilità che una famiglia risulti materialmente deprivata al tempo t dipende dalla probabilità che questa sia stata deprivata al tempo $t - 1$. In termini estremamente pratici, tramite il modello LM si vogliono assegnare le osservazioni $Y_i = (Y_{i1}^t, Y_{i2}^t, \dots, Y_{iR}^t)$ a uno dei due stati latenti U_{it} sulla base delle configurazioni osservate nei diversi item. Nella prossima sezione verranno commentati e interpretati i risultati del modello.

2.2.2.3 I risultati del modello

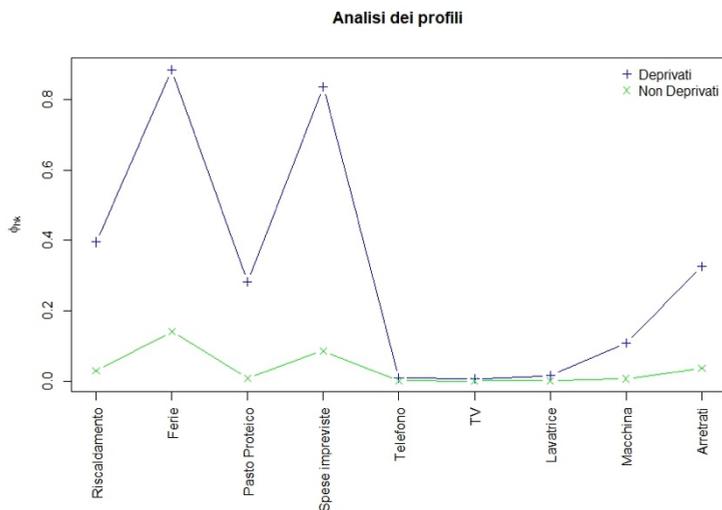
La parte principale dell'output del modello sul quale ci si concentrerà con maggiore attenzione è quella relativa alla *clusterizzazione* delle osservazioni, ovvero la loro suddivisione nei due stati latenti. La Tabella 2.3 riporta la percentuale di famiglie che sono state assegnate allo stato latente di deprivazione materiale nei diversi anni.

La Figura 2.2 riporta invece l'analisi dei profili nei due diversi stati latenti ovvero, in altri termini, le probabilità che i singoli item assumano valore pari ad 1 all'interno dei due stati latenti.

L'analisi dei profili latenti risulta piuttosto chiara: lo stato latente 1 (rappresentato dalla linea verde e dai punti aventi forma 'x') contiene le famiglie non materialmente deprivate: ovvero tutte quelle famiglie la cui probabilità di deprivazione nei singoli item presenta valori particolarmente bassi se non, al limite, vicini a zero. Di contro, lo stato latente 2, contiene le famiglie materialmente deprivate. Al suo interno le probabilità di deprivazione assumono valori generalmente più alti e in alcuni casi vicini al valore 1. Particolarmente significativi, a nostro avviso, risultano i due item relativi alla possibilità di trascorrere una settimana di ferie in un luogo diverso dal luogo di residenza e alla possibilità di affrontare spese impreviste, che sembrano essere i due item che maggiormente caratterizzano le famiglie materialmente deprivate.

Figura 2.2

Analisi dei profili latenti: Probabilità di deprivazione per ogni singolo item nei diversi stati latenti



Fonte: elaborazione dell'autore

Poiché l'analisi dei profili non tiene conto delle differenze tra i diversi Paesi si riporta, nella Tabella 2.3, la percentuale di famiglie materialmente deprivate stratificata per anno e Paese di appartenenza. Queste percentuali sono state ottenute direttamente dall'output del modello dalle probabilità a posteriori di appartenenza agli stati latenti. Come atteso, la Grecia è il Paese con la maggior percentuale di famiglie materialmente deprivate e il Regno Unito, di contro, è il Paese in cui registriamo il minor numero di famiglie deprivate per ogni anno.

Tabella 2.3 – Percentuale di famiglie aventi probabilità maggiore di 0.5 di appartenere allo stato latente delle famiglie materialmente deprivate

| Paese | 2010 | 2011 | 2012 | 2013 |
|-------------|------|------|------|------|
| Grecia | 51.4 | 46.7 | 46.9 | 43.0 |
| Italia | 28.0 | 32.0 | 37.0 | 34.7 |
| Regno Unito | 24.3 | 24.2 | 25.4 | 25.7 |

Fonte: Dotto *et al.* (2019)

I risultati ottenuti con il modello LM sono in una qualche misura comparabili con i risultati ottenuti con l'approccio della Statistica ufficiale, sebbene ci siano delle differenze piuttosto marcate.

Più nello specifico, le stime delle percentuali delle famiglie materialmente deprivate ottenute con il metodo del conteggio degli item, che sono riportate nella Tabella 2.4, risultano significativamente inferiori alle stime fornite dal modello LM.

Tabella 2.4 – *Percentuale di famiglie deprivate al variare della variabile Paese, Anno e Numero di Item*

| Regno Unito: numero di item in cui le famiglie sono deprivate | 2010 | 2011 | 2012 | 2013 |
|--|-------------|-------------|-------------|-------------|
| Uno o più | 38.61 | 40.03 | 40.87 | 40.42 |
| Due o più | 24.76 | 25.55 | 26.72 | 26.78 |
| Tre o più | 13.57 | 12.93 | 15.29 | 16.33 |
| Quattro o più | 7.14 | 5.11 | 6.82 | 7.07 |
| Italia: numero di item in cui le famiglie sono deprivate | 2010 | 2011 | 2012 | 2013 |
| Uno o più | 50.11 | 55.95 | 57.38 | 55.42 |
| Due o più | 30.54 | 36.96 | 40.73 | 37.49 |
| Tre o più | 14.75 | 22.07 | 24.58 | 23.15 |
| Quattro o più | 6.76 | 10.80 | 14.21 | 11.92 |
| Grecia: numero di item in cui le famiglie sono deprivate | 2010 | 2011 | 2012 | 2013 |
| Uno o più | 72.08 | 66.94 | 70.07 | 70.74 |
| Due o più | 53.63 | 49.04 | 50.19 | 48.36 |
| Tre o più | 38.28 | 31.75 | 34.90 | 31.64 |
| Quattro o più | 22.88 | 16.42 | 19.00 | 15.95 |

Fonte: Dotto *et al.* (2019)

Al fine di capire le ragioni di tale discrepanza, nonché ai fini della validazione della clusterizzazione, sono stati fatti alcuni controlli incrociati con variabili quali il reddito disponibile equivalente, la percezione relativa al proprio stato di salute e la capacità percepita di tenere in ordine i conti nella famiglia. In questa sede ci si limita a riportare i risultati relativi al reddito disponibile equivalente relativo all'anno 2013.

Tabella 2.5 – *Matrice di confusione tra la classificazione Eurostat e la classificazione LM e reddito disponibile equivalente*

| Eurostat | Modello LM | | | | | |
|--------------------|------------|-------|--------|-------|-------------|-------|
| | Grecia | | Italia | | Regno Unito | |
| | D | ND | D | ND | D | ND |
| Deprivati (D) | 6031 | 7242 | 12335 | 14912 | 14674 | 16108 |
| Non Deprivati (ND) | - | 11189 | - | 21933 | - | 26346 |

Fonte: Dotto *et al.* (2019)

La Tabella 2.5 rappresenta una variante di una matrice di confusione in cui è stato riportato il reddito medio equivalente delle famiglie a seconda che queste siano state classificate come materialmente deprivate usando i due approcci fin qui commentati: il conteggio degli item e il modello LM.

Come prima cosa è possibile osservare che nessuna famiglia classificata dall'Eurostat come deprivata viene classificata, tramite il modello LM, come non deprivata. In seconda battuta è possibile notare che il reddito medio equivalente delle famiglie non deprivate secondo il criterio Eurostat ma deprivate secondo il modello LM sia significativamente inferiore rispetto al reddito delle famiglie classificate come non materialmente deprivate. Le ulteriori variabili esterne utilizzate confermano quanto detto sopra. Pertanto, è possibile concludere che, adottando il criterio Eurostat, ci sia una tendenza a sottostimare la percentuale di famiglie materialmente deprivate. Una delle possibili cause di questa distorsione dei risultati potrebbe essere dovuta al fatto che tutti gli item di deprivazione, adottando il criterio Eurostat, hanno lo stesso peso nel classificare una famiglia come materialmente deprivata.

La Tabella 2.6 riporta le associazioni tra i diversi item di deprivazione e la variabile latente. Per ogni Paese oggetto di analisi, la prima colonna riporta la probabilità di una famiglia di essere manchevole in un determinato item condizionatamente all'essere assegnata al sottoinsieme delle famiglie deprivate. Tale probabilità, indicata con \widehat{p}_{2r} , misura la sensibilità dell' r -esimo item. Di contro, la quantità data da $1 - \widehat{p}_{1r}$ rappresenta la specificità dell' r -esimo item, ovvero la probabilità che una famiglia ha di non essere manchevole in un determinato item condizionatamente all'essere stata assegnata al sottoinsieme delle famiglie non deprivate. I risultati sono in linea con quelli rappresentati nella Figura 2.2 ma aggiungono qualche spunto in più. In prima battuta osserviamo come il potere

discriminante degli item subisca notevoli variazioni al variare del Paese. Questo evidenzia, di fatto, un'ulteriore criticità dell'approccio utilizzato dalla statistica ufficiale: ovvero quello di adottare lo stesso criterio di classificazione a prescindere dal Paese. In seconda battuta osserviamo come i beni durevoli, quali il telefono, la lavatrice o la televisione siano piuttosto specifici ma ben poco sensibili.

Concludiamo la sezione di commento dei risultati con una breve discussione riguardante il parametro k che rappresenta il numero degli stati latenti. Tale quantità costituisce uno dei parametri di calibrazione del modello e deve essere fissato dal ricercatore. Solitamente, per scegliere un valore opportuno per tale parametro si confrontano uno o più criteri di valutazione dell'adattamento del modello ai dati. Nel caso specifico dell'applicazione proposta si è scelto di utilizzare BIC - *Bayesian Information Criterion* – proposto in Schwarz, 1978 – che ha portato a imporre $k = 2$ stati latenti.

Tabella 2.6 – Sensibilità e specificità degli item di deprivazione

| item | Grecia | | Italia | | Regno Unito | |
|--|--------------------|------------------------|--------------------|------------------------|--------------------|------------------------|
| | \widehat{p}_{2r} | $1 - \widehat{p}_{1r}$ | \widehat{p}_{2r} | $1 - \widehat{p}_{1r}$ | \widehat{p}_{2r} | $1 - \widehat{p}_{1r}$ |
| Riscaldare la propria casa | 49.6 | 92.9 | 43.4 | 98.0 | 21.8 | 98.1 |
| Permettersi una settimana di ferie lontano da casa | 88.9 | 76.0 | 92.4 | 82.4 | 81.0 | 95.7 |
| Fare un pasto proteico | 31.7 | 99.0 | 30.8 | 98.9 | 20.9 | 99.8 |
| Possesso di un televisore a colori | 87.3 | 88.8 | 83.4 | 90.3 | 85.3 | 91.5 |
| Possesso di una lavatrice | 1.2 | 100.0 | 0.8 | 100.0 | 0.2 | 100.0 |
| Possesso di un'automobile | 0.1 | 100.0 | 0.8 | 100.0 | 0.3 | 100.0 |
| Possesso di un telefono | 2.5 | 99.7 | 0.9 | 100.0 | 1.6 | 100.0 |
| Potersi permettere una macchina | 15.5 | 97.6 | 7.9 | 99.8 | 17.9 | 99.2 |
| Essere in regola con i pagamenti | 58.5 | 82.9 | 26.8 | 98.3 | 28.7 | 99.5 |

Fonte: Dotto *et al.* (2019)

2.2.2.4 Modelli LM e deprivazione materiale: alcuni spunti aggiuntivi

La sezione presente riporta una parte dei risultati metodologici riportati in Dotto *et al.* (2019) con alcune considerazioni aggiuntive di stampo prettamente empirico. Si può dimostrare che gli approcci Eurostat e LM, sebbene diano risultati differenti, siano molto collegati. Nell'approccio Eurostat, lo score di deprivazione materiale è dato da una somma non

ponderata di item mentre, tramite nell'approccio LM, lo score di deprivazione è una somma ponderata di item i cui pesi sono funzione *response probabilities* degli item. Per ulteriori approfondimenti si invita il lettore alla consultazione di Dotto *et al.* (2019) e della sua relativa bibliografia. Le sezioni seguenti saranno invece incentrate sulla possibilità di inserire covariate aggiuntive all'interno del modello.

2.3 I modelli LM con le covariate

2.3.1 Introduzione

Nella trattazione fin qui discussa si è approfondita la relazione che intercorre tra un insieme di indicatori e un tratto latente. Tuttavia, il paradigma dei modelli LM prevede la possibilità di inserire una o più covariate che, congiuntamente agli indicatori del modello, possano contribuire alla stima delle probabilità iniziali degli stati latenti e delle probabilità di transizione (si rinvia a Bartolucci e Farcomeni 2009, Bartolucci *et al.* 2013, Bartolucci *et al.* 2014, Vermunt *et al.* 1999 per una trattazione più approfondita).

Nello specifico dell'applicazione presentata nelle sezioni precedenti, può essere di interesse, ad esempio, mettere in relazione lo stato di deprivazione materiale con la variabile 'Paese di appartenenza'. In tal caso la formalizzazione del modello deve essere estesa al fine di inserire opportunamente la covariata oggetto di studio. Coerentemente con quanto visto nei paragrafi precedenti, sia $\phi_{h|k} = \Pr(Y_{ith} = 1|U_{it} = k)$, la probabilità che l' h -esimo indicatore associato all' i -esima unità statistica che appartiene al k -esimo stato latente assuma valore pari a 1 e sia inoltre X_{it} una covariata (categorica o continua) osservata su tutte le unità statistiche al tempo t . Mantenendo le stesse assunzioni di base, il modello LM, in seguito all'inserimento di una covariata, risulta definito dal seguente sistema di tre equazioni simultanee. La prima di queste è data da:

$$P(Y_{it1} = y_1, \dots, Y_{itR} = y_R | U_{it} = k) = \prod_{h=1}^R \phi_{h|k}^{y_h} (1 - \phi_{h|k})^{1-y_h} \quad (2.3)$$

Questa rappresenta la probabilità congiunta degli item per l' i -esima unità statistica condizionatamente al k -esimo stato latente. La seconda equazione invece è data da:

$$\log \frac{P(U_{i1} = k | X_{i1})}{P(U_{i1} = 1 | X_{i1})} = \alpha_{1k} + \beta_{1k} X_{it} \quad (2.4)$$

e stabilisce la relazione che intercorre tra le probabilità iniziali degli stati latenti e la covariata inserita nel modello. Infine, per quanto riguarda la terza delle equazioni simultanee, questa è data da:

$$\log \frac{P(U_{it} = k | U_{i,t-1} = j, X_{it})}{P(U_{it} = j | U_{i,t-1} = j, X_{it})} = \alpha_{tkj} + \beta_{tkj} X_{it} \quad (2.5)$$

e serve a modellizzare le probabilità di transizione tra gli stati latenti in funzione di X_{it} . Le equazioni (2.3), (2.4) e (2.5) consentono di definire il ruolo delle covariate nella distribuzione di probabilità congiunta della variabile Y_i e della variabile U_i , che è data da:

$$P(Y_i | X_i) = \sum_{k_{i1}=1}^K \dots \sum_{k_{iT}}^{K_{iT}} (P(U_{i1} = k_{i1} | X_{i1}) \prod_{t=2}^T P(U_{it} = k_{it} | U_{i,t-1} = k_{i,t-1}, X_{it})) \times \prod_{t=1}^T P(Y_{it} | X_{it}, U_{it} = k_{it}) \quad (2.6)$$

L'equazione di cui sopra permette di definire la funzione di log-verosimiglianza del vettore dei parametri del modello che è data da

$$l(\theta) = \sum_{i=1}^n \log P(Y_i | X_i) \quad (2.7)$$

Massimizzando, tramite una procedura numerica, la funzione di log-verosimiglianza si ottengono, come nel caso precedente le stime di massima verosimiglianza dei parametri che costituiscono il modello ovvero i parametri α e β che governano le probabilità iniziali e di transizione e le response probabilities $\phi_{h|k}$.

2.3.2 Il ruolo delle covariate nei modelli LM

Le equazioni (2.3), (2.4) e (2.5) prevedono, per definizione, che la covariata X_{it} non abbia effetti diretti sulle response probabilities $\phi_{h|k}$. Questa condizione prende il nome in letteratura di 'measurement invariance' e implica che gli indicatori $Y_{i1}^t, Y_{i2}^t, \dots, Y_{iR}^t$ siano indipendenti tra di loro condizionatamente allo stato latente e anche che gli indicatori siano indipendenti dalle covariate a loro volta condizionatamente allo stato latente. Il concetto di 'measurement invariance' implica, in termini pratici, che

l'utilizzo dello stesso questionario (o degli stessi indicatori) in differenti gruppi riesce a misurare lo stesso costrutto latente nella stessa maniera (Chen 2008). In altri termini, come riportato in Horn e McArdle (1992), sussiste 'measurement invariance' ogni qualvolta che 'al variare delle condizioni di osservazione di un fenomeno, le rilevazioni effettuate portano alla misurazione degli stessi attributi'.

Nel contesto dei modelli LM pertanto, l'assunzione di 'measurement invariance' è verificata se gli indicatori del modello, $Y_{i1}^t, Y_{i2}^t, \dots, Y_{iR}^t$, condizionatamente allo stato latente, non dipendono dalle covariate. Nel caso della misurazione della deprivazione materiale, la probabilità che una famiglia ha di essere deprivata può cambiare in funzione del Paese di appartenenza della famiglia ma, condizionatamente al Paese di appartenenza, il significato degli item di deprivazione rimane lo stesso per tutti i Paesi. Suddetta assunzione, in molti contesti applicativi, è ben poco realistica. Quando non sussiste 'measurement invariance' alcuni item possono avere differenti *response probabilities* a seconda del valore che le covariate assumono. Formalmente, quando valori di $\phi_{h|k}$ cambiano al variare delle covariate e si parla di 'funzionamento differenziale degli item', comunemente chiamato 'Differential Item Functioning' (DIF). Lo studio della *measurement invariance* e delle sue conseguenze ha portato alla pubblicazione di un numero piuttosto esteso di pubblicazioni riguardanti modelli LM e modelli a equazioni strutturali (Kankaraš *et al.* 2010, Munck *et al.* 2018). In questa sede ci si concentrerà principalmente sui diversi tipi di DIF che possono configurarsi nei modelli LM e sulle conseguenze che questi hanno nella stima dei parametri del modello LM, ripercorrendo, a grandi linee, la trattazione presentata in Di Mari *et al.* (2022).

2.3.3 DIF: diverse declinazioni per il funzionamento differenziale degli item

La situazione di funzionamento differenziale degli item prevede che le covariate possano avere effetti diretti sulle probabilità che gli item assumano valore pari a 1 nei diversi stati latenti e nei diversi tempi t . Formalmente, definita la funzione logit come $\text{logit}(x) = \log(x/(1-x))$ il funzionamento differenziale degli item viene espresso tramite la relazione:

$$\text{logit}(\phi_{ht|k}) = \gamma_{hk} + \eta_{htk} X_{it} \quad (2.8)$$

dove il coefficiente η_{htk} rappresenta l'effetto della covariata X_{it} sulle probabilità $\phi_{ht|k}$ che varia a seconda del tempo t , dello stato latente k e a

seconda dell'item di riferimento. Detta situazione viene chiamata in letteratura con il nome di 'Full DIF' o, in alternativa, DIF non uniforme. Di contro, la situazione di measurement invariance, ovvero la situazione in cui le covariate non hanno effetto sulle *response probabilities*, può essere espressa tramite l'equazione di cui sopra imponendo $\eta_{htk} = 0$. Infine, in Di Mari *et al.* (2022), sono previste alcune situazioni intermedie: la più restrittiva impone che l'effetto delle covariate sugli indicatori sia indipendente dallo stato latente e dal tempo. In tal caso si parla DIF costante sugli stati e sul tempo ('State Time Constant DIF') e vale la relazione:

$$\text{logit}(\phi_{ht|k}) = \gamma_{hk} + \eta_{h..} X_{it} \quad (2.9)$$

Seguendo la stessa logica è possibile definire altre due forme di DIF che prendono il nome di DIF costante rispetto al tempo ('Time Constant DIF') definito come:

$$\text{logit}(\phi_{ht|k}) = \gamma_{hk} + \eta_{h..k} X_{it} \quad (2.10)$$

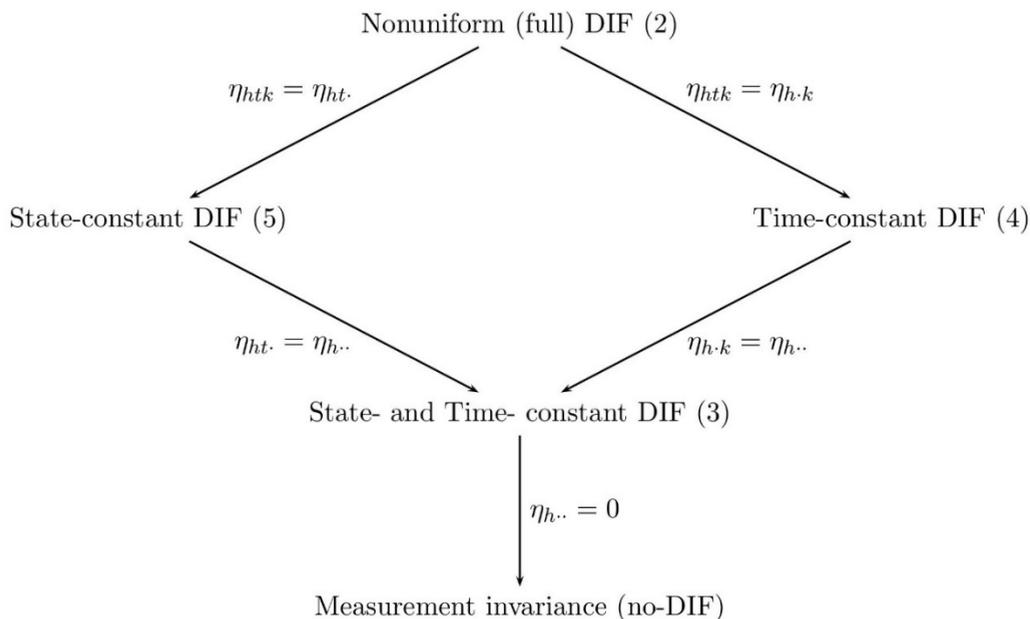
e DIF costante rispetto agli stati, altrimenti detto 'State Constant DIF' che è definito tramite la relazione:

$$\text{logit}(\phi_{ht|k}) = \gamma_{hk} + \eta_{ht..} X_{it} \quad (2.11)$$

Le parametrizzazioni di cui sopra delineano, fondamentalmente, la struttura di dipendenza che può potenzialmente intercorrere tra le *response probabilities* e le covariate: in termini estremamente sintetici questa può essere costante tra gli stati, così come nel tempo così come variare tra gli stati, nel tempo e per ogni item. Per una trattazione più approfondita si rimanda a Di Mari *et al.* 2022 dove viene dettagliatamente delineata la struttura gerarchia che intercorre tra le diverse forme di DIF e che può essere sintetizzata tramite la Figura 2.3.

Figura 2.3

La relazione gerarchica che intercorre tra i diversi modelli LM al variare del DIF



Fonte: Di Mari *et al.* (2022)

Di Mari *et al.* (2022) dimostrano, tramite un approfondito studio di simulazione (si confronti la Sezione 4 della referenza citata), come la mancata individuazione della struttura DIF sottostante i dati possa provocare forti distorsioni delle stime dei parametri e la conseguente errata classificazione delle osservazioni da parte del modello. A tal fine gli autori propongono di stimare tutte e cinque le configurazioni di modelli LM imponendo le diverse strutture di DIF e selezionare il modello migliore utilizzando il criterio di informazione Bayesiano (BIC) presentato in Schwarz (1978). Lo studio di simulazione mostra come, con altissima probabilità, il modello corretto venga selezionato e pertanto, si ottengono stime più efficienti e una classificazione delle osservazioni più precisa.

2.3.4 DIF e deprivazione materiale: una valutazione sui dati EU-SILC

Concludiamo questa sezione incentrata sul ruolo delle covariate e sulla loro relazione con gli indicatori del modello applicando i principi di cui sopra ai dati EU-SILC. Nell'analisi è stata inserita anche la covariata 'Paese

di appartenenza' e sono stati imposti i cinque tipi di DIF delineati nella precedente sezione. La Tabella 2.7 stabilisce che, nei dati EU-SILC, c'è un effetto diretto della covariata 'Paese' sugli indicatori del modello. Detto effetto diretto, rimane invariato nel tempo ma cambia da Paese a Paese nei diversi stati latenti: si tratta pertanto di DIF costante nel tempo.

Tabella 2.7 – I diversi tipi di DIF applicati ai dati SILC con i valori corrispondenti della funzione di log verosimiglianza, il numero dei parametri del modello e il corrispettivo valore del BIC

| Model | $l(\theta)$ | # parametri | BIC |
|--------------------------|---------------|-------------|--------------|
| No DIF | -45530 | 27 | 91279 |
| State Time Constant DIF | -44614 | 45 | 89616 |
| Time Constant DIF | -44384 | 63 | 89296 |
| State Constant DIF | -44451 | 99 | 89752 |
| Full DIF | -44173 | 171 | 89801 |

Fonte: elaborazione dell'autore

In termini pratici, questa informazione conferma le analisi precedenti: l'impatto della covariata Paese è diverso nei due stati latenti e sui diversi item e rimane, tuttavia, invariato nel tempo. Tale risultato non stupisce dal momento che, già dall'analisi della Tabella 2.6, è risultato piuttosto chiaro come i diversi item concorrano in maniera totalmente diversa a determinare lo stato di deprivazione materiale di una famiglia. In ultima analisi riportiamo il valore di alcuni coefficienti η_{hkt} in modo da fornire gli strumenti per interpretare il loro ruolo nel modello. Poiché gli indicatori sono variabili di tipo binario i coefficienti η_{hkt} rappresentano l'impatto della covariata X_{it} sulla probabilità che l' h -esimo indicatore assuma valore pari a 1 al tempo t . Trattandosi a tutti gli effetti di una regressione di tipo logistico l'effetto del coefficiente η_{hkt} è espresso in scala logaritmica e il suo effetto agisce sull'odds dell'evento $Y = 1$ rispetto all'evento $Y = 0$.

Tabella 2.8 – *Alcuni dei coefficienti η del modello DIF costante rispetto al tempo applicato ai dati SILC*

| Parametro | Coefficiente | Paese |
|-------------|--------------|--------|
| η_{21} | 0.540 | Grecia |
| η_{21} | 0.161 | Italia |
| η_{42} | -0.476 | Grecia |
| η_{42} | -0,087 | Italia |
| η_{81} | 0.398 | Grecia |
| η_{81} | -0.709 | Italia |
| η_{92} | 2.040 | Grecia |
| η_{92} | -0,502 | Italia |

Fonte: elaborazione dell'autore

I coefficienti nella Tabella 2.8 rispecchiano, a nostro avviso, alcune caratteristiche peculiari dei paesi oggetto di studio. Trattandosi di una covariata categorica i coefficienti sono riportati in funzione della categoria baseline che nel caso in questione è data dal Regno Unito. Nello specifico dell'interpretazione del valore dei coefficienti si ha che η_{21} è pari al valore 0.54 per quanto riguarda la Grecia. Pertanto, tenendo in considerazione che ci si sta riferendo al secondo item di deprivazione – ovvero l'item riguardante la possibilità delle famiglie di potersi permettere una settimana di ferie – e poiché $\exp(0.540)=1.71$, se ne conclude che una famiglia greca materialmente deprivata ha una probabilità del 71% maggiore rispetto a una famiglia in UK di non potersi permettere una settimana di vacanza. Particolare, inoltre, è l'interpretazione del parametro η_{81} che è riferito all'item relativo al possesso dell'automobile. Questo, per quanto riguarda l'Italia assume valore pari a -0.709 e, poiché $\exp(-0.709) = 0.49$ questo ci dice che una famiglia Italiana deprivata, ha una probabilità di non potersi permettere un'automobile del 49% inferiore rispetto a una famiglia britannica. Dal nostro punto di vista questo dato rispecchia alcune caratteristiche peculiari dei due Paesi: in Italia, il mancato possesso di un mezzo proprio, soprattutto per chi vive nelle grandi città, è sinonimo di forte deprivazione. Di contro, in Paesi come il Regno Unito, che gode mediamente di un trasporto pubblico particolarmente ben sviluppato, soprattutto nelle grandi città, questo item risulta meno discriminante per determinare lo stato di deprivazione materiale di una famiglia. In ultimo risulta particolarmente estremo il valore del coefficiente η_{92} che è pari, per

la Grecia a 2.04. Il nono item di deprivazione rappresenta la possibilità di pagare spese arretrate e il fatto che $\exp(2.04)=7.609$ ci dice che, nell'ambito delle famiglie non deprivate, una famiglia greca ha una probabilità circa 8 volte maggiore di avere arretrati rispetto a una famiglia che proviene dall'UK.

2.4 Conclusioni

I modelli Latent Markov costituiscono una possibile strategia di analisi per la risoluzione di un'ampia gamma di problemi econometrici. In questo contributo si sono voluti delineare i principali aspetti tecnici di questa classe di modelli statistici e alcune possibili applicazioni con un'attenzione particolare al problema della stima della deprivazione materiale. In seconda battuta ci si è voluto concentrare sulla possibilità di inserire una o più covariate all'interno del modello e sono state discusse le accortezze da prendere in termini del funzionamento differenziale degli item. I vantaggi dell'utilizzo dei modelli LM risiedono nella loro flessibilità e nella facilità di interpretazione dei parametri del modello. Di contro, le limitazioni di suddetta classe di modelli risiedono principalmente nell'onere computazionale necessario alla loro stima. Le procedure iterative necessarie a massimizzare la funzione di verosimiglianza del vettore dei parametri del richiedono di essere inizializzate molte volte e in punti molto distanti tra di loro all'interno dello spazio parametrico al fine di arrivare a convergenza. Per quanto riguarda invece la strategia delineata per individuare la corretta configurazione di DIF, questa, a sua volta, richiede di stimare ripetutamente il modello e valutare ex post il miglior modello disponibile in termini di adattamento ai dati e parsimonia, il tutto a danno del costo computazionale complessivo. Una possibile linea di ricerca in questa direzione è quella di includere criteri di penalizzazione della funzione di verosimiglianza direttamente nella formalizzazione del modello, quali ad esempio un approccio basato sul metodo LASSO (Tibshirani 1996) al fine di contenere il costo computazionale della procedura ma lasciando pressoché intatta l'interpretabilità del modello.

Bibliografia

- Bartolucci, F., & Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, 104(486), 816-831.
- Bartolucci, F., Farcomeni, A., & Pennoni, F. (2012). *Latent Markov models for longitudinal data*. CRC Press.
- Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test*, 23, 433-465.
- Brzezińska, J. (2016). Analysis of latent class models in economic research. *Ekonometria*, (54), 36-47.
- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of personality and social psychology*, 95(5), 1005.
- Collins, L.M., & Lanza, S.T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.
- Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1), 1-22.
- Di Mari, R., Oberski, D.L., & Vermunt, J.K. (2016). Bias-adjusted three-step latent Markov modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5), 649-660.
- Di Mari, R., Dotto, F., Farcomeni, A., & Punzo, A. (2022). Assessing measurement invariance for longitudinal data through latent Markov models. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(3), 381-393.
- Dotto, F., Farcomeni, A., Pittau, M.G., & Zelli, R. (2019). A dynamic inhomogeneous latent state model for measuring material deprivation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2), 495-516.
- European Commission. Directorate-General for Employment, & Social Affairs. Unit E. (2004). *Joint report on social inclusion*. Office for Official Publications of the European Communities.
- Guio, A.C., Gordon, D., & Marlier, E. (2012). *Measuring material deprivation in the EU: Indicators for the whole population and child-specific indicators*. Eurostat methodologies and working papers, Publications Office of

- the European Union, Luxembourg.
- Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, 18(3), 117-144.
- Kankaraš, M., Vermunt, J.K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, 40(2), 279-310.
- Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, 47(4), 637-664.
- Poulsen, C.S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing*, 7(1), 5-19.
- Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*. 461-464.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267-288.
- Townsend, P. (1987). Deprivation. *Journal of social policy*, 16(2), 125-146.
- Vermunt, J.K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24(2), 179-207.
- Wiggins, L.M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. Jossey-Bass.